

RESEARCH HANDBOOK ON
**Artificial Intelligence
and Decision Making
in Organizations**

EE
Elgar

Edited by

Ioanna Constantiou

Mayur P. Joshi • Marta Stelmaszak



RESEARCH HANDBOOK ON ARTIFICIAL
INTELLIGENCE AND DECISION MAKING
IN ORGANIZATIONS

Research Handbook on Artificial Intelligence and Decision Making in Organizations

Edited by

Ioanna Constantiou

Full Professor, Department of Digitalization, Copenhagen Business School, Denmark

Mayur P. Joshi

Assistant Professor of Information Systems, Telfer School of Management, University of Ottawa, Canada

Marta Stelmaszak

Assistant Professor of Information Systems, The School of Business, Portland State University, USA

 **Edward Elgar**
PUBLISHING

Cheltenham, UK • Northampton, MA, USA

© The Editors and Contributors Severally 2024

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical or photocopying, recording, or otherwise without the prior permission of the publisher.

Published by
Edward Elgar Publishing Limited
The Lypiatts
15 Lansdown Road
Cheltenham
Glos GL50 2JA
UK

Edward Elgar Publishing, Inc.
William Pratt House
9 Dewey Court
Northampton
Massachusetts 01060
USA

A catalogue record for this book
is available from the British Library

Library of Congress Control Number: 2023952099

This book is available electronically in the **Elgaronline**
Business subject collection
<http://dx.doi.org/10.4337/9781803926216>

ISBN 978 1 80392 620 9 (cased)
ISBN 978 1 80392 621 6 (eBook)

Contents

<i>List of figures</i>	vii
<i>List of tables</i>	viii
<i>List of contributors</i>	ix

Introduction to <i>Research Handbook on Artificial Intelligence and Decision Making in Organizations</i>	1
<i>Ioanna Constantiou, Mayur P. Joshi and Marta Stelmaszak</i>	

PART I MAKING DECISIONS ABOUT AI

1	Sourcing data for data-driven applications: foundational questions <i>Sirkka L. Jarvenpaa</i>	17
2	Data work as an organizing principle in developing AI <i>Angelos Kostis, Leif Sundberg, and Jonny Holmström</i>	38
3	Natural language processing techniques in management research <i>Mike H.M. Teodorescu</i>	58
4	Captains don't navigate with a keyboard: developing AI for naturalistic decision-making <i>Adrian Bumann</i>	80
5	Reconfiguring human–AI collaboration: integrating chatbots in welfare services <i>Elena Parmiggiani, Polyxeni Vassilakopoulou, and Ilias Pappas</i>	97
6	Circumspection as a process of responsible appropriation of AI <i>Margunn Aanestad</i>	112
7	Responsible AI governance: from ideation to implementation <i>Patrick Mikalef</i>	126

PART II MAKING DECISIONS WITH AI

8	Human judgment in the age of automated decision-making systems <i>Dina Koutsikouri, Lena Hylving, Jonna Bornemark, and Susanne Lindberg</i>	144

9	Making decisions with AI in complex intelligent systems <i>Bijona Trope, Gunnar Holmberg, and Nicolette Lakemond</i>	160
10	Addressing the knowledge gap between business managers and data scientists: the case of data analytics implementation in a sales organization <i>Stella Pachidi and Marleen Huysman</i>	179
11	Constructing actionable insights: the missing link between data, artificial intelligence, and organizational decision-making <i>Arisa Shollo and Robert D. Galliers</i>	195
12	It takes a village: the ecology of explaining AI <i>Lauren Waardenburg and Attila Márton</i>	214
13	Synthetic stakeholders: engaging the environment in organizational decision-making <i>Jen Rhymer, Alex Murray, and David Sirmon</i>	226
14	Interpretable artificial intelligence systems in medical imaging: review and theoretical framework <i>Tiantian Xian, Panos Constantinides, and Nikolay Mehandjiev</i>	240
15	Artificial intelligence to support public sector decision-making: the emergence of entangled accountability <i>Francesco Gualdi and Antonio Cordella</i>	266
16	Contrasting human–AI workplace relationship configurations <i>Miriam Möllers, Benedikt Berger, and Stefan Klein</i>	282
PART III IMPLICATIONS OF DECISIONS MADE WITH AI		
17	Who am I in the age of AI? Exploring dimensions that shape occupational identity in the context of AI for decision-making <i>Anne-Sophie Mayer and Franz Strich</i>	305
18	Imagination or validation? Using futuring techniques to enhance AI's relevance in strategic decision-making <i>Andrew Sarta and Angela Aristidou</i>	322
19	Artificial intelligence as a mechanism of algorithmic isomorphism <i>Camille G. Endacott and Paul M. Leonardi</i>	342
20	Ethical implications of AI use in practice for decision-making <i>Jingyao (Lydia) Li, Yulia Litvinova, Marco Marabelli, and Sue Newell</i>	359
	<i>Index</i>	376

Figures

0.1	An integrative framework of AI and organizational decision making	10
2.1	Data work as an organizing principle in developing AI solutions	52
3.1	Typical workflow for processing text	60
3.2	Example of a typical workflow for processing a collection of text documents with an overview of the RapidMiner interface	64
4.1	Example display of ship predictor	86
7.1	Relationship between responsible AI principles and governance	129
7.2	Conceptual overview of responsible AI governance in broader ecosystem	131
9.1	Central aspects of decision-making in organizations	164
9.2	Personalized medicine overview	169
12.1	An ecology of explaining AI	220
14.1	Process workflow of screening mammograms by radiologists	243
14.2	Process workflow of augmenting the screening of mammograms with an interpretable AI system	244
14.3	An overview of the development process of an interpretable AI system in the medical image field, and the human roles in each step	245
14.4	Tensions emerging from an Interpretable AI system	254
16.1	A framework of human–AI relationship configurations	295

Tables

3.1	Example of an output of a topic model: topics identified through LDA versus top topics identified through HDP for the 2009–2012 patent claims corpus	69
7.1	Principles of responsible AI	128
7.2	Indicative themes and research questions	136
9.1	System characteristics of personalized medicine	168
9.2	New decision-making prerequisites for CoIS	171
13.1	Locus of design range from predominately internal to predominately external	232
14.1	Three components in an interpretable AI system and their classifications and dimensions	246
14.2	Human agents: classifications and dimensions for human agents	247
14.3	The classifications and dimensions for medical imaging data	249
14.4	Classifications and dimensions for interpretable AI models	251
14.5	Summary of the features and limitations of model-centric and data-centric models	252
16.1	Exemplary illustrations of human–AI relationship configurations	298
17.1	Overview of case organizations	308
20.1	AI characteristics, ethical considerations and remedies	371

Contributors

Margunn Aanestad, Department of Informatics, University of Oslo, Norway and Department of Information Systems, University of Agder, Norway.

Angela Aristidou, UCL School of Management, University College London, UK and Stanford CASBS, Stanford University, USA.

Benedikt Berger, Department of Information Systems, University of Münster, Germany.

Jonna Bornemark, Centre for Studies in Practical Knowledge, Södertörn University, Sweden.

Adrian Bumann, Department of Technology Management and Economics, Chalmers University of Technology, Sweden.

Panos Constantinides, Alliance Manchester Business School, University of Manchester, UK.

Ioanna Constantiou, Department of Digitalization, Copenhagen Business School, Denmark.

Antonio Cordella, Department of Management, London School of Economics, UK.

Camille G. Endacott, Department of Communication Studies, University of North Carolina at Charlotte, USA.

Robert D. Galliers, Departments of Information & Process Management and Sociology, Bentley University, USA and Warwick Business School, University of Warwick, UK.

Francesco Gualdi, Department of Management, London School of Economics, UK.

Gunnar Holmberg, Department of Management and Engineering, Linköping University, Sweden and Saab Aeronautics, Sweden.

Jonny Holmström, Department of Informatics, Umeå University, Sweden.

Marleen Huysman, KIN Center for Digital Innovation, School of Business and Economics, Vrije Universiteit Amsterdam, The Netherlands.

Lena Hylving, Department of Informatics, University of Oslo, Norway and School of Information Technology, Halmstad University, Sweden.

Sirkka L. Jarvenpaa, McCombs School of Business, University of Texas at Austin, USA.

Mayur P. Joshi, Assistant Professor of Information Systems, Telfer School of Management, University of Ottawa, Canada.

Stefan Klein, Department of Information Systems, University of Münster, Germany.

Angelos Kostis, Management Science and Engineering and SCANCOR, Stanford University, USA; Umeå School of Business, Economics, and Statistics, Umeå University, Sweden, and Department of Informatics, Umeå University, Sweden.

Dina Koutsikouri, Department of Applied IT, University of Gothenburg, Sweden.

Nicolette Lakemond, Department of Management and Engineering, Linköping University, Sweden.

Paul M. Leonardi, Department of Technology Management, University of California Santa Barbara, USA.

Jingyao (Lydia) Li, Computer Information Systems, Bentley University, USA.

Susanne Lindberg, School of Information Technology, Halmstad University, Sweden.

Yulia Litvinova, Economics Group, IHK-Chair of Small and Medium-Sized Enterprises, Otto Beisheim School of Management, Germany.

Marco Marabelli, Computer Information Systems, Bentley University, USA.

Attila Márton, Department of Digitalization, Copenhagen Business School, Denmark.

Anne-Sophie Mayer, KIN Center for Digital Innovation, Vrije Universiteit Amsterdam, The Netherlands.

Nikolay Mehandjiev, Alliance Manchester Business School, University of Manchester, UK.

Patrick Mikalef, Department of Computer Science, Norwegian University of Science and Technology (NTNU), Norway and Department of Technology Management, SINTEF Digital, Norway.

Miriam Möllers, Department of Information Systems, University of Münster, Germany.

Alex Murray, Lundquist College of Business, University of Oregon, USA.

Sue Newell, Warwick Business School, University of Warwick, UK.

Stella Pachidi, Judge Business School, University of Cambridge, UK.

Ilias Pappas, Department of Computer Science, Norwegian University of Science and Technology (NTNU), Norway, and Department of Information Systems, University of Agder, Norway.

Elena Parmiggiani, Department of Computer Science, Norwegian University of Science and Technology (NTNU), and Sintef Nord AS, Norway.

Jen Rhymer, School of Management, University College London, UK.

Andrew Sarta, School of Administrative Studies, York University, Canada.

Arisa Shollo, Department of Digitalization, Copenhagen Business School, Denmark.

David Sirmon, Foster School of Business, University of Washington, USA.

Marta Stelmaszak, The School of Business, Portland State University, USA.

Franz Strich, Department of Information Systems & Business Analytics, Deakin University Melbourne, Australia.

Leif Sundberg, Department of Informatics, Umeå University, Sweden.

Mike H.M. Teodorescu, Information School, University of Washington, USA.

Bijona Troqe, Department of Management and Engineering, Linköping University, Sweden.

Polyxeni Vassilakopoulou, Department of Information Systems, University of Agder, Norway.

Lauren Waardenburg, Department of Information Systems, Decision Sciences and Statistics, ESSEC Business School, France.

Tiantian Xian, Alliance Manchester Business School, University of Manchester, UK.

Introduction to *Research Handbook on Artificial Intelligence and Decision Making in Organizations*

Ioanna Constantiou, Mayur P. Joshi and Marta Stelmaszak

THE HYPE AND THE REALITY OF AI IN DECISION MAKING

Zillow is a multi-sided digital real estate platform founded in 2006. It predominantly deals in United States properties, and has recently started listing properties from Canada. Zillow collects zillions of data about properties and digitalizes buying and selling. To exploit the data collected over the years, Zillow developed an artificial intelligence (AI)-based algorithm called Zestimate. Launched in 2006, Zestimate came with a promise to predict the future price of any house. Building on the potential of data and real estate market inefficiencies due to numerous intermediaries, Zillow adopted a vertical strategy, directly competing with these intermediaries, such as real estate agents and mortgage lenders. The company was able to make instant offers to buy a house, available for 48 hours. These offers were based on the Zestimate algorithm combining data about the property from public and private sources, and information provided by the seller. Initially very successful, over the years Zestimate has been improved to decrease error margins of price estimates. Most recently, Zillow added a neural network to increase the frequency of model updates and reduce errors in more traditional models. Zillow also explored other data sources through natural language processing. Based on these growing capabilities, in 2019 Zillow introduced a new business model: buy a property after a brief inspection, for a price recommended by Zestimate, and flip it (renovate to sell it at a higher price shortly after). An entire department grew around property purchase decisions with Zestimate, owing to the abundance of data used by the algorithm to predict buying and selling prices. Yet, in 2021 Zillow lost over \$800 million on its new business after it ended up saddled with around 7000 houses to be sold below purchase prices. In consequence, Zillow shut the Zestimate algorithm down and let go of 2000 workers, a quarter of its staff. What happened at Zillow? (See Metz, 2021; Parker and Putzier, 2021 for details.)

The enthusiasm for (and criticism of) AI among academics as well as practitioners is at its peak. Numerous companies, just like Zillow, try to harness AI, engaging researchers and industry specialists. The standard narrative in support of AI points to the extensive computing power that can be applied to the immense pools of structured as well as unstructured data to aid managerial and professional decision making in organizations (Brynjolfsson and McAfee, 2017; Davenport and Kirby, 2016; Miller, 2018; Polli, 2019). The combination of more data and more information processing makes AI-enabled technology potentially better than humans at evaluating options and optimizing choices (Davenport et al., 2012; McAfee and Brynjolfsson,

2012; Varian, 2010). Arguing against business managers' intuitive decision making, experts have conceptualized the role of AI in using data extensively to engender the practices of fact-based management in driving decisions (Davenport et al., 2010).

In the spirit of treating AI as a dynamic frontier of computing (as well as organizing) (Berente et al., 2021), we refrain from offering a specific definition of AI. Instead, we emphasize that the definition of AI is ever evolving since its earlier conceptualization (Simon, 1969), and the latest wave focuses on the ability of algorithms to mimic the human capacities to learn and act, that is, machine learning (ML) and deep learning (DL) (Brynjolfsson and Mitchell, 2017). With this ability comes the key promise of going beyond menial tasks. In earlier waves, AI technologies could only handle routine tasks, leaving the more complex reasoning to humans, making AI irrelevant to managerial decision making (Raisch and Krakowski, 2021). However, in the current wave, it is argued and also shown that AI is capable of performing not only routine and structured tasks but also non-routine and cognitive tasks. In particular, AI is seen as enabling organizations (and individuals) to make accurate (and cheap) predictions, thereby reducing uncertainty in decision making (Agrawal et al., 2018). With superlative predictive power, the latest AI offers the possibility to overcome human biases (Kahneman et al., 2021). At least there is consensus among the proponents of AI that it is much better at predicting known knowns compared to humans (Agrawal et al., 2018) (in the case of Zillow, predicting the house buying prices based on historical data). These capacities are fundamental in decision making, allowing AI algorithms to compete with human decision makers.

Despite high expectations, we have seen some anecdotal as well as well-researched academic cases of how AI often fails to deliver on some promises. Zillow's Zestimate is a good example of AI deployed to make decisions instead of humans (Troncoso et al., 2023), and its failure to do so serves as a perfect background for this *Handbook*. In this introduction, we frequently refer to this case to show both the promises and the challenges of AI in decision making. On the one hand, Zillow's story is about the power of AI-based predictions and how AI may substitute for various cognitive tasks. Zillow's decisions about the introduction of the predictive pricing service, the particular use of predictive models, the choice of datasets, and the technological updates to increase the frequency of data updates and improve prediction accuracy have transformed the real estate market. A number of experts, especially real estate agents, have been progressively replaced in selling and buying properties. They play a limited role in setting prices, and they become controllers of the predictive model's outputs. Hence, the decision making process changed significantly, and human experts, such as real estate agents, shifted to new activities due to the automation of their traditional tasks. Other experts, such as house inspectors, experienced augmentation of their activities by improving price estimates after physical property inspections.

On the other hand, the case also shows the perils of deploying AI in decision making. Zillow's predictive model worked well for some time, but the exogenous shocks to the economy because of the pandemic rendered the model inefficient, with severe implications for Zillow's bottom line (Metz, 2021). Shifts such as volatility in

personal income and changes in consumption patterns rendered predictive algorithms built on historical data inaccurate. Firms using algorithms relying on historical data experienced significant challenges because of the inaccuracy of the predictions. The Zillow case shows the boundaries of AI-powered predictions which may fail in the situations of unknown knowns (predicting buying prices when the economic factors that determine the prices are influenced by an external shock).

Complementing the Zillow example, we have seen several well-researched cases where AI does not lead to the expected results in line with its promises of data-driven decision making (Lebovitz et al., 2021, 2022; van den Broek et al., 2021; Waardenburg et al., 2022; Pachidi et al., 2021). Research has shown the aspects of professional expertise that AI cannot know (Pakarinen and Huising, 2023; Lebovitz et al., 2022), as well as how decision makers may adopt AI only symbolically but not in practice (Pachidi et al., 2021). Research has also demonstrated the implications of AI adoption for the identity of its developers (Vaast and Pinsonneault, 2021) as well as its users (Strich et al., 2021) at the individual level, as well as how it can contribute to inertia at the organizational level (Omidvar et al., 2023). We need to examine AI in much more depth in the context within which it is implemented, as AI in theory may not necessarily be the same as AI in practice (Anthony et al., 2023), especially not in the complex practice of decision making.

ORGANIZATIONAL DECISION MAKING MEETS AI

Literature on Organizational Decision Making

In order to understand how organizations, managers, and professionals make decisions with AI, we need to trace the genealogy of organizational decision making in the classic literature. Information has been recognized as an integral lever in the process of organizational decision making in information systems (IS) as well as organizational theory. Researchers in the latter field have extensively studied the role of information in organizational decision making processes for several decades. A prominent stream of literature within this domain is the Carnegie School tradition (e.g., March and Simon, 1958), which has examined the relationship between information and decision making by relying on the information processing perspective (Gavetti et al., 2007). Built on the foundational work of Herbert Simon (Joseph and Gaba, 2020), the information processing perspective affords a vast and wide literature within the organizational theory, and we only review a fragment of the literature that is relevant to this introduction.

Organizations are conceived as information processing and decision making systems (March and Simon, 1958). Organizational decision making and problem solving are conceived to have four broad stages: agenda setting, problem representation, search, and evaluation (Simon, 1947). Organizations need to structure the processing of information in terms of gathering, interpreting, and synthesizing (Tushman and Nadler, 1978) to facilitate decision making processes, which are

performed by inherently boundedly rational humans (Simon, 1997). The notion of bounded rationality implies that managers bring to work and rely on a set of simplified mental models in each of the stages of problem solving (Gavetti and Levinthal, 2000). Effective information processing is posited to include “the collection of appropriate information, the movement of information in a timely fashion, and its transmission without distortion ... [as well as] ...the ability to handle needed quantities of information according to these criteria” (Tushman and Nadler, 1978, p. 617), and advanced IS are believed to be aiding organizations in making effective information processing possible by reducing the information processing requirements (Galbraith, 1974; Huber, 1990).

Three observations can be made about the existing literature on the information processing perspective. First, managers and professionals—individuals responsible for making decisions—are the key (and the only) actors considered to be important (Turner and Makhija, 2012), and they are assumed to be cognitively bounded yet rational and hence logical. Second, the technologies that facilitate information processing are mostly the technologies of information storage, aggregation, and retrieval (Huber, 1990) that treat information as given, unlike the technologies that facilitate generating insights from data (for example, AI-enabled analytics tools). Third, even though mentioned otherwise (Tushman and Nadler, 1978), information is often treated as indistinguishable from data. In other words, these studies adopt a “token view” of information (Boell, 2017), where information is an “undifferentiated commodity of data bits that are processed” (McKinney and Yoos, 2010, p. 331).

The information processing perspective assumes bounded yet instrumental use of information in decision making which is challenged by the researchers of socio-political processes. In this view, the information processing perspective—where “information is regarded as an input to decision making and the decision maker is considered a passive recipient of this information” (Schultze, 2000, p. 3)—is considered problematic, as information is often decoupled from decision making in organizations. There are several conspicuous features that hinder the instrumental use of information (Feldman and March, 1981). Treating organizations as political coalitions, scholars in this stream of literature adopt a socio-political perspective that goes beyond individual decision makers’ cognitive limitations and demonstrates that the conflicts of interest between self-interested individuals or groups inevitably form the background of organizational decision making (March, 1962; March and Olsen, 1984). Instead of the available and relevant information, these studies show that choices are often made based on the bargaining and preferences of the most powerful actors (Pettigrew, 1973). Overall, the large body of existing literature on organizational decision making has focused on how managers or professionals process (or ignore) information in ways characterized as boundedly rational, socially situated, or politically motivated.

AI in Organizational Decision Making

The chapters in this book investigate the relationship between AI and decision making in a variety of ways. As we show below, introducing AI to decision making brings about more significant changes than deploying traditional information systems. Unlike many previous decision support systems, for example, AI can no longer be seen just as a tool or a medium (Anthony et al., 2023). Scholars have highlighted that research and practice on decision making with AI are based on “an implicit assumption [that organizations] can capture value while continuing to function as before” (Sharma et al., 2014, p. 434). But we know from the experiences of practitioners as well as scholars that after implementing AI neither organizations nor technologies function as before. Scholars have increasingly started to recognize this challenge. For instance, in the updated version of their book *Power and Prediction*, Agrawal et al. (2022) explain why some of their predictions about the widespread commercialization of AI in *Prediction Machines* (Agrawal et al., 2018) did not take shape as they envisioned:

Our focus on the possibilities of prediction machines had blinded us to the probability of actual commercial deployments. While we had been focused on the economic properties of AI itself—lowering the cost of prediction—we underestimated the economics of building the new systems in which AIs must be embedded. (Agrawal et al., 2022, p. xii)

On similar lines, Anthony et al. (2023) argue that we should conceptualize and examine AI as a counterpart. Such a view may help us to realize that managerial as well as professional decision making is perhaps not only cognitive but an inherently relational phenomenon (Pakarinen and Huising, 2023).

In this *Handbook*, we offer such a relational understanding of AI and decision making by proposing an integrative framework that explicates three interconnected facets: making decisions about AI, making decisions with AI, and implications of decision making with AI. These three facets constitute three parts in this *Handbook*, which we summarize below.

Part I: Making Decisions about AI

While the majority of research and practice focuses on making decisions with AI (and this flows naturally by conceptualizing the role of AI in theories and practice of organizational decision making), this *Handbook* demonstrates that this is not where the relationship between AI and decision making begins. Several chapters that have examined decision making with AI (Part II) have shown that the process starts not only when the insights generated by AI are ready to be consumed in decision making, but rather during generating those insights and even earlier, identifying and deploying those algorithms in the first place. In other words, making decisions about AI (Part I) emerged as a precursor to making decisions with AI. In particular, our chapters highlight four aspects of decisions about AI: making decisions about

the data, making decisions about the algorithms, making decisions about developing context-specific models, and making decisions about AI responsibly.

Making decisions about data

Jarvenpaa (Chapter 1) highlights the processes and practices of sourcing data for data-driven applications: unlike the traditional IS sourcing processes that rely on mechanisms of trust and control, data sourcing processes tend to be predicated on organizational learning at their core. This is because how, when, and what organizations can learn from sourced data depends on what data are sourced, under what arrangements, and data management established. Extending this line of argument, Kostis et al. (Chapter 2) theorize data work as an organizing principle for developing AI. The authors propose three key mechanisms of cultivating knowledge interlace, triggering data-based effectuation, and facilitating multi-faceted delegations through which data work helps to address the epistemic uncertainty inherent in AI development and deployment.

Making decisions about algorithms

In addition to data, another key decision that organizations (as well as researchers) employing AI need to make is about choosing the right algorithm for their idiosyncratic context. Teodorescu (Chapter 3) covers a great landscape of various natural language processing (NLP) and ML tools available for processing vast amounts of structured and unstructured data in the hands of practitioners as well as researchers, and maps the process of generating insights from data from a more technical perspective.

Making decisions about context-specific models

Even after making decisions about data and algorithms, organizations are still left with making decisions about the development and deployment of in-house models and artifacts that integrate those data and algorithms. Bumann (Chapter 4) explicates the challenges and mitigating strategies for AI development in the context of naturalistic decision making environments. Along similar lines, Parmiggiani et al. (Chapter 5) highlight the dynamics around the development of an AI-enabled chatbot for welfare services. Both chapters demonstrate how making decisions about AI inherently influences the decisions that are actually made with AI.

Making decisions about AI responsibly

Finally, making decisions about AI also entails considerations about the subsequent responsible implementation of AI. Aanestad (Chapter 6) draws on the actor-network theory to develop a concept of circumspection and posits it as a process of responsible appropriation of AI. The author makes a case for paying attention to how organizational capabilities, both pre-existing and novel, evolve and are impacted when organizations engage with AI. Along similar lines, Mikalef (Chapter 7) offers a framework for responsible AI governance and shows a way to translate responsible AI principles into responsible AI practices.

Summary

Overall the chapters in Part I show that it is not only “the managers that make all key decisions about AI” (Berente et al., 2021, p. 1434): the other stakeholders, especially data scientists, and developers who choose to develop specific models using specific algorithms and sourcing specific data, also play a pivotal role in the process.

Part II: Making Decisions with AI

Making decisions with AI opens up avenues for investigating how AI and related technologies are reshaping the landscape of organizational decision making. The chapters in Part II offer various ways to think about updating our theories in the wake of organizations making decisions with AI. We highlight several themes that are prevalent in the chapters: changes in the nature of information processing; going beyond traditional decision makers; stakeholders and ecosystems; as well as interpretability and explainability.

Changes in the nature of information processing

Recent literature has shown what AI and algorithms cannot know (Lebovitz et al., 2021; Pakarinen and Huising, 2023). Extending this line of reasoning, and contrary to the dominant accounts on AI alleviating the need for human judgment, Koutsikouri et al. (Chapter 8) show that human judgment, seen through the lens of *phronesis*, is all the more important in making decisions with AI. Similarly, Troqe et al. (Chapter 9) highlight the changing nature of the three key components of decision making: the decision maker, the decision making process, and the decision space, using an illustrative example of personalized medicine.

Going beyond traditional decision makers

It is increasingly becoming clear that making decisions with AI expands the scope of decision making beyond the decision makers as we understand it in the traditional sense. Research has shown that many large organizations are not only adopting ready-made AI algorithms but are also hiring data scientists who often develop such algorithms in-house. The introduction of data scientists in organizational settings opens up new dynamics around decision making. Pachidi and Huysman (Chapter 10) highlight the challenges of decision making emanating from the knowledge gap between business managers (that is, decision makers) and data scientists (that is, information producers). It then becomes interesting how such actors with distinct approaches to knowing come together to construct actionable insights for making decisions with AI in organizations (Shollo and Galliers, Chapter 11).

Stakeholders and ecosystems

Some chapters extend this line of reasoning of going beyond decision makers to broaden the horizon even further. For instance, Waardenburg and Márton (Chapter 12) argue for an ecological perspective of decision making that includes an entire ecology of unbounded, open-ended interactions and interdependencies. Rhymer et al.

(Chapter 13) offer an imaginative argument of how AI and other digital technologies can function as synthetic stakeholders, representing the environment in organizational decision making. Finally, Möllers et al. (Chapter 16) continue this trajectory by theorizing about a spectrum of making decisions with AI.

Interpretability and explainability

As we broaden the view on making decisions with AI, we can easily see the challenges related to interpretability and explainability. Xian et al. (Chapter 14) review the literature on interpretable AI in the context of medical imaging and offer a framework to guide further research in this space. A parallel to interpretability is accountability. If we cannot fully interpret an AI-powered algorithm, it raises questions about accountability. Gualdi and Cordella (Chapter 15) raise this issue and offer a conceptualization of entangled accountability in the context of public sector decision making.

Summary

Overall, these chapters show that AI is not merely a tool or a medium, but often a counterpart, an actor in organizational decision making (Anthony et al., 2023), and at times it signifies the system itself, of which humans become counterparts (see Demetis and Lee, 2018).

Part III: Implications of Decisions Made with AI

When organizations make decisions about and with AI, these decisions come with implications across multiple levels. The chapters in this section highlight the need to examine these implications.

Micro and meso level implications

Mayer and Strich (Chapter 17) demonstrate the implications at an individual level. The authors highlight that making decisions with AI influences the occupational identity of the decision makers depending on their skill level, the purpose of deployment of AI (augmentation or automation), and the extent to which the decisions are consequential to others. Making decisions with AI also influences other organizing processes and practices. Sarta and Aristidou (Chapter 18) conceptualize the implications of making decisions with AI on firms' strategy making. In particular, the authors argue that firms can employ AI either for validation or for imagination, which is conditional on whether they rely on well-established structures and processes or on a set of futuring techniques for the selection of strategic issues.

Macro level implications

The implications of making decisions with AI span beyond the organizational level. Endacott and Leonardi (Chapter 19) make a compelling case about how making decisions with AI may fuel algorithmic isomorphism at an institutional level. This new form of isomorphism can lead to challenges related to the reduction in the req-

uisite variety that is conducive to creativity and innovation, as well as the spread and reinforcement of bias. However, it may also help organizations to introspect their existing processes and practices.

Implications at all levels

Finally, across all levels, there are notable implications for ethical concerns. Li et al. (Chapter 20) take a lifecycle view of AI to explicate various ethical concerns that emanate from the long-term use of AI in decision making.

Summary

Overall, the chapters in Part III highlight the far-reaching implications of making decisions with AI. It is important to understand these implications not only because they are far-reaching, but also because they have implications on how decisions about and with AI are made.

DECISION REDISTRIBUTION: AN INTEGRATIVE FRAMEWORK OF ARTIFICIAL INTELLIGENCE IN ORGANIZATIONAL DECISION MAKING

Each chapter in this *Handbook* explicates a unique aspect of AI and decision making across its three facets: making decisions about AI, making decisions with AI, and implications of decision making with AI. We have listed each chapter under a specific facet of the phenomenon; however, most of our chapters traverse across the facets. First, it is not hard to imagine how the decisions about adopting and developing AI may influence how making decisions with AI plays out, which may in turn have implications for other organizing processes and practices. Second, the three facets also influence each other in an anticipatory manner. For instance, decisions about developing specific AI algorithms are usually influenced by what these algorithms are expected to achieve when they are deployed in decision making processes. Similarly, managers may use AI in a particular way in line with the implications they expect (for example, deploying AI such that it conforms with their identity). Sometimes, the decisions on the nature of AI could influence other organizing practices already in place, even when AI is not actually used (or only used symbolically) in making decisions.

These interdependencies indicate that AI in decision making cannot be viewed or studied just as making decisions with AI, making decisions about AI, or implications of AI in decision making. All three facets need to be considered together, both in practical implementations and in research. Figure 0.1 summarizes our integrative framework.

Our integrative framework offers an initial conceptualization of decision redistribution. We define it as the migration of human decision making (with or without an existing technology) across the three facets of AI, and organizational decision making as a result of deploying AI. In other words, when organizations implement

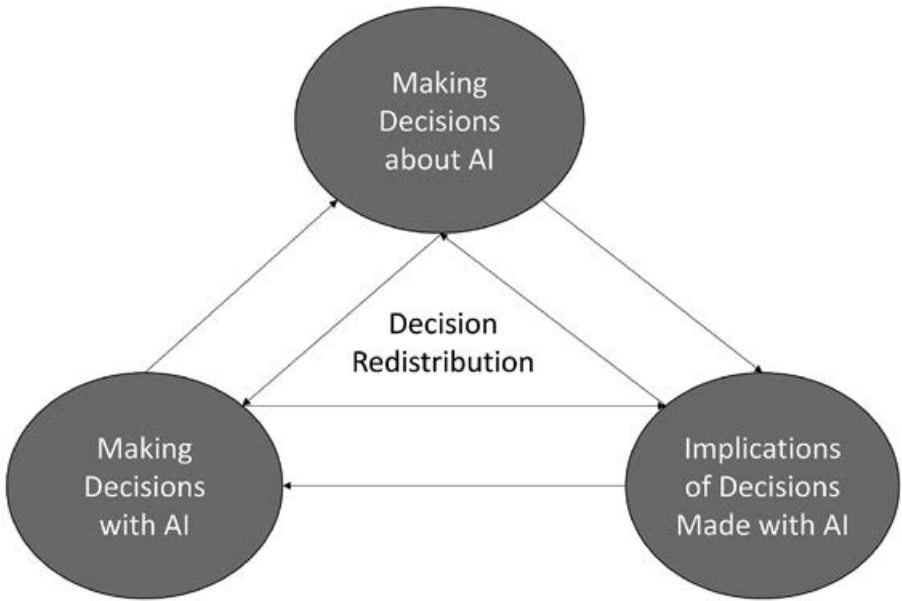


Figure 0.1 An integrative framework of AI and organizational decision making

AI for decision making, human decision making and makers do not simply disappear (even in the case of automation). Rather, the locus of decision making and the actors making decisions move around (not without changes) to the other facets, making and monitoring different decisions. This redistribution could be between old actors (for example, business managers or professionals) and new actors (for example, data scientists or ML engineers); between pre-existing information technology (IT) artifacts and new AI algorithms; between established work practices, routines, or mental models and new AI algorithms; or any combination of these. To be able to fully grasp the phenomenon, we need to pay attention to work practices or routines, identity, and epistemic stances of not only the traditional decision makers (for example, business managers, professionals) but also the developers (for example, data scientists, ML engineers). The practices, routines, and mental models of the actors are anchored on the broader organizational and institutional context and hence are influenced by as well as consequential to the broader decision ecosystem. The implications of such redistribution can be far-reaching. The realization of the promise of AI to aid data-driven decision making lies in the understanding of the dynamics around decision redistribution. In the following, we offer three scenarios to think about such redistribution.

In the first scenario, introducing decision making with AI may result in decision redistribution towards making decisions about AI and considering the implications of AI-driven decisions. This means that AI does not eliminate human decisions from the organization, but rather requires their redistribution to decision making about

AI. Such decisions may cover decisions about data (see Jarvenpaa, Chapter 1, and Kostis et al., Chapter 2, this *Handbook*) or algorithms (see Teodorescu, Chapter 3, this *Handbook*), and a number of other areas that further research should uncover. Decision redistribution is also in part what happened at Zillow, where newly hired data scientists carefully developed the Zestimate algorithm. Yet some aspects of the decision redistribution were not fully accounted for by the organization: human decision making regarding the data used was based on the past trends in a stable environment, which led Zillow to rely on easily available but insufficient datasets. As our framework indicates, practical applications of making decisions with AI should account for decision redistribution towards making decisions about AI to avoid similarly fraught outcomes.

Similarly, making decisions with AI may entail decision redistribution toward the implications of AI-driven decisions. As we have seen (for example, Li et al., Chapter 20, this *Handbook*), the variety of implications that decisions made by AI can bring often call for human oversight and accountability. For theory, this implies a complex relationship between making decisions and being accountable for them. While humans making decisions are traditionally seen as accountable for their outcomes (e.g., Simonson and Nye, 1992), who can be held to account for decisions made by AI, and how, and why? This is not solely an interesting research question, but a valid practical concern for organizations. In the end, Zillow had to fire a quarter of its workforce to recover from the bad decisions made by its AI (Metz, 2021); but who (or what) was responsible for this disastrous outcome?

In the second scenario, making decisions about AI results in decision redistribution to the other two facets: someone will need to make decisions with AI, and consider their implications. From the research perspective, decision redistribution offers a productive notion to theorize further how the decisions made about AI may dictate making decisions with AI. This approach can extend research on the various configurations of human–AI decision making (e.g., Shrestha et al., 2019; Murray et al., 2021). It can also offer productive ways to investigate the relationship between decisions made about AI and their potential implications, thus building a stronger link across design and its consequences (Costanza-Chock, 2020). In practice, it was Zillow’s developers who decided to what extent human experts could influence the Zestimate algorithm after property inspections, thus circumscribing the ways in which humans could make decisions with AI. Perhaps a more careful consideration of making decisions with AI and humans together would have resulted in a different outcome. Almost certainly, more consideration for the implications of almost exclusively AI-driven decisions would have alerted Zillow to more potential challenges.

Finally, it may be hard at first to imagine a scenario where the implications of making decisions with AI are left to AI. However, this is not a far-fetched scenario. Many organizations that do implement AI in decision making do not implement sufficient safeguards and oversight, effectively leaving AI unattended. Sometimes such algorithms could also bring rigidity (Omidvar et al., 2023). This is evident in the case of Zillow, where the algorithm had not been sufficiently adjusted and assessed in time to halt its catastrophic implications. Zestimate was doing what it was trained

to do, and the implications of its decisions had never been included in its training dataset. Organizations that intend to leave the implications to AI itself need to consider, in consequence, decision redistribution towards making decisions about AI and with AI. In other words, if AI is left to decide whether its decisions are good or bad, then human decision making needs to intensify around making decisions about AI and making decisions with AI. Theoretically, this signifies a potential direction for research that traces the implications of decisions made with AI back to those who design and use it.

It is noteworthy that choosing not to implement AI or a specific AI algorithm (e.g., Lebovitz et al., 2022) is also a decision. Within our framework, this scenario would constitute decision retention: maintaining the capability to make decisions among human decision makers without using AI. Chapters in Part III of the book indicate why decision retention may occur. Cases of deciding not to implement AI can be sometimes explained by human biases (Burton et al., 2020), general resistance to change (Markus and Robey, 1988), or specific resistance to technology (Markus, 1981). Yet, our framework highlights another possibility. We argue that organizations may decide not to adopt AI because they may not have a way to account for all the decision redistribution that such adoption may cause. Organizations may not have the resources and skillsets required in the new facets of decision making, or may foresee adverse implications of decision redistribution. In other words, when examined holistically, at times a decision not to implement AI could also be a rational decision, as our framework would suggest.

To summarize, our key argument is that AI in organizational decision making does not simply replace human decision makers, but rather it redistributes them to other facets. We have outlined how this concept can be used productively for research and guide organizations in their adoptions.

THIS BOOK'S PROMISE: WHY SHOULD YOU READ THIS *HANDBOOK*?

While Zillow's example may deter some organizations from even considering AI in decision making, we hope that this *Handbook* can provide valuable guidance and pointers toward a measured approach to AI. Our intention, and the promise of this book, is to showcase the key facets and the underlying issues that need to be considered and navigated in the area of AI and decision making. Collectively, we draw on the discipline of information systems as a way to understand AI and theorize its relationship with decision making. The specific socio-technical angle of the book offers a unique perspective on decision making where AI and human decision makers are not seen as conflicting, but rather considered in relation to each other.

The *Handbook* is targeted at three audience groups. First, the academic scholars. The book is written by researchers with academic rigor. IS as well as organizational scholars will benefit from the book, as it brings together a variety of decisions concerning AI into an integrative framework that advances a more nuanced discussion

around human and AI decision making. The three parts of the book organize research around the types of decisions that are made in relation to AI which can bring novel insights into how human–AI decision making should be configured depending on the area of application of decisions; that is, whether these decisions are made about, with, or by AI. Scholars can also use the book for the purpose of teaching, as discussed next.

Second, the students and practitioners who are concerned with the strategic use of AI. Although written by academics, the *Handbook* is written with practitioners also in mind. It showcases how business managers can make informed decisions about AI in organizations to meet specific business needs. Readers will be able to make better informed decisions about the available AI solutions. The book is equally useful for students of business administration. As such, the book can be used for self-study as well as in university courses on managing AI, decision making, and innovation. The rich diversity of conceptual and case-based chapters will make the book an excellent teaching tool.

Third, the students and practitioners who are concerned with building AI algorithms and systems. The *Handbook* is a valuable source for AI engineers and data scientists who wish to better understand how their data products or insights are used by business customers, and what value they produce to the organization. The book will enable them to develop a holistic view of their role in aiding decision making in organizations. The *Handbook* can be used to deepen the understanding of key organizational issues, and several chapters can be used in discussions and presentations to various organizational stakeholders.

REFERENCES

- Agrawal, A., Gans, J., and Goldfarb, A. (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence* (Illustrated edition). Harvard Business Review Press.
- Agrawal, A., Gans, J., and Goldfarb, A. (2022). *Power and Prediction: The Disruptive Economics of Artificial Intelligence*. Harvard Business Review Press.
- Anthony, C., Bechky, B.A., and Fayard, A.L. (2023). “Collaborating” with AI: taking a system view to explore the future of work. *Organization Science*, 34(5), 1672–1694.
- Berente, N., Gu, B., Recker, J., and Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, 45(3), 1433–1450.
- Boell, S.K. (2017). Information: Fundamental positions and their implications for information systems research, education and practice. *Information and Organization*, 27(1), 1–16.
- Brynjolfsson, E., and McAfee, A. (2017). *Machine, Platform, Crowd: Harnessing our Digital Future*. WW Norton & Company.
- Brynjolfsson, E., and Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science*, 358(6370), 1530–1534.
- Burton, J.W., Stein, M.K., and Jensen, T.B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2), 220–239.
- Costanza-Chock, S. (2020). *Design Justice: Community-Led Practices to Build the Worlds We Need*. MIT Press.
- Davenport, T.H., Barth, P., and Bean, R. (2012). How big data is different. *MIT Sloan Management Review*, 54(1), 43–46.

- Davenport, T.H. and Kirby, J. (2016). *Only Humans Need Apply: Winners and Losers in the Age of Smart Machines*. Harper Business.
- Davenport, T.H., Harris, J.G. and Morison, R. (2010). *Analytics at Work: Smarter Decisions, Better Results*. Harvard Business Press.
- Demetis, D., and Lee, A.S. (2018). When humans using the IT artifact becomes IT using the human artifact. *Journal of the Association for Information Systems*, 19(10), 929–952.
- Feldman, M.S. and March, J.G. (1981). Information in organizations as signal and symbol. *Administrative Science Quarterly*, 26(2), 171–186.
- Galbraith, J.R. (1974). Organization design: an information processing view. *Journal on Applied Analytics*, 4(3), 28–36.
- Gavetti, G. and Levinthal, D. (2000). Looking forward and looking backward: cognitive and experiential search. *Administrative Science Quarterly*, 45(1), 113–137.
- Gavetti, G., Levinthal, D. and Ocasio, W. (2007). Perspective—neo-Carnegie: the Carnegie school's past, present, and reconstructing for the future. *Organization Science*, 18(3), 523–536.
- Huber, G.P. (1990). A theory of the effects of advanced information technologies on organizational design, intelligence, and decision making. *Academy of Management Review*, 15(1), 47–71.
- Joseph, J. and Gaba, V. (2020). Organizational structure, information processing, and decision-making: a retrospective and road map for research. *Academy of Management Annals*, 14(1), 267–302.
- Kahneman, D., Sibony, O. and Sunstein, C.R. (2021). *Noise: A Flaw in Human Judgment*. Hachette UK.
- Lebovitz, S., Levina, N., and Lifshitz-Assaf, H. (2021). Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what. *MIS Quarterly*, 45(3), 1501–1525.
- Lebovitz, S., Lifshitz-Assaf, H. and Levina, N. (2022). To engage or not to engage with AI for critical judgments: how professionals deal with opacity when using AI for medical diagnosis. *Organization Science*, 33(1), 126–148.
- March, J.G. (1962). The business firm as a political coalition. *Journal of Politics*, 24(4), 662–678.
- March, J.G., and Olsen, J.P. (1984). The new institutionalism: organizational factors in political life. *American Political Science Review*, 78(3), 734–749.
- March, J.G., and Simon, H.A. (1958). *Organizations* (2nd edition). Wiley-Blackwell.
- Markus, M.L. (1981). Implementation politics: top management support and user involvement. PhD Thesis.
- Markus, M.L. and Robey, D. (1988). Information technology and organizational change: causal structure in theory and research. *Management Science*, 34(5), 583–598.
- McAfee, A. and Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard Business Review*, 90(10), 60–68.
- McKinney Jr, E.H. and Yoos, C.J. (2010). Information about information: a taxonomy of views. *MIS Quarterly*, 34(2), 329–344.
- Metz, R. (2021). Zillow's home-buying debacle shows how hard it is to use AI to value real estate. *CNN Business*, Nov, 9.
- Miller, A.P. (2018). Want less-biased decisions? Use algorithms. *Harvard Business Review* (July 26). Accessed June 5, 2023, <https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms>.
- Murray, A., Rhymer, J.E.N. and Sirmon, D.G. (2021). Humans and technology: forms of conjoined agency in organizations. *Academy of Management Review*, 46(3), 552–571.
- Omidvar, O., Safavi, M., and Glaser, V.L. (2023). Algorithmic routines and dynamic inertia: how organizations avoid adapting to changes in the environment. *Journal of Management Studies*, 60(2), 313–345.

- Pachidi, S., Berends, H., Faraj, S., and Huysman, M. (2021). Make way for the algorithms: symbolic actions and change in a regime of knowing. *Organization Science*, 32(1), 18–41.
- Pakarinen, P., and Huising, R. (2023). Relational expertise: what machines can't know. *Journal of Management Studies*. doi:10.1111/joms.12915.
- Parker, W., and Putzier, K. (2021). What went wrong with Zillow? A real-estate algorithm derailed its big bet. *Wall Street Journal*, November 17. <https://www.wsj.com/articles/zillow-offers-real-estate-algorithm-homes-ibuyer-11637159261>.
- Pettigrew, A.M. (1973). Occupational specialization as an emergent process. *Sociological Review*, 21(2), 255–278.
- Polli, F. (2019). Using AI to eliminate bias from hiring. *Harvard Business Review* (October 29). Accessed June 5, 2023, <https://hbr.org/2019/10/using-ai-to-eliminate-bias-from-hiring>.
- Raisch, S. and Krakowski, S. (2021). Artificial intelligence and management: the automation–augmentation paradox. *Academy of Management Review*, 46(1), 192–210.
- Schultze, U. (2000). A confessional account of an ethnography about knowledge work. *MIS Quarterly*, 24(1), 3–41.
- Sharma, R., Mithas, S. and Kankanhalli, A. (2014). Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organisations. *European Journal of Information Systems*, 23(4), 433–441.
- Shrestha, Y.R., Ben-Menahem, S.M. and Von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review*, 61(4), 66–83.
- Simon, H.A. (1947). *Administrative Behavior. A Study of Decision-making Processes in Administrative Organization*. Macmillan.
- Simon, H.A. (1969). *The Sciences of the Artificial*. MIT Press.
- Simon, H.A. (1997). *Models of Bounded Rationality: Empirically Grounded Economic Reason*. MIT Press.
- Simonson, I., and Nye, P. (1992). The effect of accountability on susceptibility to decision errors. *Organizational Behavior and Human Decision Processes*, 51(3), 416–446.
- Strich, F., Mayer, A.S. and Fiedler, M. (2021). What do I do in a world of artificial intelligence? Investigating the impact of substitutive decision-making AI systems on employees' professional role identity. *Journal of the Association for Information Systems*, 22(2), 304–324.
- Troncoso, I., Fu, R. Malik, N., and Proserpio, D. (2023, July 24). Algorithm failures and consumers' response: evidence from Zillow. <http://dx.doi.org/10.2139/ssrn.4520172>.
- Turner, K.L., and Makhija, M.V. (2012). The role of individuals in the information processing perspective. *Strategic Management Journal*, 33(6), 661–680.
- Tushman, M.L., and Nadler, D.A. (1978). Information processing as an integrating concept in organizational design. *Academy of Management Review*, 3(3), 613–624.
- Vaast, E., and Pinsonneault, A. (2021). When digital technologies enable and threaten occupational identity: the delicate balancing act of data scientists. *MIS Quarterly*, 45(3), 1087–1112.
- van den Broek, E., Sergeeva, A., and Huysman, M. (2021). When the machine meets the expert: an ethnography of developing AI for hiring. *MIS Quarterly*, 45(3), pp. 1557–1580.
- Varian, H.R. (2010). Computer mediated transactions. *American Economic Review*, 100(2), 1–10.
- Waardenburg, L., Huysman, M., and Sergeeva, A.V. (2022). In the land of the blind, the one-eyed man is king: knowledge brokerage in the age of learning algorithms. *Organization Science*, 33(1), 59–82.

PART I

MAKING DECISIONS ABOUT AI

1. Sourcing data for data-driven applications: foundational questions

Sirkka L. Jarvenpaa

INTRODUCTION

Artificial intelligence (AI) applications require access to large datasets, whether it is spotting tumors in medical images (Aerts, 2018), determining creditworthiness (Strich et al., 2021), or using algorithms for hiring (van den Broek et al., 2022). The statement that “algorithms without data are just a mathematical fiction” (Constantiou and Kallinikos, 2015) is now well accepted; nevertheless, our understanding of decisions that need to be made in sourcing data is limited. Jarvenpaa and Markus (2020, p. 65) define data sourcing as “procuring, licensing, and accessing data (e.g., an ongoing service or one-off project) from an internal or external entity (supplier).” Sourcing decisions affect the data that organizations access, how the problems are framed, and what solutions are sought (Teodorescu et al., 2021). Yet, data sourcing has remained a hidden issue in discussions of AI in organizations (Benbya et al., 2021). Data sourcing is rarely at the forefront, even in studies that focus on data practices or data objects in organizations (e.g., Aaltonen et al., 2021).

Rather, organizations are promised transformational effects from AI without “the need to do major surgery on their IT [information technology] infrastructure or data architecture before they begin” (Fountain et al., 2021, p. 123). Internal data are generated in their own operations (for example, customer data) and combined with data from varied external sources: commercially purchased data, data partnerships and consortia, data drawn from social media sites, and government-released datasets, among others. Data are not only aggregated by brokers, but also accessible from various data infrastructures, such as data marketplaces, data lakes, data pools, and data ecosystems (Abbas et al., 2021; Koutroumpis et al., 2020; Oliveira et al., 2019; Zuboff, 2015). The landscape of data sources and their providers is complex and constantly in flux, with few entry barriers.

In practice, data sourcing is treacherous except in the most basic situations; for example, when data are highly commoditized, structured, and from a known and fixed set of partners (Steinfeld et al., 2011; de Corbiere and Rowe, 2013; Markus, 2016). Organizations procuring external data are depicted as unprepared and as having ad hoc practices (Davenport et al., 2021). Internally, data are often locked in silos, and transferring data across units is difficult because of haphazard or inconsistent data collection and sharing practices (Jarvenpaa and Markus, 2020). Others report how data, although seemingly abundant, are in fact limited (Krasikov et al., 2022). Projects are slowed down and ventures halted because of a lack of access to

appropriate data (Rothe et al., 2019; Sporsem et al., 2021). When data are available, cries often ensue that the data are “incorrect (e.g., outdated), incomplete, biased, or irrelevant” (Polyviou and Zamani, 2022). Sourcing is problematic, and *caveat emptor* fits the context: let the buyer beware.

Others warn of data vulnerabilities. Boyd quotes Bowker: “[R]aw data is both an oxymoron and a bad idea. Data should be cooked with care.” Boyd (2020, p. 259) interprets Bowker’s quote: “within the context of AI, we need to talk about what that data is, what it looks like, where it comes from, and what the nuances are. We need to tease out these issues in a sensible way so that we can better understand what makes data legitimate.”

There are many complexities in determining data’s fit (Markus, 2001; Janssen et al., 2012): it requires learning from and about the data. What experiences do the data relate to that are relevant? Data sourcing involves connecting data to the organizational context. Much of data’s representation and meaning are lost when sourcing moves data across contexts (Eirich and Fischer-Pressler, 2022; Jussupow et al., 2021; Lebovitz et al., 2021).

This chapter argues that data sourcing is a strategic issue that requires an organizational learning perspective. Organizational learning influences the data experience, as well as the choice of problems and decisions that organizations pursue with AI. In data sourcing, considering organizational learning belongs at the outset, not as an outcome of AI applications or as a parallel process to see how humans augment machine expertise (Sturm et al., 2021; Teodorescu et al., 2021; van den Broek et al., 2022). Decisions on data sourcing require a constant inquiry into organizational learning and its context: what data can help organization members to learn, with what tasks, and using what tools?

To structure the discussion on data sourcing decisions, I turn to the IS sourcing literature. In the 30-year history of IS sourcing literature, research has rallied around foundational decisions in sourcing: making the decision to source data, designing the arrangements for it, and managing these arrangements. This chapter reviews these decisions and relates them to current issues in data sourcing. Data sourcing decisions call attention to the organizational contexts in which organizations can learn from and with data. I use behavioral learning theory to structure the discussion of organizational learning contexts and data sourcing. I conclude with directions for future research.

KEY ISSUES IN SOURCING DECISIONS

A seminal event in the information systems (IS) outsourcing literature was Kodak Corporation’s decision in late 1988 to hand over its IS functions to several sourcing providers (Pearlson et al., 1994). One of the goals of the outsourcing in the Kodak case was data integration. But subsequent research on IS sourcing focused on cost savings rather than on data integration. After all, data sourcing had long been viewed as “the routine manipulation, storage, and transfer of symbolic information within

established categories” (Azoulay, 2004, p. 1591). Recent studies have begun to recognize the complexities in data sourcing. Chen et al. (2017) addressed concerns that emerge when the client and vendors are co-owners of the database in a software development outsourcing project. In addition to data ownership issues, data sourcing comes into play in discussions of privacy and security, such as knowledge leaks (Jarvenpaa and Majchrzak, 2016). In the fifth edition of *Information Systems Outsourcing*, Hirschheim et al. (2020) categorize data sourcing and data partnerships as “emergent sourcing challenges.”

IS sourcing has been defined as “the contracting or delegating of IS- or IT-related work (e.g., an ongoing service or one-off project) to an internal or external entity (a supplier)” (Kotlarsky et al., 2018, p. 1). Kotlarsky et al. (2018) identified three primary clusters in the IS sourcing literature: (1) making the sourcing decision; (2) designing contractual structures; and (3) managing the sourcing relationship. Other reviews have paralleled this three-part structure (Gambal et al., 2022; Hanafizadeh and Zareravasan, 2020; Lacity et al., 2017). These three decisions help to organize the data sourcing conversation in the literature. “Buyer” refers generically to an organization, group, or individual that is procuring resources.

Making the Sourcing Decision

In IS sourcing, the foundational decisions are about delegation of activities (make or buy), and who and what should be delegated. In data sourcing, the foundational decisions focus less on delegation and more on what the organization can learn with the sourced data, and what is required for learning to take place.

IS sourcing decision

The first decisions in IS sourcing focus on questions of whether to source, and if so, from whom to source. In this realm, questions of governance, ownership, and location must be considered. How these sourcing decisions play out depends on the underlying motivation, transaction attributes, and vendor and buyer attributes (Kotlarsky et al., 2018; Lacity et al., 2017; Mani et al., 2010). Buyers’ motivations may involve commercial and/or societal concerns, such as impact sourcing (Carmel et al., 2016). Transaction attributes include costs, service standards, and service complexity, among others; vendor attributes include the supplier’s domain and technical expertise; and buyer attributes include prior experience with outsourcing relationships and technical experience relative to the supplier. The IS sourcing literature assumes the availability of information about or knowledge of these attributes *ex ante*.

The heterogeneity of the transactions, relationships, and tasks in sourcing decisions is significant with varied risks and benefits. Much of the related literature has focused on governance of these risks and benefits from a perspective of either transactional control or relational trust (Kotlarsky et al., 2018). Transactional logic focuses on control mechanisms (for example, in contracting, investing, monitoring, or enforcing), and on the strategies for managing expectations and sanctions that are specified *ex ante*. Meanwhile, relational logic focuses on mechanisms for commu-

nication (for example, openness, explanations), information sharing, demonstrations of goodwill, and procedural consistencies that can help to manage uncertainties that cannot be known at the outset. Theoretical perspectives have been drawn primarily from transaction cost economics (Williamson, 1985), a resource-based view of the firm (Kogut and Zander, 1992), and psychological and sociological perspectives on trust (Kramer and Tyler, 1996).

Data sourcing decision

In the initial data sourcing decision, organizations determine what data they can access and use; either internally or externally. Motivations for sourcing the data are broad: beyond commercial organizational performance, firms may seek scientific breakthroughs or have broader collective and societal purposes in mind. Local and state authorities might source private data for purposes such as crisis preparedness or management. Having broad and flexible motivations leaves room for clarity to emerge as firms learn from the sourced data. Hence, rather than specifying motivations and controls *ex ante*, as in much of IS sourcing literature, data sourcing literature allows motivations to be shaped by data.

Except in cases of highly commoditized data, such as financial market data or click-stream marketing data, buyers may not know transaction attributes or be able to assess quality when they make sourcing decisions. Martins et al. (2017, p. 1) note that “in data marketplaces, customers have little knowledge on the actual provided data.” Thus, data quality cannot be assessed or verified, independently of use. Many providers hesitate to share data before a purchase (Stahl et al., 2017). Kennedy et al. (2022, p. 33) state that “[b]uyers ... cannot directly observe data without payment.” Hence, buyers face much uncertainty, not only about what data is available, but also in whether the data can be used, by whom in the organization, and for how long.

Some data providers offer summaries of metadata (annotated documentation and schema), but the unique and differentiated nature of data offerings can make assessments and comparisons across datasets difficult (Stahl et al., 2017). Assessing data quality and data fit is impeded when buyers do not know how a provider created or obtained the data, or what types of transactions and transformations may have been involved (Thomas et al., 2022). Like many ethnographers, Boyd (2020) has concerns over the effects of data acquisition on data quality; coerced data is different, both in quality and in reliability, from voluntarily provided data.

Buyers that have access to trusted intermediaries may be able to exchange information about transaction attributes, perceived quality, and costs (Perkmann and Schildt, 2015). However, whether the data have these attributes and quality at reasonable cost also relates to the users in the organization, the tasks for which the data will be used, and the tools needed for that use. Importantly, many of the antecedents to IS sourcing decisions (for example, quality) are unknowable until the organization has sourced and put the data to use. In addition, buyers may lack the necessary domain knowledge, data analysis skills, or tools to process and analyze data for relevant tasks.

Designing Sourcing Arrangements

The second cluster of sourcing decisions identified by Kotlarsky et al. (2018) formalizes the incentives and the division both of responsibilities and of risks in the outsourcing governance. In IS sourcing, the governance is interorganizational: between the provider and the buyer. In data sourcing, the governance includes relationships between the data provider and the buyer. The buyer may be external, internal, or both. In addition, strategic risks from data sourcing often are internal, arising from fit between the data and tasks or decisions in the organization.

IS sourcing arrangements

Either a control (transactional) perspective or a trust (relational) perspective, or a combination of both, becomes the premise for designing formal (legal) contracts in IS sourcing. Contracts most often are bilateral, even in a multi-sourcing context. Importantly, the contracts are negotiated with and tailored to the buyer based on the specific services provided. They define both pricing and the *ex ante* control points, such as quality standards, penalties for the vendor's failing to perform, and termination rules. Contracts also include the ownership stipulations and are restricted to certain time periods and geographies (Kotlarsky et al., 2018; Lacity et al., 2017).

Data sourcing arrangements

The data governance models may involve transactional, relational, network, or commons sourcing arrangements (e.g., van den Broek and van Veenstra, 2018; Azkan et al., 2020; Bergman et al., 2022; Gelhaar et al., 2021). Transactional arrangements involve data markets that trade data on digital platforms (for example, Amazon Web Services) and data brokerages that aggregate data from different sources and then resell it (for example, Acxiom). Relational arrangements involve data partnerships and data bartering, while network arrangements involve data consortia. Commons include data pools (for example, government-released open data). Pools and consortia are often based on industry verticals (for example, insurance, maritime) that are used to manage shared industry risks, which may be of a regulatory, commercial, or some other nature (Koutroumpis et al., 2020). The same datasets can be governed by both formal and informal agreements.

In the transactional governance model, data providers dictate the arrangements and contingencies for buyers (Kennedy et al., 2022). We know little about these data contracts, apart from practitioner reports indicating that contracts are standardized regardless of the buyer, or that they are ad hoc or informal. In general, the buyer has little recourse if it receives poor-quality data or if delivery of data is incomplete, even when the transaction is covered by a bilateral contract.

Relational governance can involve reciprocal data-sharing commitments and joint work products (for example, a joint research paper) (Jarvenpaa and Markus, 2020). Relational and network governance also may include an obligation to return the data, perhaps with any enhancements or with reports of key findings (Jarvenpaa and Markus, 2018). Buyers might need to report their intentions before they can access

the data. In relational and network governance modes, arrangements can be rather informal. For example, they may be based on interpersonal agreements, professional accreditation, or an individual or organizational membership in a professional association (van den Broek and van Veenstra, 2018). Sporseem and Tkalich (2020, p. 2) astutely observed that none of the three modes of data governance described by van den Broek and van Veenstra (2018) (that is, bazaar, hierarchy, and network) involved organizations “buying or selling through regular contracts.” The lack of formal (legal) contracts is not surprising, given the buyer’s inability to verify transaction attributes and address many unknowns about the data. Only weak protections attach to data, whether for creators or users. Data “ownership” may be ambiguous or nonexistent. Data access and use rights are relevant, but the rights can be transient as data move across contexts in sourcing. The data may come from individuals who have requested that it be eliminated or “forgotten” at a certain point. As data are reused, forwarded, and resold, the rights and obligations to the data (and any ownership claims) become distributed and can be hard to enforce (Leonelli, 2015).

Data sourcing arrangements are asymmetrical, so that buyers must beware. Transactional data arrangements suffer from quality problems and a lack of pricing transparency. Although the data providers set the rules for access and use, the rules vary for different buyers, depending on different data sensitivities. Restrictions can vary based on the data’s form of access (for example, cloud, on-premises), which buyers can have access, and the types of intended tasks and how they are related to either public or private interests (Lee, 2017). Similarly, the tools for analyzing the data may be limited and dictated by the data provider. Buyers may be able only to send data queries, and to get the returned results but not to explore the data more broadly (Gainer et al., 2016; Jarvenpaa and Markus, 2020). In desiring access to data, buyers have to concede to data providers’ governance arrangements.

Managing Sourcing Arrangements

The third cluster of sourcing decision identified by Kotlarsky et al. (2018) builds on the two prior decisions and addresses uncertainties that emerge over time and that cannot be managed by governance arrangements.

Managing IS sourcing

As in the two previous clusters, the perspective in managing IS sourcing, in terms of how buyers and providers interact on a daily basis, is either transactional (control) or relational (trust). Both interaction modes assume that some goal conflict always remains between the provider and the buyer. In relational outsourcing deals, such as business process and strategic innovation outsourcing, contracts can be at a high level; they may lack well-defined service-level agreements and quality standards, thus necessitating frequent (and sometimes even daily) interactions and reciprocal feedback to comply with and enforce the contract. Transactional outsourcing includes penalty clauses for violations of the contract. Managing IS sourcing arrangements also involves issues related to contract termination, switching providers, and

back-sourcing. Although organizations may use multiple IS providers, the quantity of providers generally is limited (Chatterjee, 2017).

Managing data sourcing

Whether data are sourced for one-time use or for continuing use can affect uncertainties to be managed, particularly those related to power balances in the buyer–supplier relationship. Data arrangements may become nonviable if industry norms or societal regulations on data access and use rights change (Winter and Davidson, 2019). For example, data regulations may restrict the movement of certain data or may demand that data be available or deleted within a certain timespan.

Even if the data are available and fit for the intended purpose, buyers may not have the necessary skills and competencies to learn from them. The appropriate data scientists may not be available, or domain experts may be too busy or lack incentives to attend to sourced data (Teodorescu and Yao, 2021). Data acquisition can lead to intensive data processing and manipulation; even involving daily interactions between data scientists and internal (or external) domain experts to learn what the organization can gain from the data.

When buyers source data from many providers (for example, as many as dozens or hundreds), interactions are multiplied, and the learning challenges become even more formidable. For example, buyers of genome data can source from more than 30 000 genome data providers (Contreras and Knoppers, 2018). These providers are highly heterogeneous, ranging from individuals to trade associations, from commercial start-ups to large multinational companies to governments. Buying organizations face difficulties in interfacing with so many external providers. Practitioners recommend that buyers establish a cross-functional group to interface with these external providers (e.g., Aaser and McElhaney, 2021). However, examples of such a data management strategy are difficult to find. In one pharmaceutical company, each research group and product group managed its own data providers, without any coordinated effort (Jarvenpaa and Markus, 2020). Integrating external and internal data and learning from it also present many challenges. Rothe et al. (2019) highlight that external data—particularly from open data pools—rarely produce data services or commodities that can be commercially appropriated without internal data.

In summary, how, when, and what organizations can learn from sourced data—whether by humans or machines, or in combination—depends on what data are sourced, under what arrangements, and what data management arrangements are established. The questions call attention to organizational learning needs.

DATA SOURCING FROM AN ORGANIZATIONAL LEARNING PERSPECTIVE

Argote and Miron-Spektor (2011) offer a pragmatic organizational learning framework that centers on organizational experience and the context of this experience. This behavioral learning framework “parse[s] organizational learning to make it more

tractable analytically” (Argote and Miron-Spektor, 2011, p. 1124). The framework maintains that “organizations learn from experience, not from knowledge” (Argote and Todorova, 2007, p. 220), and that this experience takes place in the context of task performance. Such experience has cognitive and informational components as well as motivational and social components. The outcome of organizational experience is knowledge creation, retention, and transfer; without experiential learning, organizations are unable to produce, retain, and transfer knowledge (Argote, 2013). Thus, experiential learning in organizations affects the repertoire of available actions (Argote, 1999; Maula et al., 2023), and leads to changes in employees’ cognition, routines, and practices, expanding the range of organizations’ potential or actual performance (Argote and Todorova, 2007; Huber, 1991).

The foundational tenet of this framework is that learning is contextual. Organizational contexts have both active and latent qualities, as does the broader environment. The active learning context comprises learners, tools, and tasks; the learners perform tasks that generate organizational experience that produces knowledge. The experience gained comes not only from the direct performance of tasks, but also from interactions with others, resulting in indirect (that is, vicarious) learning and knowledge transfer (Argote and Ingram, 2000).

The latent or higher-order organizational context “affects which individuals are members of the organizations, what tools they have, and which tasks they perform” (Argote and Miron-Spektor, 2011, p. 1125). The latent context is critical to organizational learning; the past history of task completion (that is, by individuals and groups), and how this experience accumulates into routines or heuristics, matter (Maula et al., 2023). This context comprises macro-concepts, such as organizational culture, structure, and strategy. The difference between the overt and latent contexts is their capacity for action. Learners and tools perform tasks: they do things. In contrast, “the latent context is not capable of action” (Argote and Miron-Spektor, 2011, p. 1125). Beyond the active and latent organizational contexts is the environment: the broader context comprising competitors, regulators, clients, and other institutions.

Learning Experience in Active, Latent, and External Contexts

Questions about how companies can manage their active, latent, and external contexts, and operate effectively as these environments change, demand sufficient attention to data sourcing through an organizational learning lens. Task performance occurs in an active context that produces learning experience. Learning experience is multidimensional, incorporating the context of both the direct and the indirect task performance, the content of the tasks and relationships, and both spatial and temporal elements.

Direct or indirect organizational experience

Learning experience is multidimensional, incorporating the context of both the direct and indirect task performance, the content of the tasks and relationships, and both spatial and temporal elements. Direct experience involves learners’ performance

and completion of the task for the organization. Indirect experience comes from observing others' experience or delegating the task performance to a machine. When an algorithm automates task performance, individuals and groups in the organization learn indirectly because they do not engage in the task; instead, they see the outcome of the technology's task performance. Whether this indirect experience is meaningful and produces experiential knowledge depends on whether potential learners reflect on the outcome or results (Argote and Todorova, 2007; Maula et al., 2023). Perceived relationships between the data and outcomes or results affect the meaning perceived and the knowledge constructed. Direct experience can maximize variance-seeking behaviors and lead to an improved, often more complex understanding.

Nature of task and relationships

The second dimension of the framework involves the content of the learning experience. Tasks and relationships produce different learning experiences based on their novelty (newness), heterogeneity (across repetitions), and ambiguity (interpretive potential). To illustrate, novelty might pertain to the actions for performing the task or to the task itself. Heterogeneity of a task is about how it changes: is the same or a similar task repeated over time? Although homogeneity can be beneficial for novices or at the beginning of task performance, heterogeneity (that is, change) is needed for learning. Such learning expands knowledge, decreasing learners' "uncertainty about the true relationship between cause and effects" (Argote and Todorova, 2007, p. 1126). As uncertainty or ambiguity increases in task performance, learning becomes more challenging, as learners have to reflect on and interpret the experience and its meaning.

Space of learning experience

The third dimension pertains to spatial elements, such as how geographically concentrated or dispersed the learning experience is. The spatial dimension may influence learners' level of motivation: how relevant is the learning to their context? What are the communication and interpretation challenges across contexts? Can learners develop mutual understandings or find "common ground" in different contexts (Argote and Todorova, 2007)? These contextualization issues become paramount as data move into new contexts (Rothe et al., 2019).

Temporality of learning experience

The fourth dimension of the organizational learning framework pertains to time and temporality, and captures whether learning happens before action (learning before doing), during action (learning while doing), or after action (reflecting on action). This dimension relates to the social experience of learning: how much attention have organizational leaders allocated to problems and solutions, and which organizational members have also been expected to attend to them? Argote and Miron-Spektor (2011) generally note various obstacles to learning from the past; particularly as contexts change, when prior experience has decayed, or knowledge is lost because individuals with direct experience of the tasks have left. Learning before doing,

such as with research and experimentation, can be effective; such learning generally requires consensus on and understanding of relationships (Pisano, 1994). When such relationships are unclear, task performance requires learning while doing. Learning after doing (that is, reflexivity and review) needs to incorporate learning from both successful and less successful cases (Ellis and Davidi, 2005).

Data Sourcing Questions from Organizational Learning Perspective

There are benefits of taking a fine-grained approach to learning experiences (Argote and Miron-Spektor, 2011). Such benefits include identifying experiences that have positive effects and negative effects on organizational processes and outcomes (see March, 2010); understanding the relationships among different types of learning experiences; and designing intentional experiences to promote effective and meaningful organizational learning. The organizational learning framework supports the following foundational questions: (1) What data can the organization access and use, for what tasks, with what members, and with what tools? (2) How do data sourcing decisions frame organizational problems and change decisions? and (3) How do organizations interface with data providers, and how do these interactions influence the interdependence of the organizations and their broader environment?

In sourcing data, what data can the organization access and use, by whom, for what tasks, and with what tools?

Organizations may not be able or allowed to access the data they want to pursue. Even if they can, they still need to determine the data's fit for tasks, for learners, and for the available tools. The combination of tasks, learners, and tools, in turn, determines the nature of the experience and the knowledge the data can generate.

This first question relates to the active learning context in the framework of Argote and Miron-Spektor (2011). In IS sourcing, Lacity et al. (2017) note that "a client is more likely to experience better sourcing outcomes when they themselves have mastered the technologies and processes associated with providing the service." The lack of experience complicates the buyer's control and trust mechanisms in the IS sourcing decisions. In less complex situations, learning might be delegated to outsourcing advisors, who aggregate and transfer the learning across the outsourcing relationships between different buyers and providers (Poppo and Lacity, 2002). But such advisors may not be appropriate in business process or strategic innovation sourcing, where contexts can vary dramatically, as does the learning they require. Similarly, in data sourcing, context is crucial. Complex settings can create learning experiences that produce tacit knowledge and that make transferring difficult via arm's-length relationships (Argote et al., 2003).

In the design of sourcing arrangements, how do data sourcing decisions frame problems and shape decisions?

This second question relates to the latent organizational context discussed in the framework of Argote and Miron-Spektor (2011). Organizations require data to make

data-driven decisions, and changes in available data not only may change the decisions to be made, but also may affect organizational strategy. In addition, changes in collaborations or teams (for example, for data sourcing purposes) can affect organizational structure and even identity. Such structural changes influence “which individuals are members of the organizations, what tools they have, and which tasks they perform” (Argote and Miron-Spektor, 2011, p. 1125).

In managing data sourcing arrangements, how do organizations interface with data providers, and how do these interactions influence the interdependencies of the organization with its broader environment?

The third question maps to the broader environmental context in the organizational learning framework of Argote and Miron-Spektor (2011). With external data sourcing, organizations might be exposed to manipulated or biased data, and they may not have generated the necessary experience and the learning to produce knowledge about the attendant risks (De-Arteaga et al., 2021; Teodorescu et al., 2021). A data source may become the target of a data scandal (Aaen et al., 2022). To illustrate, Monteiro and Parmiggiani (2019) describe how data are used as a political device: organizations regulate how others can leverage their open datasets, and by doing so, they ensure that data are used for their desired political purposes.

DATA SOURCING AND ORGANIZATIONAL EXPERIENCE: SOME EXAMPLES

I offer three examples to illustrate how the framework can be useful for analyzing an organization’s learning experience in sourcing data. The examples are based on published cases that are reinterpreted here through the behavioral learning framework. The first example focuses on how challenges related to data access and use changed what could be learned with data, and how the algorithm was developed. The second example highlights the incentive challenges that can arise, not just in interorganizational relationships, but also internally. The third example reveals issues related to power and ethics that can challenge continued use of data.

Maritime Trade (Grønsund and Aanestad, 2020)

The case chronicled and analyzed the development, introduction, and post-introduction of an algorithm after the algorithm achieved satisfactory accuracy in a ship brokerage organization. The algorithm automated the maritime trade table generation, predicting seaborne trade flows and arbitrage opportunities that the company then sold as intelligence to its customers, via analysts who were domain experts but also users of the trade table. The company’s motivations involved competitive desires to remove delays caused by human-generated forecasts, thus improving arbitrage opportunities. (The forecasts leveraged temporary changes in the prices of commodities.)

Data were externally sourced because they were “openly available and accessible” (Grønsund and Aanestad, 2020, p. 3):

data source ... is the Automatic Identification System (AIS), a global, standardized communication system enabling exchange of data on ships' navigational status and voyages ... The default transmission rate is every few seconds, and the message includes automatically updated (that is, dynamic) data on ships' identity, position coordinates, course, and speed. In addition, the message carries manually entered (and more static) data, such as cargo, point of departure, destination, and estimated time of arrival.

This widely and standardized data produced major challenges in the development of the algorithm, despite the fact that the same data source was already used to develop timetables manually by the domain expert (most recently, using Excel spreadsheets). Moreover, in the development process, the data scientist had access to the cumulative experiences of the domain expert. In terms of the active learning context, members included the domain expert (“She knows almost all ships and ports inside and out”, p. 6); a data scientist; users (for example, analysts who used the researcher's outputs); and later, a data analyst. The tools involved various online tracking systems, rule bases, and a Hadoop-based data warehouse repository. During the development of the algorithm, the researcher continued ongoing manual generation of the trade table, which came to represent the reference against which the algorithmic trade table was evaluated. The tasks of the data scientist were to audit and detect anomalies between the researcher's manually generated timetable and the algorithm, and to alter the algorithm accordingly. Later, when the data analyst was hired, the analyst provided feedback to the data scientist on the performance of the algorithm.

Although the data source was the same as it had been in the manual processes, the company transacted with the service provider to increase the frequency of AIS data—from daily updates to hourly updates—to provide a more fine-grained view. This finer-grained view was seen as critical in creating value-added services for customers. Yet, these high-frequency AIS data were seen as “messy, erroneous, missing, or obfuscated,” and “the problem was not resolvable as the data sources were not under ShipCo's control” (p. 11). Problems resulted from faulty transmitters, noise, and errors, as well as from purposeful hiding or misrepresentation of the whereabouts of the ships to avoid pirates. The data service provider was unable to offer help.

The company was unprepared. It had expected the data gathering and processing to be automatic, based on well-understood rules. Hence, the plans had to be modified, and involved joint learning meetings between the domain expert, the data scientist, and the users of the trade table. The number and range of people in these spatially proximal meetings initially increased rather than decreased. The algorithm development demanded significant synchronous learning-by-doing through joint meetings. Hence, the right members in the organization had to be incentivized and engaged to address anomalies and ensure adequate timeliness and accuracy of data.

As the algorithm was expanded to new segments (for example, regions and markets of different commodities, such as from oil to gas), the task changed. The data acquisition and processing had to be reworked again, involving joint meetings between

the domain expert, data scientist, and users. The different patterns of behavior of oil and gas tankers had to be reflected in the algorithm. In sum, despite the similarities in task outputs of the manual process and the algorithm, and the team's familiarity with the AIS data, there were many unanticipated problems that challenged organizational learning from the data. Ongoing interactions in an active learning context were required between the data scientist, the domain expert, and the users of the timetable.

The higher-level organization (the latent context) and the external environment were also affecting and affected by the project. The algorithm development introduced new members (for example, the data analyst), new tools (algorithmic system), and new tasks (auditing and altering the algorithm) to the organization. The broader environment was becoming increasingly competitive as customers became competitors. Some customers had started to source the AIS data and invest in research and data science themselves. To increase the value of its services, the shipbroker's future plans called for data integration of external data with internal data. The strategy was contingent on first "building a culture of learning and experimentation." The new strategy depended on forging new partnerships internally and externally.

Data Labeling at an Automotive Company (Eirich and Fischer-Pressler, 2022)

The maritime trade table example highlights the continuous involvement of a domain expert and users of the timetables internally to ensure the fit and legitimacy of the "openly available and accessible" data. This involvement cannot be taken for granted.

Data label work in an automotive company (Eirich and Fischer-Pressler, 2022) highlights the intense interaction between data scientists and domain experts in an active learning context when the task itself is novel, heterogeneous, and ambiguous. A large automobile company embarked on an effort to digitize manufacturing processes, and to develop data-driven tools to improve new product line production (electric motors). Here, the data were internally sourced and required labeling by experts to render the data useful. With considerable fine-grained analysis, the case examines the company's data label work, which required developing tools that would automatically label data for new machine learning algorithms to assist engineers in optimizing manufacturing processes. Data labels allow for "attaching a certain attribute to an instance in a data set." In fact, the data labels themselves represented aggregated expert knowledge of the engineers, who were the domain experts. The company had no means for internal or external validation to ensure accurate labeling. The sourcing of data labels required close interaction and much dialog between data scientists and domain experts in the active learning context. The authors write that "[d]uring the process of providing labels, domain experts gain a better understanding of the data they analyze and thus create new knowledge, which can be externalized in the form of labels." The data—and hence the data labels—varied across different engine parts and different versions of the same engine parts. The high dynamicity of data further complicated the data label work.

At the latent organizational level, the novelty of the data work produced challenges. These challenges resulted from nonaligned incentives and a lack of relevant

domain knowledge about the high-volume production of electric engines. They also resulted from mistrust in the data work itself; mistrust that was attributed at least partly to differences in the interpretations and belief systems of the data scientists and the engineers. The stochastic models of the data scientists were questioned by the engineers, who were accustomed to deterministic ways of thinking. Differences also emerged in the level of ambiguity tolerated by the data scientists and the engineers. Although these alignment issues have been discussed in the literature on algorithm design (Martin, 2019; Morse et al., 2021), how nonaligned incentives affect the sourcing of data for algorithms has received little attention.

In contrast to the case of the maritime trade tables, the data label work at the automobile manufacturer represented a novel task. Relevant domain knowledge may not have been available or appropriable. Eirich and Fischer-Pressler (2022) underscored the importance of human domain knowledge data pipelines. Humans are needed to be willing human trainers to render technologies that can act as “lifelong learners” (Agrawal et al., 2019). These trainers are expected to maintain a learning experience of “subject matter experts” who continue to be “coaches who guide” or “laboratory scientists who experiment” (Seidel et al., 2019, p. 57). Hence, the challenges are broader than aligning incentives in appropriate relationships.

Danish General Practitioners’ Database (Aaen et al., 2022)

For my third example, an interorganizational data-driven system highlights the broader environmental context of power, politics, and ethics in continued sourcing of the data. As new actors join, the tasks and tools change, as do the risks from sourcing the data. The case chronicles a study of the Danish General Practitioners’ Database (DAMD), a multiyear and multimillion dollar project initially established by and set up to support general practitioners (GPs) themselves. The database was to contain data on one disease that the GPs themselves treated, but over time it grew to cover more than 700 diseases, and extended its reach of users to a broad range of stakeholders, including pharmaceutical companies and various public data authorities. At the end of the case narrative, the same database was described as “*golden egg, indispensable, and a game changer*,” as well as “*corrupted, toxic, and illegal*” (emphasis in the original). The authors raised the question: “How can something as innocent as data (everlasting, persistent, and agnostic), used with the very purpose of saving lives, turn into something toxic?” Any potential opportunities to learn with and from the data ended abruptly when the database was legally directed to be destroyed. The uses had expanded to areas that violated the country’s laws.

Originally, GPs collected data on their own patients in the context of care. GPs received advisory summary reports on these patients’ data. The motivation was to support practitioners in relation to patients who failed to seek treatment, or who were suboptimally treated, and hence to “improve quality in general practice” (p. 295). The database was envisioned as providing a more holistic picture of the patients, and particularly of more complex or chronic disease cases.

As the database grew in terms of its functions and its disease categories, the data were repurposed for research projects. This purpose required increased attention to data collection quality across GPs, which led to standardized coding. GPs were offered coding lessons and mentoring by other users to ensure that each patient contact was diagnosed and recorded in a consistent way. In addition, the data collection became increasingly automated for the benefit of the “centralized datahub.” What had been a voluntary data-sharing decision by GPs became mandated; now, users included professionals who were not directly treating the patients from whom data were collected.

The data users included patients themselves, for the purpose of giving them more “new action possibilities” via self-care. The case study quotes the Association of Danish Patients: “No one has died from the misuse of data. But it is possible that you can die if data sharing does not occur.” As the use of the database expanded, there was less concern about whether GPs who had initiated the data collection were maximizing their learning. Aaen et al. (2022) suggest that “the function of delivering data to health care professionals was only partially implemented” (p. 13).

As the project swelled in terms of its functions, stakeholders, and data, other parties joined in, with the goal of using the data to control the data providers; that is, the GPs. The purpose of the system had shifted from providing after-action learning for GPs, to research, and then to health administration. Because the data were no longer being collected just for care delivery, the legality of the data collection began to be questioned, including by the GPs. The Danish Data Authority instructed the large, nationwide Danish health database to erase the “already collected data,” although “the Danish National Archives determin[ed] that it was worth preserving and archiving” (p. 291). The Parliament intervened, and the database was deleted. As noted, the authors raised the question: “How can something as innocent as data (everlasting, persistent, and agnostic), used with the very purpose of saving lives, turn into something toxic?”

The case highlights how the active learning context, the latent organizational contexts, and the broader environment change the availability of and perspectives on data. As the number of stakeholders grew, data were sourced for different organizational experiences. As the case illustrates, these new organizational experiences not only undermined but also conflicted with the original purpose, thus also undermining the legitimacy and fit of the data for the original power holders (that is, the GPs). New interested audiences bring along their own power hierarchies (that is, Parliament). Zuboff (2015, p. 76) writes that as data extraction is automated and data are used to inform, “new contexts” lead to “new opportunities for learning and therefore new contests over who would learn, how, and what.” Thus, what Zuboff labels the “division of learning” becomes a potent and complex issue in sourcing data.

CONCLUSION AND FUTURE RESEARCH

This chapter has examined data sourcing for data-driven applications in organizations. I sought to build on the strengths of the well-established IS sourcing literature by reviewing its key decisions and relating them to data sourcing, thus to further develop and contribute to a cumulative research base. I discerned specific questions related to data sourcing based on the general IS sourcing questions about making a sourcing decision, designing sourcing arrangements, and managing sourcing arrangements.

I argued that an organizational learning perspective can serve as a theoretical perspective to offer insight into data sourcing. The specific questions from this perspective highlight how organizations learn from data. I relied on a behavioral learning framework (Argote and Miron-Spector, 2011) to highlight how an organization produces and accumulates organizational experience with data through task performance and how the context of this experience matters. Three brief examples illustrated how active and latent organizational learning contexts, as well as the broader external environment, affect data acquisition and processing and, thereby, problems and decisions that organizations pursue.

In light of the current work, future research is encouraged on the different elements and dimensions of the behavioral learning framework. The work can advance testable propositions about the mechanisms and contingencies of the learning experience, based on whether data are sourced internally or externally in relation to data-driven applications. Future work also needs to be attentive to new and changing dimensions of organizational experiences with data.

Clearly, contexts matter in data sourcing decisions. Contexts change perceptions of what the data are, how the data might be used, and for what purpose. They change the meaning, the experiences, and the knowledge that the data produce (Aaltonen and Penttinen, 2021; Rothe et al., 2019; Tuomi, 1999). Methodologically, longitudinal studies are needed because learning is a process that happens over time (March, 1991). “Learning begins with experience” (Argote and Miron-Spektor, 2011, p. 1126). Human experience with learning is very different from that of machines (Balasubramanian et al., 2022). Human experience has cognitive, social, and motivational elements. This chapter is a call for research into how organizational learning experiences are shaped by and shape data sourcing decisions, design of arrangements, and their management.

REFERENCES

- Aaen, J., Nielsen, J.A., and Carugati, A. (2022). The dark side of data ecosystems: a longitudinal study of the DAMD project. *European Journal of Information Systems*, 31(3), 288–312.
- Aaltonen, A., Alaimo, C., and Kallinikos, J. (2021). The making of data commodities: data analytics as an embedded process. *Journal of Management Information Systems*, 38(2), 401–429.

- Aaltonen, A., and Penttinen, E. (2021). What makes data possible? A sociotechnical view on structured data innovations. *Proceedings of the 54th Hawaii International Conference on System Sciences*.
- Aaser, M., and McElhaney, D. (2021). Harnessing the power of external data. McKinsey & Co. www.mckinsey.com/business-functions/mckinsey-digital/our.
- Abbas, A.E., Agahari, W., Van de Ven, M., Zuiderwijk, A., and De Reuver, M. (2021). Business data sharing through data marketplaces: a systematic literature review. *Journal of Theoretical and Applied Electronic Commerce Research*, 16(7), 3321–3339.
- Aerts, H.J. (2018). Data science in radiology: a path forward. *Clinical Cancer Research*, 24(3), 532–534.
- Agrawal, A., Gans, J., and Goldfarb, A. (2019). Prediction, judgment, and complexity: a theory of decision-making and artificial intelligence. In Agrawal, A., Gans, J., and Goldfarb, A. (eds), *The Economics of Artificial Intelligence* (pp. 89–114). University of Chicago Press. <https://doi.org/10.7208/9780226613475-005>
- Argote, L. (1999). Organizational learning: creating. *Retaining and Transferring*, 25, 45–58.
- Argote, L. (2013). *Organizational Learning: Creating, Retaining and Transferring Knowledge*, 2nd edn. New York: Springer.
- Argote, L., and Ingram, P. (2000). Knowledge transfer: a basis for competitive advantage in firms. *Organ. Behav. Human Decision Processes*, 82(1), 150–169.
- Argote, L., McEvily, B., and Reagans, R. (2003). Introduction to the special issue on managing knowledge in organizations: creating, retaining, and transferring knowledge. *Management Science*, 49(4), v–viii.
- Argote, L., and Miron-Spektor, E. (2011). Organizational learning: from experience to knowledge. *Organization Science*, 22(5), 1123–1137.
- Argote, L., and Todorova, G. (2007). Organizational learning. *International Review of Industrial and Organizational Psychology*, 22, 193.
- Azkan, C., Iggena, L., Gür, I., Möller, F.O., and Otto, B. (2020). A taxonomy for data-driven services in manufacturing industries. In *PACIS 2020 Proceedings*.
- Azoulay, P. (2004). Capturing knowledge within and across firm boundaries: evidence from clinical development. *American Economic Review*, 94(5), 1591–1612.
- Balasubramanian, N., Ye, Y., and Xu, M. (2022). Substituting human decision-making with machine learning: Implications for organizational learning. *Academy of Management Review*, 47(3), 448–465.
- Benbya, H., Pachidi, S., and Jarvenpaa, S.L. (2021). Artificial intelligence in organizations: implications for information systems research. *Journal of Association for Information Systems* (special issue editorial), 22(2), 281–303.
- Bergman, R., Abbas, A.E., Jung, S., Werker, C., and de Reuver, M. (2022). Business model archetypes for data marketplaces in the automotive industry. *Electronic Markets*, 32(2), 1–19.
- Boyd, D. (2020). Questioning the legitimacy of data. *Information Services and Use*, 40(3), 259–272.
- Carmel, E., Lacity, M.C., and Doty, A. (2016). The impact of impact sourcing: framing a research agenda. In Nicholson, B., Babin, R., and Lacity, M.C. (eds), *Global Sourcing with Social Impact* (pp. 16–47). New York: Palgrave Macmillan.
- Chatterjee, J. (2017). Strategy, human capital investments, business-domain capabilities, and performance: a study in the global software services industry. *Strateg. Manag. J.* 38(3), 588–608.
- Chen, Y., Bharadwaj, A., and Goh, K-Y. (2017). An empirical analysis of intellectual property rights sharing in software development outsourcing. *MIS Quarterly*, 41(1), 131–161.
- Constantiou, I.D., and Kallinikos, J. (2015). New games, new rules: big data and the changing context of strategy. *Journal of Information Technology*, 30, 44–57.

- Contreras, J.L., and Knoppers, B.M. (2018). The genomic commons. *Annual Review of Genomics and Human Genetics*, 19, 429–453.
- Davenport, T., Evgeniou, T., and Redman, T.C. (2021). Your data supply chains are probably a mess. Here's how to fix them. *Harvard Business Review*, June 24, 1–6.
- De-Arteaga, M., Dubrawski, A., and Chouldechova, A. (2021). Leveraging expert consistency to improve algorithmic decision support. *Workshop on Information Systems*, Austin, TX.
- de Corbière, F. and Rowe, F. (2013). From ideal data synchronization to hybrid forms of inter-connections: architectures, processes, and data. *Journal of the Association for Information Systems*, 14(10), 550–584.
- Eirich, J., and Fischer-Pressler, D. (2022). The life cycle of data labels in organizational learning: a case study of the automotive industry. *Proceedings of European Conference on Information Systems*.
- Ellis, S., and Davidi, I. (2005). After-event reviews: drawing lessons from successful and failed experience. *J. Appl. Psychol.* 90(5), 857–871.
- Fountain, T., McCarthy, B., and Saleh, T. (2021). Getting AI to scale. *Harvard Business Review*, May–June, 116–123.
- Gainer, V.S., Cagan, A., Castro, V.M., et al. (2016). The Biobank Portal for Partners Personalized Medicine: a query tool for working with consented biobank samples, genotypes, and phenotypes using i2b2. *Journal of Personalized Medicine*, 6(11), Article 11. doi: 10.3390/jpm6010011.
- Gambal, M.J., Asatiani, A., and Kotlarsky, J. (2022). Strategic innovation through outsourcing—a theoretical review. *Journal of Strategic Information Systems*, 31(2), 101718.
- Gelhaar, J., Groß, T., and Otto, B. (2021). A taxonomy for data ecosystems. HICSS.
- Grønsund, T., and Aanestad, M. (2020). Augmenting the algorithm: emerging human-in-the-loop work configurations. *Journal of Strategic Information Systems*, 29(2), 101614.
- Hanafizadeh, P., and Zareravasan, A. (2020). A systematic literature review on IT outsourcing decision and future research directions. *Journal of Global Information Management (JGIM)*, 28(2), 160–201.
- Hirschheim, R.A., Heinzl, A., and Dibbern, J. (2020). *Information Systems Outsourcing: The Era of Digital Transformation*. Cham: Springer.
- Huber, G.P. (1991). Organizational learning: the contributing processes and the literatures. *Organ. Sci.*, 2(1), 88–115.
- Janssen, M., Charalabidis, Y., and Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Inf. Syst. Manag.* 29, 258–268.
- Jarvenpaa, S.L., and Majchrzak, A. (2016). Interactive self-regulatory theory for sharing and protecting in interorganizational collaborations. *Academy of Management Review*, 41(1), 9–27.
- Jarvenpaa, S.L., and Markus, M.L. (2018, January 5). Data perspective in digital platforms: three tales of genetic platforms. *Hawaii International Conference on System Sciences (HICSS)*, Big Island, Hawaii.
- Jarvenpaa, S.L., and Markus, M.L. (2020). Data sourcing and data partnerships: opportunities for IS sourcing research. In Kirschheim, R., Heinzl, A., and Dibbern, J. (eds), *Information Systems Outsourcing: The Era of Digital Transformation, Progress in IS* (pp. 61–79). Cham: Springer.
- Jussupow, E., Spohrer, K., Heinzl, A., and Gawlitza, J. (2021). Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. *Information Systems Research*, 32(3), 713–735.
- Kennedy, J., Subramaniam, P., Galhotra, S., and Fernandez, R.C. (2022). Revisiting online data markets in 2022: a seller and buyer perspective. *ACM SIGMOD Record*, 51(3), 30–37.
- Kogut, B., and Zander, U. (1992). Knowledge of the firm, combinative capabilities, and the replication of technology. *Organ. Sci.*, 3(3), 383–397.

- Kotlarsky, J., Oshri, I., Dibbern, J., Mani, D. (2018). IS sourcing. In Bush, Ashley and Rai, Arun (eds), *MIS Quarterly Research Curations*, July 1. <http://misq.org/research-curations>.
- Koutroumpis, P., Leiponen, A., and Thomas, L.D. (2020). Markets for data. *Industrial and Corporate Change*, 29(3), 645–660.
- Kramer, R., and Tyler, R. (1996). *Trust in Organizations*. Thousand Oaks, CA: SAGE Publications.
- Krasikov, P., Eurich, M., and Legner, C. (2022). Unleashing the potential of external data: a DSR-based approach to data sourcing. *Thirteenth European Conference on Information Systems (ECIS 2022)*, Timisoara, Romania.
- Lacity, M.C., Khan, S.A., and Yan, A. (2017). Review of the empirical business services sourcing literature: an update and future directions. *Journal of Information Technology*, 31, 269–328.
- Lebovitz, S., Levina, N., and Lifshitz-Assaf, H. (2021). Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what. *MIS Quarterly*, 45(3), 1501–1526.
- Lee, P. (2017). Centralization, fragmentation, and replication in the genomic data commons. In Strandburg, K.J., Frischmann, B.M., and Madison, M.J. (eds), *Governing Medical Research Commons* (pp. 46–73). New York: Cambridge University Press.
- Leonelli, S. (2015). What counts as scientific data? A relational framework. *Philosophy of Science*, 82(5), 810–821.
- Mani, D., Barua, A., and Whinston, A.B. (2010). An empirical analysis of the impact of information capabilities design on business process outsourcing performance. *MIS Quarterly*, 34(1), 39–62.
- March, J.G. (2010). *The Ambiguities of Experience*. Ithaca, NY: Cornell University Press.
- Markus, M.L. (2001). Toward a theory of knowledge reuse: types of knowledge reuse situations and factors in reuse success. *Journal of Management Information Systems*, 18(1), 57–93.
- Markus, M.L. (2016). 11 obstacles on the road to corporate data responsibility. In Sugimoto, Cassidy R., Ekbis, Hamid R., and Mattioli, M. (eds), *Big Data is Not a Monolith: Policies, Practices, and Problems* (pp. 143–161). Cambridge, MA: MIT Press.
- Martin, K. (2019). Designing ethical algorithms. *MIS Quarterly Executive*, 18(2), 129–142.
- Martins, D.M.L., Vossen, G., and de Lima Neto, F.B. (2017, August). Intelligent decision support for data purchase. In *Proceedings of the International Conference on Web Intelligence* (pp. 396–402).
- Maula, M., Heimeriks, K.H., and Keil, T. (2023). Organizational experience and performance: a systematic review and contingency framework. *Academy of Management Annals*, 17(2), 546–585.
- Monteiro, E., and Parmiggiani, E. (2019). Synthetic knowing: the politics of the Internet of Things. *MIS Quarterly*, 43(1), 167–184.
- Morse, L., Teodorescu, M.H.M., Awward, Y., and Kane, G.C. (2021). Do the ends justify the means? Variation in the distributive and procedural fairness of machine learning algorithms. *Journal of Business Ethics*, 181(4), 1–13.
- Oliveira, S., Barros, M.I., Lima, G.D.F., and Farias Lóscio, B. (2019). Investigations into data ecosystems: a systematic mapping study. *Knowledge and Information Systems*, 61, 589–630.
- Pearlson, K., Applegate, L.M., and Ibarra, H. (1994). The Kodak Outsourcing Agreement (A) and (B). Harvard Business School.
- Perkmann, M., and Schildt, H. (2015). Open data partnerships between firms and universities: the role of boundary organizations. *Research Policy*, 44, 1133–1143.
- Pisano, G.P. (1994). Knowledge, integration, and the locus of learning: an empirical analysis of process development. *Strategic Management Journal*, 15(S1), 85–100.

- Polyviou, A., and Zamani, E.D. (2022). Are we nearly there yet? A desires and realities framework for Europe's AI strategy. *Information Systems Frontiers*, 25(1), 143–159.
- Poppo, L., and Lacity, M.C. (2002). The normative value of transaction cost economics: what managers have learned about TCE principles in the IT context. In Hirschheim, R., Heinzl, A., and Dibbern, J. (eds), *Information Systems Outsourcing: Enduring Themes, Emergent Patterns and Future Directions* (pp. 253–276). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-04754-5_12
- Rothe, H., Jarvenpaa, S.L., and Penninger, A.A. (2019). How do entrepreneurial firms appropriate value in bio data infrastructures: an exploratory qualitative study. In *Proceedings of the 27th European Conference on Information Systems (ECIS), Stockholm and Uppsala, Sweden* (pp. 1–17).
- Seidel, S., Berente, N., Lindberg, A., Lyytinen, K., and Nickerson, J.V. (2019). Autonomous tools and design: a triple-loop approach to human–machine learning. *Communications of the ACM*, 62(1), 50–57.
- Sporsem, T., and Tkalic, A. (2020). Data sourcing in the context of software product innovation. SINTEF, Trondheim, Norway.
- Sporsem, T., Tkalic, A., Moe, N.B., and Mikalsen, M. (2021). Understanding barriers to internal startups in large organizations: evidence from a globally distributed company. *2021 IEEE/ACM Joint 15th International Conference on Software and System Processes* (pp. 12–21).
- Stahl, F., Schomm, F., Vomfell, L., and Vossen, G. (2017). Marketplaces for digital data: quo vadis? *Computer and Information Science*, 10(4).
- Steinfeld, C., Markus, M.L., and Wigand, R.T. (2011). Through a glass clearly: standards, architecture, and process transparency in global supply chains. *Journal of Management Information Systems*, 28(2), 75–108.
- Strich, F., Mayer, A-S., and Fiedler, M. (2021). What do I do in a world of artificial intelligence? Investigating the impact of substitutive decision-making AI systems on employees' professional role identity. *Journal of the Association for Information Systems*, 22(2), 9.
- Sturm, T., Gerlach, J.P., Pumplun, L., Mesdah, N., Peters, F., and Tauchert, C. (2021). Coordinating human and machine learning for effective organizational learning. *MIS Quarterly*, 45(3), 1581–1602.
- Teodorescu, M.H., and Xinyu Y. (2021). Machine learning fairness is computationally difficult and algorithmically unsatisfactorily solved. In *2021 IEEE High Performance Extreme Computing Conference (HPEC)* (pp. 1–8).
- Teodorescu, M.H., Morse, L., Awwad, Y., and Kane, G.C. (2021). Failures of fairness in automation require a deeper understanding of human–ML augmentation. *MIS Quarterly*, 45(3), 1483–1500.
- Thomas, L.D.W., Leiponen, A., Koutroumpis, P. (2022). Profiting from data products. In Cennamo, C., Dagnino, G.B., and Zhu, F. (eds), *Handbook of Research on Digital Strategy* (pp. 255–272). Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing.
- Tuomi, I. (1999). Data is more than knowledge: implications of the reversed knowledge hierarchy for knowledge management and organizational memory. *Journal of Management Information Systems*, 16(3), 103–117.
- Van den Broek, E., Sergeeva, A., and Huysman, M. (2022). When the machine meets the expert: an ethnography of developing AI for hiring. *MIS Quarterly*, 45(3), 1557–1580.
- Van den Broeck, T., and van Veenstra, A.F. (2018). Governance of big data collaborations: how to balance regulatory compliance and disruptive innovation. *Technological Forecasting and Social Change*, 129, 330–338.
- Williamson, O.E. (1985). Assessing contract. *Journal of Law, Economics, and Organization*, 1(1), 177–208.
- Winter, J.S., and Davidson, D. (2019). Big data governance of personal health information and challenges to contextual integrity. *Information Society*, 35(1), 36–51.

Zuboff, S. (2015). Big other: surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, 30(1), 75–89.

2. Data work as an organizing principle in developing AI

Angelos Kostis, Leif Sundberg, and Jonny Holmström

INTRODUCTION

A growing number of scholars explore organizing with and through data (Abbasi et al., 2016; George et al., 2014). The proliferation of data and recent advancements in emerging technologies have pushed us to move beyond approaching data as administrative support tools and means for organizational streamlining that operate in the background. Instead, scholars have started acknowledging data as relational entities that actively shape processes of knowing and accomplishing organizational objectives (Alaimo and Kallinikos, 2022; Bailey et al., 2022). However, research on data science and big data analytics has often misleadingly portrayed data as objective, raw facts, or faithful representations of external phenomena which are highly available to the contemporary organizations. As such, data are treated as a source of competitive advantage and innovation if organizations manage to coax untapped treasure out of the enormity of available data. A central assumption in those discourses has been that data are naturally occurring entities and marked by an inherent representational capacity, making learning from data a mere analytics problem (Busch, 2014; Davenport, 2006; Smith, 2020). Yet, the processes through which data are embedded and interweave with organizing and “the actual conditions and practices through which data are made to matter in organizational and industry settings” (Aaltonen et al., 2021, p. 6) have received limited attention.

Acknowledging this limitation, a growing body of research has recently painted a more dynamic picture around the role of data and has argued that data are embedded in a socio-technical sphere influencing how organizing occurs, how knowing is achieved, and how work practices are accomplished (Alaimo et al., 2020; Alaimo and Kallinikos, 2022; Jones, 2019; Mikalsen and Monteiro, 2021). Building on this stream of research, we view data as incomplete, cooked, and non-faithful representations of external phenomena which require data work, meaning “extensive, complex, and ongoing tasks involved in finding and curating the digital data ... that happens before and during data science analytics in exploration and production” (Parmiggiani et al., 2022, p. 3). To this definition of data work, we also add the creative tasks of working with data to define and shape potential trajectories of an artificial intelligence (AI) initiative and make decisions about AI. This suggests that data are relational (Bailey et al., 2022); data are woven into and alter everyday situated work practices (Gitelman, 2013; Günther et al., 2017) and are not mere representations of their physical referent, but active shapers of processes of knowing and emerging forms of

organizing (Alaimo and Kallinikos, 2022). Drawing on this view, in this chapter we focus attention on the role of data as a relational entity and on data work (in terms of working on and with data) as a dynamic organizing principle in AI initiatives.

Developing AI is a challenging and highly uncertain endeavor where various issues such as unfairness may emerge (Teodorescu et al., 2021), requiring iterative interactions, exchanges, and involvement of not only AI experts and their know-how, but also domain experts and their know-what (van den Broek et al., 2021; Lebovitz et al., 2021; Teodorescu et al., 2021). Although advancements in machine learning (ML) (Jordan and Mitchell, 2015), and more specifically, deep learning (DL) (LeCun et al., 2015), have extended the boundaries of what AI can perform compared to previous generations of intelligent systems, developing AI solutions is fundamentally dependent on how domain and AI experts engage in knowledge integration and mutual learning to handle the tension between independence (that is, producing knowledge without relying on domain expertise) and relevance (that is, producing knowledge that is useful to the domain) (van den Broek et al., 2021). AI development is thus a collective process where domain and AI experts enter dialogical exchanges to identify potential use cases and leverage AI opportunities by building on their distinct disciplinary knowledge. Nonetheless, developing AI is not a straightforward, friction-less, linear process where AI and domain experts identify use cases, select appropriate data for training ML models, and leverage AI opportunities. Instead, we argue that developing AI is a highly iterative process characterized by a high degree of uncertainty rooted in the epistemic differences between the two different types of expertise regarding both the availability and features of data, and algorithmic performance.

In this chapter, we argue that developing AI requires coping with epistemic uncertainty, which constitutes a key organizing challenge in this context given the simultaneous presence of epistemic boundaries and the interdependence and need for knowledge exchanges among domain and AI experts. Specifically, building on Mengis et al. (2018), we argue that AI development is typical of “situations where the relevant parameters of a phenomenon are unknown, causal relations become unclear and interdependencies among parties become unpredictable, generating ‘epistemic uncertainty’” (p. 596). Epistemic uncertainty is uncertainty rooted in ignorance of pertinent knowledge that is knowable in principle (Packard and Clark, 2020) and represents a knowledge problem (Townsend et al., 2018). In the context of AI development, epistemic uncertainty is related not only to key actors’ ignorance of domain knowledge, domain specificities, and of the availability, features, and meanings of existing data, but also to ignorance of AI possibilities and limitations. As such, epistemic uncertainty inflicts AI initiatives and the interactions between domain and AI experts, making it hard for decisions about AI solutions to be made, and for AI solutions to be developed. The two types of expertise—domain experts (DEs) and AI/ML experts (MLEs)—involved in AI initiatives are epistemic communities with specialist forms of knowledge that need to be integrated for mutual learning to be achieved and meaningful-to-the-domain AI solutions to be created. Yet, “specialized practice generates boundaries across epistemic communities” (Mengis et al., 2018,

p. 597) and creates epistemic uncertainty which organizations need to cope with for complex innovations to be pursued (Tuertscher et al., 2014). With the high demands on data, the uncertain outcomes of machine learning, and the requirement for AI experts to heavily rely on domain experts' inputs, views, and interpretations, we argue that epistemic uncertainty is amplified in AI contexts. However, prior research has paid limited attention to epistemic uncertainty in the context of AI development, which is an important knowledge gap that needs to be addressed to further enrich and extend insights of the intricate dynamics and mutual learning processes taking place in AI initiatives (van den Broek et al., 2021). If developing AI solutions requires different experts to intersect their specialized knowledge in various combinations yet this process is inflicted by epistemic uncertainty, more knowledge is needed regarding how organizations can cope with epistemic uncertainty in an effective way for AI opportunities to be leveraged. Shedding light on this issue may also provide further insights into why so many data science projects fail to deliver (Joshi et al., 2021).

Prior research suggests that in the light of uncertainty, organizations need organizing principles (Ouchi, 1980), which represent "the logic by which work is coordinated and information is gathered, disseminated, and processed within and between organizations" and guide organizations' interpretations and behaviors (McEvily et al., 2003, p. 92). Among others, organizing principles can be trust, norms, and hierarchy (Powell, 1990), which also play an important role in AI development too as they influence coordination of work among the different experts by facilitating knowledge sharing, establishing specialized roles, supporting problem solving in cases of unanticipated and unsettling events, and motivating domain and AI experts to combine their resources in joint efforts. Yet, given the significant role of data in the process of developing AI, and building on the emerging stream of research looking at data as woven into organizing (Alaimo and Kallinikos, 2022), we argue that data work serves as a unique organizing principle guiding domain and AI experts' interpretations and behaviors. Against this background, we ask: *what are the mechanisms through which data work serves as an organizing principle in developing AI?*

By answering the above research question, we contribute to the emerging data work literature (Aaltonen et al., 2021; Alaimo and Kallinikos, 2022; Jones, 2019; Kallinikos and Constantiou, 2015; Mikalsen and Monteiro, 2021; Parmiggiani et al., 2022; Passi and Jackson, 2018; Stelmaszak, 2022; Waardenburg et al., 2022b) in three ways: by expounding the performative role of data as a relational entity, by providing a processual view on data's interweaving with organizing, and by deciphering the nature of data work as an organizing principle that is collectively accomplished. The rest of this chapter is structured as follows. First, we explain how epistemic uncertainty emerges as a key organizing challenge in AI initiatives. Second, we briefly describe the role of organizing principles in coping with epistemic uncertainty. Third, we connect to the literature on data work which provides key insights on the ontology of data and the nature and importance of data work for developing AI solutions. Then, building on those insights, we develop a theoretical framework deciphering three main mechanisms through which data work serves as an organizing principle in AI initiatives.

EPISTEMIC UNCERTAINTY IN DEVELOPING AI SOLUTIONS

Epistemic uncertainty refers to mitigable ignorance of pertinent knowledge that is knowable in principle (Packard and Clark, 2020), and emerges particularly in situations where specialized knowledge is required and needs to be shared and integrated with others' specialized knowledge for a solution to a problem to be created on the basis of interdisciplinary collaboration (Ewenstein and Whyte, 2009; Mengis et al., 2018). Epistemic uncertainty emerges in the light of "a competence gap in problem solving" (Dosi and Egidi, 1991, p. 769), or else when relevant factors and their interrelations are unknown to key actors involved in an initiative. Building on this, we argue that in the context of interdisciplinary collaboration, epistemic uncertainty is related to ignorance of pertinent knowledge or lack of relevant expertise to overcome existing knowledge boundaries in the interactions between the engaged experts. Knowledge boundaries and disciplines' specificities make it challenging for experts residing in different and often unconnected epistemic communities to engage in such knowledge sharing and integration processes (Mengis et al., 2018; Tuertscher et al., 2014). Epistemic uncertainty and the associated challenge of sharing and integrating knowledge within constellations of actors belonging to different disciplines become especially pronounced as actors largely rely on tacit knowledge to perform their work (Polanyi, 1962; Tsoukas, 2005), which is challenging to articulate and bring to the foreground in its entirety (Tsoukas, 2009). Consequently, this contributes to emergence of knowledge boundaries and increasing epistemic uncertainty that are infused in complex innovation processes (Tuertscher et al., 2014). Knowledge boundaries can inflict central activities of innovation, such as defining product concepts, articulating the procedure for creating the product, and projecting forward to anticipate surprises or potential breakdowns (Dougherty and Dunne, 2012; Kostis and Ritala, 2020). Thus, epistemic uncertainty arises as there is unpredictability stemming from lack of pertinent knowledge of "relevant alternative moves and parameters" (Grandori, 2010, p. 482) important for grasping the complexities of a situation.

In this chapter, we argue that epistemic uncertainty is a key organizing challenge in AI initiatives, given that developing AI involves what Mengis et al. (2018) describe as "situations where the relevant parameters of a phenomenon are unknown, causal relations become unclear and interdependencies among parties become unpredictable, generating 'epistemic uncertainty'" (p. 596). The process of developing AI is inflicted by complexity, dynamism, and stochasticity, which are the three main sources of epistemic uncertainty (Packard and Clark, 2020). First, developing AI is marked by complexity given the requirement to make sense of and grasp both domain specificities and fundamentals of AI. Second, developing AI is inherently dynamic, as changes occur constantly and surprises can fundamentally challenge interrelations between (seemingly) relevant factors or even the relevance of the expertise and/or existing data. Third, changes and surprises do not surface based on specific patterns, meaning that stochasticity inhibits predictability and triggers epistemic uncertainty. Also, as there is no way to "know" the output of a ML/DL model before it has been

trained, the output of the model triggers interpretative practices in the wake of algorithmic affordances. Predictions generated by intelligent systems in organizations are subject to interpretations of key individuals (“algorithmic brokers”), and the consequences of the output rely heavily on those interpretations, and on who these individuals are (Waardenburg et al., 2022a), which can further instigate epistemic uncertainty.

Moreover, as Tuertscher et al. (2014) note, collaboration between unconnected epistemic communities and knowledge integration are particularly challenging in the context of engaging with emerging technologies. Against the above background, we argue that epistemic uncertainty is indeed a key organizing challenge in AI initiatives. However, while prior research on AI use has reported that domain experts experience increasing uncertainty due to lack of ability to understand the reasoning behind a prediction from an AI tool, or else AI knowledge claim (Lebovitz et al., 2022), limited attention has been paid to epistemic uncertainty emerging in developing AI. This is rather surprising, as developing AI fundamentally relies on interdisciplinary collaboration where domain experts and AI experts are involved in knowledge integration and engage in mutual learning (van den Broek et al., 2021). The development of intelligent systems that learn from data is far from linear and involves multiple competences from both the domain in which the system is to be implemented, and data science knowledge. As noted, the interaction between these types of expertise is a highly ambiguous and interactive process characterized by a high degree of uncertainty. This uncertainty not only relates to technical issues such as the availability and features of data and algorithmic performance, but is largely epistemic in nature. Whereas DEs possess situated knowledge about the specificities and processes in their domain, the existing data and the meanings of those data for the domain, AI experts possess knowledge about AI technologies, their underlying algorithms, and the associated limitations and possibilities. Thus, to develop AI solutions, both DE and AI experts need to deal with what Townsend et al. (2018) call a knowledge problem of agentic nature, meaning that either of the epistemic communities “does not possess certitude regarding either the relevant factors or likely consequences of action” (p. 670) to provide relevant contributions.

To sum up, prior research has consistently highlighted the need for AI experts to involve and build on DEs’ expertise, that is, the skills and knowledge accumulated through prior learning within a domain (Choudhury et al., 2020) to ensure production of knowledge that is relevant to the domain (van den Broek et al., 2021; Lebovitz et al., 2021), and attention to biases due to input incompleteness (Choudhury et al., 2020). Yet, such mutual learning and interdisciplinary collaboration is easier said than accomplished in developing AI, as there is an inherent unpredictability associated with the difficulty in even defining what constitutes relevant expertise, in making sense of the features of available data, and in grasping AI possibilities and limitations in connection to the domain.

ORGANIZING PRINCIPLES IN THE LIGHT OF UNCERTAINTY: STRUCTURING AND MOBILIZING PATHWAYS

In the wake of increased digitalization in various industries, there is a need to combine and align different epistemic communities to manage complexity (Dougherty and Dunne, 2012) and other factors giving rise to epistemic uncertainty. Such knowledge integrations in novel settings with various expertise involved are important, yet also challenging and uncertain. While novel practices may lead to contestation within practice communities (Mørk et al., 2010), creative processes where different knowledge communities meet and engage in activities to convert past experiences into new and valuable insights in collective processes may also take place (Hargadon and Bechky, 2006). To effectively cope with epistemic uncertainty, organizations rely on organizing principles, such as trust, norms, clans, and hierarchy (McEvily et al., 2003; Ouchi, 1980; Powell, 1990). Organizing principles provide the core logic behind organizational actors' interpretations and behaviors, and the "interaction patterns and processes that enable and constrain the coordination of work among individuals" (McEvily et al., 2003, p. 94). Those are expected to be relevant in processes of developing AI solutions as they shape the relations and positions of different actors within a social context and encourage key actors to contribute much-needed resources. Organizing principles influence coordination of work among different experts by facilitating knowledge sharing, establishing specialized roles, supporting problem solving in cases of unanticipated and unsettling events, and motivating domain and AI experts to combine their resources in joint efforts. However, emerging digital technologies, with their unique characteristics, have created the need to rethink how organizing occurs, and to articulate new organizing principles. For instance, Hanseth and Lyytinen (2010) articulate adaptation as an organizing principle, as distributed actors adapt to their environment through changes in tasks, technology, and relations. Following the pervasive role that data play in all societal aspects (Alaimo and Kallinikos, 2022; Berente et al., 2019; Constantiou and Kallinikos, 2015), and given the critical role of data for emerging technologies to learn and perform a wide variety of functions, we argue that data work should be understood as an organizing principle.

Given the significant role of data and data work practices in the process of developing AI, we build on an emerging stream of research looking at data as woven into organizing (Alaimo and Kallinikos, 2022) and argue that in developing AI, data work serves as a unique organizing principle guiding domain and AI experts' interpretations and behaviors. To discuss how data work serves this role, we build on the idea that organizing principles influence organizing through two key pathways: structuring and mobilizing (McEvily et al., 2003). Regarding the structuring pathway, we theorize how data work shapes the development of a system of relative positions and links among key actors by considering the delegation dynamics involved in data work. We do so because delegation, meaning the transferring of rights and responsibilities to others (Baird and Maruping, 2021) and even to AI (Mann and O'Neil,

2016), cultivates a system of relative positions and links among entities involved in AI initiatives and shapes those entities' power or even authority to make decisions about the trajectory of the initiative. Thus, by paying attention to various facets of delegation that occur in data work, we offer novel theoretical insights into how data work influences organizing through the so-called structuring pathway. Regarding the mobilizing pathway, to theorize how data work as an organizing principle encourages domain and AI experts to contribute their resources and undertake joint activities for developing AI, we draw on two key concepts. First, given the importance of sharing, integrating, and combining domain-specific and AI-specific knowledge for developing AI, we focus on how the experts' efforts for knowledge interlace—that is, creation of “pockets of shared knowledge interwoven within and across subsystem communities” (Tuertscher et al., 2014, p. 1579)—further motivates domain and AI experts to act and shape the trajectories of the AI initiative. Second, we take a step back and consider how the actors involved construct AI opportunities while performing data work, without having a clear-cut, pre-defined goal in mind. Those opportunities are fundamental motors behind the actors' mobilization of resources to develop AI, and understanding the process through which they emerge will provide insight into how data work influences organizing through the mobilizing pathway. Against this background, to shed light on the process through which AI opportunities are constructed, we utilize the concept of effectuation. Here, it is important to note that the list of mechanisms which we focus on in this chapter (that is, delegation, knowledge interlace, and effectuation) is representative of mechanisms based on which data work influences the structuring of AI initiatives and the mobilizing of resources within such initiatives, but by no means is it comprehensive. We now turn to the literature on data work and present key findings from prior research.

DATA WORK

An emerging body of literature has recently criticized the rational and simplistic explanations provided in computer science and data science discourses regarding the ontology and role of data in organizational life. Data have traditionally been seen as stable and mere inputs to algorithms and as instrumental for value creation purposes if appropriate statistical procedures are in place. As Sambasivan et al. (2021, p. 1) note, “paradoxically, for AI researchers and developers, data is often the least incentivized aspect, viewed as ‘operational’ relative to the lionized work of building novel models and algorithms.” However, a nascent body of research has turned attention to data science as an embedded process and to the data work undertaken backstage by multiple different actors (Aaltonen et al., 2021; Jones, 2019; Kallinikos and Constantiou, 2015; Mikalsen and Monteiro, 2021; Passi and Jackson, 2018; Waardenburg et al., 2022b). Data work studies focus on the “extensive, complex, and ongoing tasks involved in finding and curating the digital data ... that happens before and during data science analytics in exploration and production” (Parmiggiani et al., 2022, p. 3), and address both how data are created and how data are used in practice

(Alaimo and Kallinikos, 2022; Aaltonen et al., 2021). For instance, Mikalsen and Monteiro (2021) identify three data work practices, namely accumulating (that is, triangulating one kind of data by connecting it to other supporting data), reframing (that is, contesting existing interpretations and changing existing models by incorporating new data), and prospecting (that is, producing and deliberating multiple and often contradictory interpretations and possibilities based on different sets of data), all of which are fundamentally entangled with and alter the work of domain experts. Thus, data work scholars place emphasis on the relational and performative nature of data (Jones, 2019), and acknowledge that data have an ambivalent ontology (Kallinikos et al., 2013), are ambiguous, “human-made and bound up with specific practices and institutional settings” (Aaltonen and Penttinen, 2021, p. 5924). Further, data science involves both objectivity and subjective judgments (Joshi, 2020), which testifies to the relational nature of data and that their value depends on what practices and contexts they are entangled with and how they perform towards different ends. Particularly relevant to our argument that data work serves as an organizing principle is the idea that data science is performed not only by skillful, competent data scientists who possess technical knowledge, but also by domain experts who generate, curate, compile, and interpret data alongside their daily work and in collaboration with AI experts (van den Broek et al., 2021). For instance, while data are utilized by data scientists who create new knowledge and alter work practices (Pachidi et al., 2021), research also reports that data work is often performed by other new professions (Pine and Bossen, 2020) and other actors too, such as clerical workers, medical professions (Knudsen and Bertelsen, 2022), police officers (Waardenburg et al., 2022b), and oil experts (Monteiro and Parmiggiani, 2019).

Building on the above, we argue that data work plays a critical role in making AI development a collective and fruitful endeavor where data scientists fundamentally rely not only on their knowledge but also on domain experts’ knowledge, their involvement with data, and their assertive, situated, and creative interventions between data and algorithms. Complex sets of practices of discovering, aggregating, and preparing data are performed by actors holding different types of expertise (Parmiggiani et al., 2022), and inputs by both AI experts and DEs are required (Lebovitz et al., 2021; van den Broek et al., 2021).

In sum, the data work research stream emphasizes the performative, open-ended, and relational nature of data and directs scholarly attention to the question of how data interweave with organizing (Aaltonen et al., 2021). Drawing upon this line of work, we address the following research question: what are the mechanisms through which data work serves as an organizing principle in developing AI?

DATA WORK AS AN ORGANIZING PRINCIPLE: THREE INTRICATE MECHANISMS

The key argument in this chapter is that data work serves as an organizing principle in the process of developing AI solutions, and supports organizations in coping with

epistemic uncertainty that inflicts AI initiatives. We argue that data work serves as an organizing principle through three main mechanisms, namely cultivating knowledge interlace, triggering data-based effectuation, and facilitating multi-faceted delegations. These three mechanisms are set in motion due to the performative, relational, and ambivalent ontology of data, and the situated data work practices undertaken by the two epistemic communities to collect, curate, and label data, to frame AI problems, and to rely on algorithms as organizational realities. In the following, we present each of the three mechanisms based on which data work influences the organizing of AI initiatives.

Cultivating Knowledge Interlace

The first mechanism through which data work serves as an organizing principle is cultivating knowledge interlace. As mentioned earlier, interlaced knowledge refers to “pockets of shared knowledge interwoven within and across subsystem communities” (Tuertscher et al., 2014, p. 1579). We argue that such overlaps in knowledge of distinct epistemic communities emerge through cycles of contestation and justification. Different experts, who possess distinct specialization and are supposed to work together on developing a complex technological solution, may compete and argue for certain paths and features that they believe are relevant and important. Through this process, interlaced knowledge can occur as each of the epistemic communities may reach a deeper understanding of the rationale and knowledge grounds based on which different epistemic communities argue for certain path choices or technological features. We argue that in the process of developing AI solutions, interlaced knowledge can emerge within the epistemic communities of domain and AI experts due to the role of data and work undertaken collectively by the two epistemic communities.

Data have an ambivalent ontology (Kallinikos et al., 2013) and are performative, editable, and re-purposeful, allowing data work to serve as a unique organizing principle in AI initiatives. Data are incomplete, inviting data work by epistemic communities and integration of their distinct knowledge assets. The incompleteness of data encourages further exploration of what data work practices are needed to decrease the scope of incompleteness, better explicate domain knowledge into data, and “craft data into data” (Mikalsen and Monteiro, 2021, p. 2) for developing AI solutions. Data serve as active epistemic objects demonstrating agency and the “capacity to generate questions and to open up different issues to different stakeholders” (Ewenstein and Whyte, 2009, p. 28), thereby encouraging both AI experts and domain experts to engage in data work.

Data incompleteness triggers AI experts not only to ask questions about the existing data and about the domain knowledge those incorporate, but also to demand potential changes in the data to make them usable for AI purposes. Therefore, data incompleteness and their editability feed into cycles of contestation and justification, and cultivate necessary conditions for interlaced knowledge to emerge as domain experts probe for existing data use and AI experts interrogate their meaning and usability for AI purposes. In this process, domain experts are motivated to reflect

upon both domain knowledge incorporated in existing data, and knowledge that is not yet, but could be, incorporated through collecting new data, curating existing data, and aggregating existing, slack, and new data. Consequently, domain experts are involved in self-distanciation, namely “taking distance from their customary and unreflective ways” (Tsoukas, 2009, p. 943) of seeing and approaching existing data. Through self-distanciation, domain experts can draw new distinctions, look at their data and even their domain in different ways than they did in the past, and become motivated to improve datafication processes (Stein et al., 2019) in order for data to better capture their domain knowledge and its specificities. Similarly, AI experts are encouraged to use existing data as a frame of reference to reflect upon algorithms’ possibilities and limitations and their demands in terms of data, educating thereby domain experts on fundamentals of AI. Such dialogic data work, namely engaging with data to boost fruitful dialogue and integration of knowledge held by domain and AI experts, triggers actors to explicate their knowledge and collectively attempt to increase the representational capacity of the data, and make them not only more faithful to their domain (to the extent that this is possible) but also more suitable for AI purposes. Moreover, through data work the attention of the different expert groups is directed towards different aspects of developing AI solutions. While domain experts focus on explicating and incorporating domain knowledge assets, AI experts focus on the possibilities (or limitations) to develop AI solutions with existing and/or new data. Thus, the two epistemic groups reach overlaps in their knowledge while they engage in a similar process, described by Tuertscher et al. (2014) as decomposing and working on “different subsystems in a distributed yet parallel fashion” (p. 1580).

An illustration of the cultivation of knowledge interlace can be found in van den Broek et al (2021), where the blending and integration of knowledge bases from various domains occurs as different groups of actors, such as machine learning developers and domain experts, critically examine the suitability of the ML system and the experts’ contributions to knowledge production. In the case, ML developers and HR professionals recognized the value of the ML system’s activities in generating insights characterized by objectivity, novelty, and efficiency. However, they also identified its shortcomings in terms of domain relevance, especially when the system disregarded valuable practice-based knowledge. These realizations led to the cultivation of an interlaced knowledge base, where the developers iteratively adapted the ML system’s activities by striking a balance between excluding and including domain expertise over time. This process of blending and integrating knowledge from different fields helps to create a more robust and effective system, as it combines the strengths of both machine learning algorithms and human expertise. In turn, this interlaced knowledge supports better decision-making and drives innovation in various domains.

Overall, the mechanism of cultivating knowledge interlace entails that data work, as an organizing principle, does not simply facilitate knowledge sharing between the two types of expertise, but supports identification and development of knowledge overlaps. Data’s characteristics invite intense situated data work, interpretations, and contributions from each of the expert groups involved in developing AI solutions,

which explicate experts' knowledge assets not only to the other group of experts but even to themselves.

Triggering Data-Based Effectuation

The second mechanism through which data work serves as an organizing principle is what we call triggering data-based effectuation. Building on the notion of effectuation (Sarasvathy, 2001; Reuber et al., 2016), as described in entrepreneurship theory, we argue that data work serves as a key process through which possibilities of AI are articulated and problems that AI can solve are constructed. As explained by Sarasvathy (2001), effectuation processes fundamentally differ from causation processes: "causation processes take a particular effect as given and focus on selecting between means to create that effect. Effectuation processes take a set of means as given and focus on selecting between possible effects that can be created with that set of means" (p. 245).

According to Galkina et al. (2022), while effectuation is "means-driven non-predictive logic of entrepreneurial reasoning, in contrast with goal-driven causal logic" (p. 575), the two logics can also lead to synergies as organizations face competing yet interrelated demands, such as starting from goals versus starting from means, and exploiting pre-defined knowledge versus leveraging contingencies. We argue that in the process of developing AI solutions, data work is the main means through which possible effects are articulated, the very AI solutions are defined, and potential areas in which AI can be used are chosen. Data work increasingly plays this role, for two main reasons that are closely associated with the ambivalent ontology of data. First, as data are characterized by open-endedness, which makes it possible to repurpose data to new uses than the ones for which data were initially collected (Aaltonen et al., 2021), data work is vital for such novel uses to be identified and new opportunities to be constructed. Second, data are increasingly approached by many organizations as "dough" for AI solutions. Given the inherent difficulty in defining in advance relevant problems that are solvable with AI, many organizations engage in AI initiatives due to what Boland and Collopy (2004) call design attitude: engage in AI development driven by the idea either that AI can provide better solutions than the current ones, or that AI may solve problems they have not yet paid attention to. This, however, means that domain and AI experts need to perform creative data work practices and come up with potential use cases by engaging with existing data.

An illustration of triggering data-based effectuation can be found in a recent study by Alaimo and Kallinikos (2022), who focus on how data interweave with organizing in an online music discovery platform. The authors show how online forms of organizing are fundamentally built on clustering of data that unleash new opportunities for creating and offering value to users. As noted, "platform operations converge around certain types of data which are first engineered and subsequently deployed as the basis for categorizing music taste" (Alaimo and Kallinikos, 2022, p. 3). The making of basic objects out of data is highlighted as a key process given that it enables the construction and diffusion of tailor-made music recommendations,

supporting the organization in explorations of new opportunities, and accomplishing their organizational objective of music discovery. By standardizing user interfaces and automating backend operations, the identities of organizations such as Last.fm are increasingly entwined with the technologies they deploy, which makes it difficult to distinguish between human and machine roles, and to clearly separate what entity performs data work.

Moreover, being curious about possible solutions and the potential value that AI can offer, organizational members approach the idea of developing AI by juggling with existing data, rather than by having a clear focus in terms of a problem they face, which can be solved with the use of AI. In other words, the available data which can be repurposed provide domain experts and AI experts the basis for experimentation and “muddling through” (Lindblom, 1959); intentionally constructing and framing ML problems through data work while achieving a more in-depth understanding of the situation at hand. As such, data work stimulates the effectuation practices of juggling with data and leveraging contingencies, as it goes beyond merely selecting or labeling data for a specific algorithm, but extends to influence the construction of the very problem AI is supposed to solve in an iterative and experimental fashion. Here, similarities can be drawn to the notion of organizational decision-making as highly fluid processes involving “choices looking for problems,” “solutions looking for issues to which they might be an answer,” and “decision makers looking for work” (Cohen et al., 1972, p. 1).

Facilitating Multi-Faceted Delegations

The third mechanism through which data work serves as an organizing principle in developing AI is what we call facilitating multi-faceted delegations. Data are unbounded (Ekbja, 2009), distributed, fluid, and transfigurible (Kallinikos et al., 2013). Given those characteristics, data: (1) can float among domain experts, AI experts, and agentic AI systems; and (2) can be morphed and transfigured not only by humans, but also by technologies and algorithms. The fluid nature of data entails also re-allocation of responsibilities and decision-making rights to either AI or human agency. Data are thus carriers of authority and when entities perform data work they can trigger different types of delegation, such as domain experts-to-AI experts delegation, human-to-AI delegation, AI-to-human delegation and even AI-to-AI delegation.

The notion of delegation has traditionally been a key theme in management and leadership studies (see Akinola et al., 2018; Chen and Aryee, 2007) but given the growing agency of contemporary IS artifacts and AI in particular (see Iansiti and Lakhani, 2020), delegation has received increased scholarly attention (Baird and Maruping, 2021; Candrian and Scherer, 2022; Fuegener et al., 2022). Despite the different epistemological assumptions of those research streams and their distinct units of analysis, there is convergence on what delegation is. For instance, IS scholars focus on the interactions between human and agentic IS artifacts and define delegation in terms of “transferring rights and responsibilities for task execution

and outcomes to another” (Baird and Maruping, 2021, p. 317). Similarly, while management and leadership studies focus on different units of analysis (that is, supervisor–subordinate dyadic interactions), they define delegation in a similar manner: “assignment of responsibilities to subordinates and conferral of authority to carry out assigned tasks” (Akinola et al., 2018, p. 1467). In essence, delegation draws on the idea of interchangeability of human and technical agency (cf., Latour, 1988). Acts of “technical” delegation function by altering the material world, whereas acts of “social” delegation seek to shape organizational action by more traditionally human means. Building on those definitions, we view delegation in the context of AI as a process of distributing work and transferring authority, decision-making rights, and task responsibilities between humans and AI. Importantly, we argue that in this process, data work plays a pivotal role as it is the engine behind such rights assignments, distributions, and transfers. Through data work, data can move across the boundaries between domain experts, AI experts, and AI algorithms, and thereby different types of delegation can take place.

First, domain experts-to-AI experts delegation occurs when data move from the boundary space of domain experts to the boundary space of AI experts, who in turn use the data to train an ML model; as AI expertise comes to the forefront, there is a delegation of authority and decision-making rights to AI experts who are now responsible for performing the required work and train the machine to make reliable predictions. Knowledge expressed as data always manifests in a digital form, which makes it at least partially determined by the technological conditions under which data are constructed. For example, data prepared for ML are generated in a professional culture of computer and data scientists. Here, as domain knowledge is subject to algorithmic processing, Fonseca (2022) describes how the domain, at least temporarily, plays an auxiliary role. When the data have been processed, however, and translated into deployed models, data science and AI experts take an auxiliary role as these algorithms become organizational realities (cf. Østerlie and Monteiro, 2020).

Second, human-to-AI delegation occurs when data are algorithmically manipulated and thus transfigured by AI systems that demonstrate agency. In line with previous literature that emphasizes the role of “agentic” IS artifacts (Baird and Maruping, 2021; Shrestha et al., 2019), we argue that humans delegate tasks and responsibilities to intelligent systems as they rely on the deployed models to make predictions and enable automation and augmentation in the organization. As noted by Baird and Maruping (2021), not all IS agents are equal, though, and the type and degree of delegation largely depends on the properties and purpose of the intelligent agent in the case of ML. Hence, we suggest that delegation generates a “distance” between the domain and intelligent artifacts, and that this distance is dependent on the data and associated data work in which domain experts can be engaged. For example, most ML applications today are based on supervised approaches where data are labeled and annotated by humans to “teach” a machine to distinguish features, such as objects in an image. These activities can be contrasted to unsupervised procedures where this annotation is “delegated” to a machine via clustering algorithms. However, with the advent of generative AI, the distance between the domain and intelligent artifacts

is likely to become irrelevant as generative AI possesses the capacity to create context-specific content as it can be fine-tuned on open-ended data that capture the specificities of a domain.

Third, AI-to-human delegation occurs when the AI solution is operative in providing domain experts with actionable predictions based on the training data. In this process, domain experts need to judge algorithmic predictions and decision-making, consider the AI knowledge claims (Lebovitz et al., 2022), and calibrate the model by engaging in data work, in case that is needed. Data work is thus the basis for authority to be transferred to algorithms and for algorithmic decisions to be made, yet such decisions serve as input to human decision-making (see Shrestha et al., 2019). In this way, new cycles of delegation may emerge, but also new knowledge interlace processes may be triggered. In the following, we further extend our argument regarding AI-to-human delegation. We acknowledge that in the face of increasingly complex interconnections among humans and AI, there is an urgent need to shed further light on instances where agency is passed back and forth across the shifting line between humans and AI. While the idea of delegation in AI work has been discussed by pointing at how AI can delegate agency to humans, and humans can delegate agency to AI (e.g., Fuegener et al., 2022), we have only scratched the surface regarding when, how, and why delegation should occur in order to be productive. To elaborate on this, we briefly examine variations in the locus of agency in data work and the locus of agency in organizational decision making. This helps us to identify four situations where delegation may vary depending on whether data work and decision-making are accomplished by humans and/or AI. First, when data work is supported by assisting technologies, delegation is wielded by humans in both a locus of agency in data work and a locus of agency in decision-making. Second, in situations where data work is supported by augmenting technologies, delegation is comprised by a focal concern on technology related to a locus of agency in data work, and a locus of human agency in decision-making. Third, data work delegated to arresting technologies occurs, for instance, when data workers are using data to train a ML model, and this model serves as an “arresting technology” (e.g., Murray et al., 2021), as the locus of agency in decision-making context lies with the technology. Finally, in situations where data work is delegated to automating technologies, delegation is different from the first type in that it builds on technological agency in data work. As such, this type of agency links to our discussion above on the role of “agentic” IS artifacts (see Baird and Maruping, 2021).

As an illustration of the complexities associated with delegation, Morse et al. (2021) examine how unfairness from algorithmic decision-making may be eliminated. Specifically, the authors seek to gain insight into unfairness in algorithmic decision-making by exploring fairness perceptions of algorithmic criteria. In so doing, the authors encourage organizations to recognize that managing fairness in machine learning systems is complex, and that adopting a one-size-fits-all mentality will not work. Instead, firms should seek to avoid adopting a one-size-fits-all approach toward algorithmic fairness criteria. Taken together, one could argue that the three generic types of delegation, which manifest differently depending on when,

how, and why transferring of data work and decision-making rights occurs, are a remedy to such a one-size-fits-all approach toward algorithmic fairness criteria. They help to explain the multi-faceted nature of delegation in the process of developing AI solutions. Finding functional processes of delegating helps in closing the gap between an organization’s current data work capabilities and those needed to meet external expectations to develop AI solutions. In this way, organizations can better cope with epistemic uncertainty and overcome challenges associated with AI/data science initiatives (see Joshi et al., 2021; Sundberg and Holmström, 2023).

CONCLUDING REMARKS

This chapter took as a point of departure the pervasive role of data and data work in contemporary organizing. We argue that developing AI requires coping with epistemic uncertainty, which constitutes a key organizing challenge in this context given the interdependence and need for knowledge exchanges among domain and AI experts. In the context of AI development, epistemic uncertainty is related not only to key actors’ ignorance of domain knowledge, and the availability and features of existing data, but also to ignorance of AI’s possibilities and limitations. We reflect on the pervasive role of data in contemporary organizing, and epistemic uncertainty as a key organizing challenge, and present some food for thought for scholars to consider data work as a key organizing principle in AI initiatives. We argue that data work serves as an organizing principle through three interrelated mechanisms, namely cultivating knowledge interlace, triggering data-based effectuation, and facilitating multi-faceted delegations. These three mechanisms are set in motion due to the ontology of data and situated data work loops in which DEs, MLEs, and agentic AI engage (see Figure 2.1).

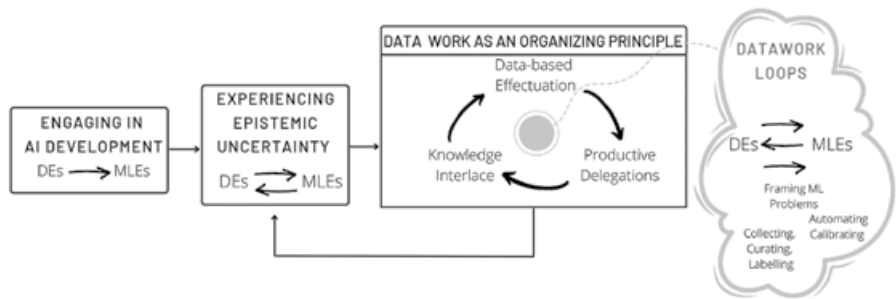


Figure 2.1 Data work as an organizing principle in developing AI solutions

We contribute to the emerging stream of research on data work in three distinct ways. First, we extend the view that data are relational and performative (Aaltonen et al., 2021) by showing how their characteristics encourage organizations to approach data

work as a unique organizing principle in AI initiatives. Second, by deciphering the mechanisms through which data work interweaves with organizing and facilitates coping with epistemic uncertainty, we question computer science and data science approaches to data as mere inputs to algorithms and analytics, and provide a processual view on the role of data. Third, we further explain the collective and hybrid practices in developing AI (van den Broek et al., 2021) by establishing that situated data work practices are collectively undertaken by domain experts, AI experts, and agentic AI. Building on our arguments, future research can empirically investigate how data work influences organizing, and how different data work practices enable or hinder coping with epistemic uncertainty or other types of uncertainty that emerge in AI initiatives, especially in constructing training datasets. In this way, organizations and managerial practice can also be supported to make more informed decisions about AI.

As a final note, the framework we develop in this chapter offers important insights and guidance to organizations seeking to implement AI initiatives. The framework places emphasis on several aspects that need to be considered when making decisions about AI, such as decisions regarding how to construct AI opportunities, how to collaborate with AI experts, and how to delegate responsibilities to others in the process of building training datasets. One overarching implication is that organizations need to prioritize data work when making those decisions, and to embrace data as a relational entity that can co-shape processes of knowing. Creating solutions that are meaningful and relevant to the domain requires firms to work on and with their data, which implies that to make effective decisions about AI, organizations need to devote resources to build data work skills and to create a team with a diversity of expertise to perform a wide array of data work practices. While there is a broadly accepted claim regarding increasing data availability, our framework highlights that data need to be re-worked, re-purposed, curated, aggregated, and often re-shaped and labeled. Such laborious and creative data work needs to be collective, and firms need to establish routines to collaborate and exchange knowledge with AI experts, and even educate them on the domain and the sources and meanings of their data. Collaborating with AI experts is also key when constructing AI opportunities and building the training dataset. Both types of experts bring unique knowledge and understandings to the AI initiative, and working together on and with data can help in dealing with epistemic uncertainty, in grasping the linkages between data and algorithms, and in ensuring the creation of a solid training dataset. Regarding decision-making about the construction of AI opportunities, our framework emphasizes the importance of engaging in an AI initiative by deeply considering the existing or slack data and the possibilities that exist with those. Our framework highlights that AI opportunities emerge in the interaction space between AI experts, domain experts, and existing data which are treated as the means towards potential ends. Yet, firms need to be ready to change trajectories and come up with new use cases, as through data work they may identify a need to move beyond existing data, and consider collecting new data that are not readily available but can be relevant to the AI initiative. This, however, requires domain experts to be educated by AI experts and embrace more reflexive ways

of seeing and approaching existing data and established data practices. It requires creative thinking and a willingness to explore new sources of data, which can be cultivated based on interactions with AI experts, who bring knowledge regarding AI's possibilities and limitations and can assess the potential of data. Finally, our framework suggests that organizations need to make decisions regarding the extent to which responsibilities are delegated to other entities, as productive delegations can be critical in dealing with epistemic uncertainty, in constructing the training dataset, and in effectively developing AI. To sum up, data work as an organizing principle of an AI initiative guides domain experts and all the other entities involved by providing the core logic through which decisions about AI are made and the trajectories of AI are shaped over time.

REFERENCES

- Aaltonen, A., Alaimo, C., and Kallinikos, J. (2021). The making of data commodities: Data analytics as an embedded process. *Journal of Management Information Systems*, 38(2), 401–429.
- Aaltonen, A., and Penttinen, E. (2021). What makes data possible? A sociotechnical view on structured data innovations. In *Proceedings of the 54th Hawaii International Conference on System Sciences*.
- Abbasi, A., Sarker, S., and Chiang, R. H. (2016). Big data research in information systems: Toward an inclusive research agenda. *Journal of the Association for Information Systems*, 17(2), 3.
- Akinola, M., Martin, A.E., and Phillips, K.W. (2018). To delegate or not to delegate: Gender differences in affective associations and behavioral responses to delegation. *Academy of Management Journal*, 61(4), 1467–1491.
- Alaimo, C., and Kallinikos, J. (2022). Organizations decentered: Data objects, technology and knowledge. *Organization Science*, 33(1), 19–37.
- Alaimo, C., Kallinikos, J., and Aaltonen, A. (2020). Data and value. In Nambisan, S., Lyytinen, K., and Yoo, Y. (eds), *Handbook of Digital Innovation*. Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing, 162–178.
- Bailey, D.E., Faraj, S., Hinds, P.J., Leonardi, P.M., and von Krogh, G. (2022). We are all theorists of technology now: A relational perspective on emerging technology and organizing. *Organization Science*, 33(1), 1–18.
- Baird, A., and Maruping, L.M. (2021). The next generation of research on IS use: A theoretical framework of delegation to and from agentic IS artifacts. *MIS Quarterly*, 45(1), 315–341.
- Berente, N., Seidel, S., and Safadi, H. (2019). Research commentary—Data-driven computationally intensive theory development. *Information Systems Research*, 30(1), 50–64.
- Boland, R., and Collopy, F. (eds) (2004). *Managing as Designing*. Stanford, CA: Stanford Business Books.
- Busch, L. (2014). Big data, big questions. A dozen ways to get lost in translation: Inherent challenges in large scale data sets. *International Journal of Communication*, 8, 18.
- Candrian, C., and Scherer, A. (2022). Rise of the machines: Delegating decisions to autonomous AI. *Computers in Human Behavior*, 134, 107308.
- Chen, X.Z., and Aryee, S. (2007). Delegation and employee work outcomes: An examination of the cultural context of mediating processes in China. *Academy of Management Journal*, 50(1), 226–238.

- Choudhury, P., Starr, E., and Agarwal, R. (2020). Machine learning and human capital complementarities: Experimental evidence on bias mitigation. *Strategic Management Journal*, 41(8), 1381–1411.
- Cohen, M.D., March, J.G., and Olsen, J.P. (1972). A garbage can model of organizational choice. *Administrative Science Quarterly*, 17(1), 1–25.
- Constantiou, I.D., and Kallinikos, J. (2015). New games, new rules: Big data and the changing context of strategy. *Journal of Information Technology*, 30(1), 44–57.
- Davenport, T.H. (2006). Competing on analytics. *Harvard Business Review*, 84(1), 98.
- Dosi, G., and Egidi, M. (1991). Substantive and procedural uncertainty: An exploration of economic behaviours in changing environments. *Journal of Evolutionary Economics*, 1, 145–168.
- Dougherty, D., and Dunne, D.D. (2012). Digital science and knowledge boundaries in complex innovation. *Organization Science*, 23(5), 1467–1484.
- Ekbja, H.R. (2009). Digital artifacts as quasi-objects: Qualification, mediation, and materiality. *Journal of the American Society for Information Science and Technology*, 60(12), 2554–2566.
- Ewenstein, B., and Whyte, J. (2009). Knowledge practices in design: The role of visual representations as epistemic objects. *Organization Studies*, 30(1), 7–30.
- Fonseca, F. (2022). Data objects for knowing. *AI and Society*, 37(1), 195–204.
- Fuegener, A., Grahl, J., Gupta, A., and Ketter, W. (2022). Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *Information Systems Research*, 33(2), 678–696.
- Galkina, T., Atkova, I., and Yang, M. (2022). From tensions to synergy: Causation and effectuation in the process of venture creation. *Strategic Entrepreneurship Journal*, 16(3), 573–601.
- George, G., Haas, M.R., and Pentland, A. (2014). Big data and management. *Academy of Management Journal*, 57(2), 321–326.
- Gitelman, L. (2013). *Raw Data is an Oxymoron*. Cambridge, MA: MIT Press.
- Grandori, A. (2010). A rational heuristic model of economic decision making. *Rationality and Society*, 22(4), 477–504.
- Günther, W.A., Mehrizi, M.H.R., Huysman, M., and Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. *Journal of Strategic Information Systems*, 26(3), 191–209.
- Hanseth, O., and Lyytinen, K. (2010). Design theory for dynamic complexity in information infrastructures: The case of building internet. *Journal of Information Technology*, 25, 1–19.
- Hargadon, A.B., and Bechky, B.A. (2006). When collections of creatives become creative collectives: A field study of problem solving at work. *Organization Science*, 17(4), 484–500.
- Iansiti, M., and Lakhani, K.R. (2020). *Competing in the Age of AI: Strategy and Leadership when Algorithms and Networks Run the World*. Boston, MA: Harvard Business Press.
- Jones, M. (2019). What we talk about when we talk about (big) data. *Journal of Strategic Information Systems*, 28(1), 3–16.
- Jordan, M.I., and Mitchell, T.M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255–260.
- Joshi, M.P. (2020). Custodians of rationality: Data science professionals and the process of information production in organizations. *AMCIS 2020 Proceedings*.
- Joshi, M.P., Su, N., Austin, R.D., and Sundaram, A.K. (2021). Why so many data science projects fail to deliver. *MIT Sloan Management Review*, 62(3), 85–89.
- Kallinikos, J., Aaltonen, A., and Marton, A. (2013). The ambivalent ontology of digital artifacts. *MIS Quarterly*, 37(2), 357–370.
- Kallinikos, J., and Constantiou, I.D. (2015). Big data revisited: A rejoinder. *Journal of Information Technology*, 30(1), 70–74.

- Knudsen, C., and Bertelsen, P. (2022). Maintaining data quality at the hospital department level—The data work of medical secretaries. *Scandinavian Conference on Health Informatics* (pp. 159–165).
- Kostis, A., and Ritala, P. (2020). Digital artifacts in industrial co-creation: How to use VR technology to bridge the provider–customer boundary. *California Management Review*, 62(4), 125–147.
- Latour, B. (1988). A relativistic account of Einstein’s relativity. *Social Studies of Science*, 18(1), 3–44.
- Lebovitz, S., Levina, N., and Lifshitz-Assaf, H. (2021). Is AI ground truth really “true”? The dangers of training and evaluating AI tools based on experts’ know-what. *Management Information Systems Quarterly*, 45(3), 1501–1525.
- Lebovitz, S., Lifshitz-Assaf, H., and Levina, N. (2022). To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization Science*, 33(1), 126–148.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- Lindblom, C.E. (1959). The science of “muddling through.” *Public Administration Review*, 19(2), 79–88.
- Mann, G., and O’Neil, C. (2016). Hiring algorithms are not neutral. *Harvard Business Review*, 9, 2016.
- McEvily, B., Perrone, V., and Zaheer, A. (2003). Trust as an organizing principle. *Organization Science*, 14(1), 91–103.
- Mengis, J., Nicolini, D., and Swan, J. (2018). Integrating knowledge in the face of epistemic uncertainty: Dialogically drawing distinctions. *Management Learning*, 49(5), 595–612.
- Mikalsen, M., and Monteiro, E. (2021). Acting with inherently uncertain data: Practices of data-centric knowing. *Journal of the Association for Information Systems*, 22(6), 1715–1735.
- Monteiro, E., and Parmiggiani, E. (2019). Synthetic knowing: The politics of the internet of things. *MIS Quarterly*, 43(1), 167–84.
- Morse, L., Teodorescu, M.H.M., Awwad, Y., and Kane, G.C. (2021). Do the ends justify the means? Variation in the distributive and procedural fairness of machine learning algorithms. *Journal of Business Ethics*, 181, 1083–1095.
- Mørk, B.E., Hoholm, T., Ellingsen, G., Edwin, B., and Aanestad, M. (2010). Challenging expertise: On power relations within and across communities of practice in medical innovation. *Management Learning*, 41(5), 575–592.
- Murray, A., Rhymer, J., and Sirmon, D.G. (2021). Humans and technology: Forms of conjoined agency in organizations. *Academy of Management Review*, 46(3), 552–571.
- Nicolini, D., Mengis, J., and Swan, J. (2012). Understanding the role of objects in cross-disciplinary collaboration. *Organization Science*, 23(3), 612–629.
- Østerlie, T., and Monteiro, E. (2020). Digital sand: The becoming of digital representations. *Information and Organization*, 30(1), 100275.
- Ouchi, W.G. (1980). Markets, bureaucracies and clans. *Administrative Science Quarterly*, 25, 129–141.
- Pachidi, S., Berends, H., Faraj, S., and Huysman, M. (2021). Make way for the algorithms: Symbolic actions and change in a regime of knowing. *Organization Science*, 32(1), 18–41.
- Packard, M.D., and Clark, B.B. (2020). On the mitigability of uncertainty and the choice between predictive and nonpredictive strategy. *Academy of Management Review*, 45(4), 766–786.
- Parmiggiani, E., Østerlie, T., and Almklov, P.G. (2022). In the backrooms of data science. *Journal of the Association for Information Systems*, 23(1), 139–164.
- Passi, S., and Jackson, S.J. (2018). Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human–Computer Interaction*, 2(CSCW), 1–28.

- Pine, K.H., and Bossen, C. (2020). Good organizational reasons for better medical records: The data work of clinical documentation integrity specialists. *Big Data and Society*, 7(2), 1–13.
- Polanyi, M. (1962). *Personal Knowledge*. Chicago, IL: University of Chicago Press.
- Powell, W.W. (1990). Neither market nor hierarchy: Network forms of organization. *Research in Organizational Behavior*, 12, 295–336.
- Reuber, A.R., Fischer, E., and Coviello, N. (2016). Deepening the dialogue: New directions for the evolution of effectuation theory. *Academy of Management Review*, 41(3), 536–540.
- Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., and Aroyo, L.M. (2021, May). “Everyone wants to do the model work, not the data work”: Data cascades in high-stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–15).
- Sarasvathy, S.D. (2001). Causation and effectuation: Toward a theoretical shift from economic inevitability to entrepreneurial contingency. *Academy of Management Review*, 26(2), 243–263.
- Shrestha, Y.R., Ben-Menahem, S.M., and Von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review*, 61(4), 66–83.
- Smith, G. (2020). Data mining fool’s gold. *Journal of Information Technology*, 35(3), 182–194.
- Stein, M.K., Wagner, E.L., Tierney, P., Newell, S., and Galliers, R.D. (2019). Datification and the pursuit of meaningfulness in work. *Journal of Management Studies*, 56(3), 685–717.
- Stelmaszak, M. (2022). Inside a data science team: Data crafting in generating strategic value from analytics. ECIS 2022 Research Papers, 86.
- Sundberg, L., and Holmström, J. (2023). Democratizing artificial intelligence: How no-code AI can leverage machine learning operations. *Business Horizons*, 66, 777–788.
- Teodorescu, M.H.M., Morse, L., Awwad, Y., and Kane, G.C. (2021). Failures of fairness in automation require a deeper understanding of human–ML augmentation. Do the ends justify the means? Variation in the distributive and procedural fairness of machine learning algorithms. *MIS Quarterly*, 45(3), 1483–1499.
- Townsend, D.M., Hunt, R.A., McMullen, J.S., and Sarasvathy, S.D. (2018). Uncertainty, knowledge problems, and entrepreneurial action. *Academy of Management Annals*, 12(2), 659–687.
- Tsoukas, H. (2005). Do we really understand tacit knowledge? *Managing Knowledge: An Essential Reader*, 107, 1–18.
- Tsoukas, H. (2009). A dialogical approach to the creation of new knowledge in organizations. *Organization Science*, 20(6), 941–957.
- Tuertscher, P., Garud, R., and Kumaraswamy, A. (2014). Justification and interlaced knowledge at ATLAS, CERN. *Organization Science*, 25(6), 1579–1608.
- van den Broek, E., Sergeeva, A., and Huysman, M. (2021). When the machine meets the expert: An ethnography of developing AI for hiring. *MIS Quarterly*, 45(3), 1557–1580.
- Waardenburg, L., Huysman, M., and Sergeeva, A.V. (2022a). In the land of the blind, the one-eyed man is king: Knowledge brokerage in the age of learning algorithms. *Organization Science*, 33(1), 59–82.
- Waardenburg, L., Sergeeva, A., and Huysman, M. (2022b). Juggling street work and data work: An ethnography of policing and reporting practices. *Academy of Management Proceedings*, 2022(1), 16697.

3. Natural language processing techniques in management research

Mike H.M. Teodorescu

INTRODUCTION

During the last few decades, various management disciplines became heavily dependent on machine learning methods and tools. Domains such as marketing (Gans et al., 2017; Struhl, 2015), financial markets (Tetlock, 2007; Tan et al., 2007; Bollen et al., 2011), risk management (Hu et al., 2012; Chen et al., 2012), knowledge management (Williams and Lee, 2009; Li et al., 2014; Balsmeier et al., 2016), and operations research (Jordan and Mitchell, 2015), among others, are inconceivable today without the use of vast quantities of data and machine learning tools. Machine learning is the study of methods that make it possible to find patterns in data and the subsequent use of these patterns to construct predictions and inferences and to make decisions. The purpose of this chapter is to give an overview of text processing methods which can consume text data, as well as their applications to management research.

In keeping with the precedent of several methods of review articles in the management literature (Hayton et al., 2004; Sardeshmukh and Vandenberg, 2017; Tonidandel et al., 2018), this chapter aims to serve the interested reader as a tutorial with fundamental methodological tools via steps and examples that are accessible and easily reusable. The interested reader will also find targeted references to in-depth methodological content expanding the methods surveyed here, and to a set of relevant articles in our management literature that showcase some of these methods. The references are aimed at a broad audience, with applications across multiple fields within business research. An understanding of statistics is useful, as is familiarity with some programming.

Given text-based methods' growing use in our field and their partial independence of the other machine learning methods, the chapter first presents and exemplifies textual analysis methods following concepts from statistical natural language processing such as term frequency, textual similarity, corpora considerations, and sentiment analysis (Manning and Schütze, 1999). The chapter then covers an overview of some additional methods which involve topic modeling, classification, and word embeddings. Some of the references included point to examples of implementation, as well as to an easy-to-use machine learning toolkit¹ that requires no programming background or toolkits in popular languages such as R or Python. This subsection covers libraries and toolkits for the interested reader.

TEXTUAL ANALYSIS FOR INFORMATION PROCESSING AND DECISION MAKING

This section of the chapter is dedicated to natural language processing methods, for a variety of reasons, including that vast amounts of business and scientific information are recorded in the form of text, and that text analysis techniques can be perceived as more intuitive to use than general machine learning methods. As a source of business information, text can substantially increase the amount of data available for regression or classification models. Textual information can also unlock new research questions for management scholars.

Many of the main sources of information in business decisions come in textual form, such as corporate filings (e.g., Li, 2010a), financial disclosures (e.g., Loughran and McDonald, 2011, 2014), customer messages (e.g., Struhl, 2015; Balazs and Velásquez, 2016; Piryani et al., 2017; Gans et al., 2017), internal corporate documents such as corporate emails (e.g., Srivastava et al., 2017) and chief executive officer (CEO) diaries (e.g., Bandiera et al., 2020), and patents (e.g., Hall et al., 2001; Trajtenberg et al., 2006; Kaplan and Vakili, 2015; Li et al., 2014; Balsmeier et al., 2016). The use of the information contained in text collections is based on methods pertaining to the domain of natural language processing (NLP).

NLP is the interpretation of text using automated analytical methods. A non-exhaustive list of subfields of NLP includes language parsers and grammars, text and speech recognition, sentiment analysis (including its impacts on firm and individual behavior), document classification (including insurance fraud detection, spam detection, and news manipulation), analysis of customers' and investors' sentiment tendencies (Lugmayr, 2013), search query disambiguation (for example, handling of word associations, abbreviations, and polysemy), market segmentation, customer churn modeling, and many more. Researchers in management, strategy, marketing, and accounting have all found applications of NLP relevant to understanding consumer, firm, government, and individual executive behavior.

Text Analysis Workflow

Text requires a sequence of processing stages to be quantified into variables which can then be used in regressions or classifications. A typical text processing workflow is depicted in Figure 3.1. The first step of any textual analysis is to determine the sample of interest, which is generally referred to as a collection of documents, where a document refers to an observation. A document or observation can be as short as a tweet, or span tens of thousands of words such as a financial report or a patent specification. The analysis of documents requires the comparison of their features with those of corpora, which are comprehensive bodies of text representing a field or a natural language.

The text preprocessing steps consist of tokenization, lemmatization or stemming, and stop words removal. Tokenization means segmenting a text, which is essentially a string of symbols including letters, spaces, punctuation marks, and numbers, into



Figure 3.1 Typical workflow for processing text

words and phrases. For example, a good tokenizer treats expressions such as “business model” as a single token, and processes hyphenation (Manning et al., 2008).

The other two preprocessing steps are used depending on the purpose of the analysis. For example, when one wishes to differentiate between the specific languages used by two authors, one may wish to determine how frequently they use common words such as “the,” “and,” “that.” These words are called “stop words” and serve grammatical purposes only. In contrast, when one is interested in sentiment analysis, words that carry semantic meaning matter; stop words are generally held not to carry semantic meaning, so for such analyses they should be removed in preprocessing. Lemmatization, the reduction of the words to their lemma (the dictionary form of the word), helps to lessen both the computational task and the duration of the analysis. It also disambiguates the semantic meaning of the words in a text by assigning words with the same meaning to their lemma. In sentiment analysis, for example, “improve,” “improved,” “improvement,” and “improves” all point equally to an optimistic sentiment and share the same root; differentiating them would serve no purpose for a sentiment analysis task. The lemmatizer does distinguish between different parts of speech, and notes whether the word is used as a verb or a noun. A typical lemmatizer is the WordNet lemmatizer; several other stemmers and lemmatizers are described in Manning et al. (2008).

In other cases, information on the part of speech is not relevant for the analysis, and a simple removal of the prefixes and suffixes to reach the stem of the word is sufficient. The stem is the root of the word, the smallest unit of text that conveys the shared semantic meaning for the word family. For example, the stem of “teaching” is “teach.” Because stemmers do not look up meaning in the context of parts of speech, verbs and nouns resolve to the same root, which reduces complexity, but at the cost of a loss of information. Stemmers are standard in any programming language or toolkit that enables text analysis. For example, the Porter Stemmer (Porter, 1980) produces on the corpus of patent titles the token “autom,” which when applied to the standard American English corpus used in the literature, the Brown corpus (Kučera and Francis, 1967), finds that the stem corresponds to “automobile,” whereas the expected word is “automate.”

While there is no generalized rule in the literature about where to use a stemmer versus a lemmatizer, all text preprocessing workflows should include at least one of the two. For complex technical texts, such as patents, lemmatization is recommended. Further background in grammars, lemmatizers, stemmers, and text processing in general can be found in the comprehensive textbook by Manning and Schütze (1999) and in Pustejovsky and Stubbs (2012).

Vector Space Model

The preprocessing steps allow us to prepare a document for consumption by a variety of numerical methods. The standard representation of a document is called the “vector space model,” as each distinct word in the document becomes a feature of the document; the text can then be represented as a vector of words, with each word assigned a value. If the collection of documents is represented in an N -dimensional space, where N is the total number of distinct words across the collection (its vocabulary V), then each individual document is represented as a point within this N dimensional space. Each dimension (axis in the corresponding diagram) represents a different word from the vocabulary of this collection. The numerical values on the axes for each document may be calculated in different ways. For a comprehensive review of term-weighting methods, see Salton and Buckley (1988). There are four typical methods:

1. Binary weighting at the document level assigns a value of 1 for the word’s presence in the document and 0 for the word’s absence in the document. This is useful in document classification tasks, where the presence or absence of a term is what matters in assigning the document to a particular topic (Albright et al., 2001).
2. Raw term frequency is the raw count of the word in the document. It is useful in applications of sentiment analysis, where counts of positive and negative words are taken to determine the overall sentiment of the text. Widely used annotated dictionaries include SENTIWORDNET² (Baccianella et al., 2010) and the University of Illinois at Chicago’s Opinion Lexicon³ (Hu and Liu, 2004).
3. The relative term frequency (TF) is calculated as the ratio between the number of occurrences of a word in a document, and the number of times the word appears in the entire collection of documents. The tokenizer preprocessing step is essential for creating the proper list of words for each document, as it removes punctuation and non-word text. Stop words typically are removed prior to calculating TF. Inflections of a word would artificially lower the TF, which makes lemmatization/stemming critical. Importantly, the TF measure does not account for words that are common across documents.
4. Using the TF one can create a separate set of weights called term frequency-inverse document frequency (TF-IDF) that takes into account the number of documents in which the word appears through a separate measure called inverse document frequency (IDF). Denoting the number of documents in the collection as D and the number of documents containing the i^{th} word in the alphabetically ordered vocabulary vector as D_i , the IDF is $IDF[i] = \log_2(D/D_i)$. It is apparent that words that are common to all documents would lead to an IDF of 0. The TF-IDF is thus defined for each word i in document D_i as $TFIDF[D_i, i] = TF[D_i, i] \cdot IDF[i]$. The effect of multiplying the term frequencies for each word in each document by the inverse document frequency of that word is that words that are common across documents are weighted down, as they receive a low IDF value. In contrast, uncommon terms that reveal specifics about a document, such as the

methodological and technical terms that make a particular document unique, are weighted up by multiplication by the IDF. This is particularly useful when determining the extent of the difference between pairs of documents, and is the standard method used in the NLP literature.

For text analysis applications that target rare features in a document, TF-IDF is the method of choice. For instance, patents use a specialized language in which common words are generally irrelevant. Younge and Kuhn (2016) performed TF-IDF on the entire patent corpus and determined the differences across patents using cosine similarity on the word vectors associated with each patent. Another application of TF-IDF in management is the comparison of corporate financial forms such as 10-Ks and 10-Qs (Li, 2010b), where words common to most firms or forms are not particularly useful for extracting features of the firm's strategy.

Textual Similarity Measures

The common similarity measures used in text analysis are cosine similarity, the Pearson correlation, the Jaccard similarity, and the Dice similarity. Cosine similarity has been used to compare texts for the past 30 years (Salton and Buckley, 1988; Manning and Schütze, 1999). The cosine similarity is computed as the cosine of the angle of the pair of word vectors representing the two texts, denoted as $\vec{w}_1$ and $\vec{w}_2$. The components of these vectors are usually word counts (Manning and Schütze, 1999, p. 301):

$$\cos(\vec{w}_1, \vec{w}_2) = \frac{\sum_i w_{1i} \cdot w_{2i}}{\sqrt{\sum_i w_{1i}^2} \cdot \sqrt{\sum_i w_{2i}^2}}$$

The cosine similarity defined above may use TF or TF-IDF as a weighting method to create the values in each vector (see Salton and Buckley, 1988, for an extensive review). Unlike cosine similarity and the Pearson correlation coefficient, the Jaccard and Dice similarity indices require binary weighting for both vectors, thus acting at the set-level of the vocabularies W of the texts. The Jaccard similarity measures the number of shared components between the two sets of words, and is defined (using set notation) as:

$$Jaccard(W_1, W_2) = \frac{|W_1 \cap W_2|}{|W_1 \cup W_2|},$$

where W_1 , W_2 are the vocabularies for the two texts. Dice similarity is defined likewise, with the key difference that it rewards shared word pairs while simultaneously penalizing pairs of texts that share fewer pairs of words relative to the total text sizes (Manning and Schütze, 1999, p. 299):

$$Dice(W_1, W_2) = \frac{2 \cdot |W_1 \cap W_2|}{|W_1| + |W_2|}.$$

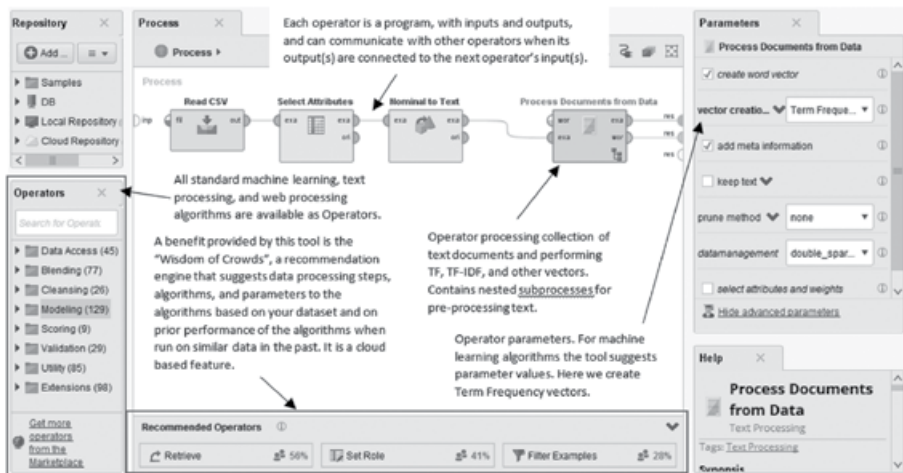
Both Jaccard and Dice indices are used in information retrieval tasks, such as classification of documents and querying (Willett, 1988). Overviews of these and other typical measures are in Manning and Schütze (1999, pp. 294–307), Salton and Buckley (1988), and Huang (2008).

Similarity measures are key to Hoberg and Phillips's (2010) study showing that firms with products very similar in textual descriptions to those of their rivals have lower profitability, and to Younge and Kuhn's (2016) study of how patent text similarities can predict future innovation. Arts et al. (2018) apply Jaccard similarity to the patent corpus to determine technological similarity classes. Textual similarity measures can also be helpful in creating comparison groups and identifying new classification structures. For example, they can help to find companies that create comparable products despite being in different SIC codes (Hoberg and Phillips, 2010), companies with similar customer review sentiments, or companies that have received comparable news coverage. These comparison groups can then be used in regression analysis.

MACHINE LEARNING TOOLS AND PROGRAMMING LANGUAGES

In the category of toolkits for data mining and processing, there are several, such as WEKA, RapidMiner, and KNIME, that are convenient due to the ease of use and fast learning curve. Several, such as RapidMiner and KNIME, enable users to run full machine learning algorithms applying just a drag-and-drop interface, while also providing suggestions for the best parameters for the algorithms. In RapidMiner, the recommendations are based on the inputted data and also on input from a cloud-based platform to which the software is connected, which compares the performance of various algorithms on millions of datasets. While regular programming languages that support machine learning packages, such as Python, C#, and R, provide more functionality, they may require a higher learning cost than toolkits, while providing more flexibility in terms of models that can be applied. Data collection from the Internet is automated in some packages, a further advantage for the management researcher. In this section, I also provide an overview of two natural language processing packages useful to management researchers working with text: NLTK and AYLIEN, a cloud-based toolkit.⁴ Under languages supporting machine learning, I briefly survey Python, C#, and Java. A review of these and other languages and tools is available in Louridas and Ebert (2017).

A programming task in these tools may reduce to selecting a sequence of prebuilt operators to create a sequence of linked operators that forms a process. Figure 3.2 depicts a process that computes the term frequency vectors for a collection of text documents in RapidMiner, a machine learning toolkit which integrates with Python (a commonly used language for machine learning applications), but also provides data processing and algorithm functionalities through a visual interface which may be easier to start with for the reader first tackling a machine learning algorithm.



Note: This annotated image of the RapidMiner User Interface is for illustrative academic purposes only.

Figure 3.2 Example of a typical workflow for processing a collection of text documents with an overview of the RapidMiner interface

In Figure 3.2, the input for the collection of documents is specified by an operator from the Data Access list. The Select Attributes operator allows selections for the columns to be used as input variables to the text processing algorithm. The actual document processing occurs in the Process Documents operator, which can take a wide variety of inputs, for example from a collection of files, from Twitter/X, or from a custom website.

Most toolkits include naïve-Bayes, tree-based algorithms, nearest neighbor, and support vector machine algorithms, neural networks, and others. Toolkits also provide standard statistical models and methods, including a suite of regression, segmentation, and correlation operators. Toolkits offer a wide variety of web mining tools (Kotu and Deshpande, 2014), including tools that gather data from any website given search parameters, gather data from websites with authentication, gather data from Twitter, and collect emails from an email server. The latter two have proven especially useful data sources for recent strategy research. For instance, Gans et al. (2017) analyzed sentiment in customer tweets to predict firm behavior. Srivastava et al. (2017) applied a tree-based machine learning approach to a firm's email server and applied a tree-based approach to determine how well employees matched the firm's email culture, and how differences in culture may impact employee turnover. The methods in these two papers could be implemented in current toolkits such as RapidMiner with just a few dragged-and-dropped operators, without the need to learn a programming language, and this may greatly lower the barrier to entry for additional researchers in using text analysis and machine learning techniques.

A unique feature of some toolkits compared to programming languages with support for machine learning is the ability to incorporate into the algorithms previous successful experience and knowledge from other sources. The use of the “wisdom of crowds” has been applied in many fields, such as biology, medicine, and NLP (Savage, 2012). A “wisdom of crowds” cloud engine (see the lower part of Figure 3.2, RapidMiner implementation) is a useful complement; it provides suggestions for parameter values for the operators as well as a sequence of operators that construct a program to analyze the inputted data.

The ability to visualize data and results is built into many tools, such as Tableau, Qlik, SAS, MATLAB, and RapidMiner. However, RapidMiner is more limited in its visualization capabilities than visualization tools such as Tableau and Qlik, or visualization packages such as D3 or Python’s Matplotlib. Tableau provides a free academic license for students and faculty. Good overviews of MATLAB for finance and economics include Anderson (2004) and Brandimarte (2006).

Statistical languages such as R provide machine learning packages, such as the “caret” library which acts as an interface for machine learning algorithms contained in other packages. R requires individual packages for different algorithms, as each package is relatively limited in scope (packages are available at CRAN). For example, “rpart” is used for basic classification algorithms, but ensemble methods require additional packages, such as “party” or “randomforest.” Other packages are built around specific algorithms, such as neural networks in “nnet” and kernel-based machine learning models in “kernLab.” Generally, these require a bit more research and learning than the prebuilt packages in MATLAB, RapidMiner, or SAS. Two good resources for working with machine learning algorithms in R are Friedman et al. (2001) and the associated datasets and packages, and the UC Irvine Machine Learning Repository (Lichman, 2013).

SAS makes possible the statistical data analysis, data management, and visualization that are widely used in business intelligence. It claims a more accessible interface than R, with targeted packages for specific fields. Such specialized packages are not free, but provide a wide array of tools, as in the case of Enterprise Miner, which provides a comprehensive set of machine learning tools, overviewed in Hall et al. (2014), the closest equivalent in terms of functionality to the tools already discussed. Like RapidMiner and the freeware R, SAS has a free version for academic use called SAS OnDemand for Academics. The general-purpose programming languages Python, C#, and Java all have a variety of machine learning, text analysis, and web mining packages. For example, in Python, the typical packages covering machine learning functionality include NLTK for natural language processing, scikit-learn and pylearn2 for machine learning methods, BeautifulSoup for web parsing, pandas for data parsing from files, and Matplotlib (MATLAB-like interface) and Seaborn for data visualization. For C#, a good library for machine learning is Accord.NET, and a good library for natural language processing is Stanford’s CoreNLP. Machine learning package examples for Java include the user-friendly freeware Weka and Java-ML.

In terms of packages specifically targeted to natural language processing, NLTK is a comprehensive text analysis platform for Python, whereas AYLIEN is a cross-language cloud-based text processing toolkit with advanced sentiment analysis, news parsing, and named entity extraction abilities. NLTK is better for corpus analytics, as it incorporates over 100 text corpora from different fields,⁵ contains a lemmatizer based on WordNet, and has extensive functionality for sentence parsing based on grammars. For an exhaustive overview of NLTK capabilities and examples, see Bird et al. (2009).

For the management researcher interested in easily collecting data about firms and then analyzing the data for sentiment or for entity extraction (locations, individuals, company names, product names, currency amounts, emails, or telephone numbers) from news sites, Twitter, documents, or websites in general, AYLIEN is available as a text extension for RapidMiner and as a Python, Java, and C# package. The news and Twitter parsers allow the user to connect these entities to collections of text documents, which can then be linked to events such as stock prices or product launches, and assigned a sentiment value through the prebuilt sentiment analyzer.

Sentiment Analysis and the Naïve-Bayes Classifier Using NLP

Investor sentiment is known to affect stock returns (Lee et al., 1991), and investors themselves are known to be influenced by the sentiment of news articles (Tetlock, 2007; Devitt and Ahmad, 2007), by the sentiment of conventional media (Yu et al., 2013), by social media (Bollen et al., 2011), and by nuances of optimism about future events as reported in standard financial filings (Li, 2010b). Attitudes and sentiments are detected by counting “positive” and “negative” words and expressions, using specific “bags” (sets) of sentiment/opinion words in lexicon-based detection methods (e.g., Taboada et al., 2011; Ravi and Ravi, 2015), and calculating sentiment scores as the ratios of these counts (Struhl, 2015). The second class of methods for sentiment detection pertains to machine learning. Various types of supervised classifiers are used in the literature to mine for the sentiments in a text, such as neural networks (NN), support vector machines (SVM), rule-based (RB) systems, naïve-Bayes (NB), maximum entropy (ME), and hybrids. Ravi and Ravi (2015) provide details on the classifiers and related machine learning techniques in opinion mining. N-grams, which are uninterrupted sequences of N tokens, are often used in sentiment analysis to classify the sentiment of expressions. In the cases of online data sources, tokens may include punctuation constructs in the form of emoticons, and N-gram analysis considers the affect that an emoticon carries, such as through the use of bi-grams (pairs of tokens) to analyze consumer behavior and sentiment with regard to actions in the airline industry (Gans et al., 2017).

In most sentiment analysis applications, a classification decision must be made regarding the type of sentiment (positive, negative, or neutral) at the document level or the sentence level. The typical classifier used in this context (Li, 2010b; Gans et al., 2017) is the naïve-Bayes, a fast, general-purpose classifier popular in sentiment analysis (e.g., Melville et al., 2009; Dinu and Iuga, 2012).

The naïve-Bayes classifier works by using Bayes's rule for each classification decision under the assumption that all predictors are independent of each other. The name of the classifier is drawn from this assumption, which yields an oversimplification in many situations but also makes this the simplest classifier, with no parameters to tune. The lack of parameter tuning makes it one of the fastest classifiers, which is especially useful in problems in which real-time analysis is needed, such as stock trading and question-answering bots. The naïve-Bayes algorithm calculates the posterior probability for each class given a predictor, and picks the class with the highest posterior probability as the outcome of the classification.

Textual Corpora

Business applications such as marketing sentiment and shareholder analysis require large corpora composed of collections of messages and documents. A linguistic corpus is a “systematically collected” set of “machine-readable texts,” “representative of ‘standard’ varieties of [a language]” (Leech, 1991, p. 10; Pustejovsky and Stubbs, 2012, p. 8). A corpus should be a representative sample of the overall language, which may be a natural language or a specialized language such as those used in patents, individual scientific fields, financial reports, consumer reviews, or short online texts (Tweets, product descriptions, firm mission statements). A comprehensive list of the most used and freely available corpora is provided in the appendix of Pustejovski and Stubbs (2012). It is widely accepted that the standard American English corpus is the Brown corpus (Kučera and Francis, 1967), which was created as a representative sample of the English language by a group at Brown University in the 1960s.

The selection of an appropriate corpus for the research setting is essential, as using a general-purpose corpus can lead to misleading results (Li, 2010a). Corpora may be tagged and annotated to enhance their analytical usefulness. For example, words may be tagged with their part of speech to enable statistical analysis of the inherent grammar of the language, as in the case of Treebanks, whose gold standard is the Penn TreeBank (Marcus et al., 1993). Such tags can then be used as an input to allow a classification algorithm to learn to classify a wider body of text.

Topic Modeling

Topic Modeling encompasses a set of methods that extract themes from bodies of texts; these methods can be used to classify texts as well as to summarize them. Topic models are conceptually a layer above the vector space model described earlier, as topic models also look at the relationships between words, such as words often occurring in close proximity, whereas the vector space model looks at raw term counts. Probabilistic topic modeling approaches assume that words follow a distribution over topics, and that topics follow a distribution over documents in a corpus. In other words, documents are obtained through a generative process in which topics are generated from a distribution, and those topics themselves are generated from

another distribution of words. The approach that has become a standard in the field is the Latent Dirichlet Allocation (LDA) of Blei et al. (2003), in which the generation of every document in a corpus is described as a probabilistic generative process (for the full derivation, see Blei et al., 2003, pp. 996–1006).

The assumption in LDA is that the number of topics of the corpus—the corpus being the entire collection of documents—is known. The documents are the observations; the topics, topic-document, and word-topic assignments are all hidden variables. There are multiple convergence approaches, of which one of the most popular is collapsed Gibbs sampling (Porteous et al., 2008; Asuncion et al., 2009; Xiao and Stibor, 2010). Variational Bayes has also been growing in popularity due to its speed on very large online corpora (Hoffman et al., 2010). One can implement LDA using variational Bayes through the Python Gensim⁶ package.

A topic model can be a useful way to categorize a large collection of documents. For an example in this section, I have chosen to run two topic models on the corpus of all United States Patent claims between 2009 and 2012; a publicly available dataset⁷ of millions of documents (Marco et al., 2019). The implementation is in Python and imported the libraries NLTK for natural language processing tools and Gensim for topic modeling. A nonparametric extension of LDA is HDP—Hierarchical Document Process—which solves the limitation of LDA, in which the number of topics had to be prespecified (Teh et al., 2007). HDP can be utilized to determine the optimal number of topics for a given corpus. However, the reader should not draw general assumptions about the interpretability of topics from this limited example; the interpretability of topics generated by topic modeling remains an active area of research in computer science, with approaches shifting to neural networks such as long-short term memory recurrent neural networks algorithms (Ghosh et al., 2016; Liu et al., 2016). The example in Table 3.1 utilizes the variation of HDP from Wang et al. (2011), implemented in Python's Gensim library. Table 3.1 shows topics that are interpretable yet differ across the two methods. This is to be expected, as different topic modeling approaches can yield different statistical models of the corpus. As the machine does not know what a “correct” set of topics would be from the standpoint of interpretability, this is left to the researcher reviewing the output of the topic model. The output is only as useful as it is interpretable; for large numbers of topics, topics with lower rank in the data may not be very interpretable.

Topic models on large bodies of texts, such as patents, firm press releases, or even firm internal documents, are useful for performing a finer comparison of knowledge transfers, competition, and differences in firm strategies. This is an area of active recent research in strategic management (Younge and Kuhn, 2016; Arts et al., 2018; Teodorescu, 2018), and there is an opportunity to create a similarity measure and comprehensive dataset.

Recent efforts in strategic management combine topic modeling with convolutional neural networks (CNN) applied to recognizing facial expressions in order to ascertain CEO communication styles (Choudhury et al., 2019). This and related works convert the spoken speech of a CEO to written text which can then be analyzed using text analysis techniques such as topic modeling (this paper utilized standard

Table 3.1 Example of an output of a topic model: topics identified through LDA versus top topics identified through HDP for the 2009–2012 patent claims corpus

Rank	LDA all patents 2009–2012 top topics	HDP all patents 2009–2012 top topics
1	(data, end, structure, group, level)	(apparatus, image, member, data, surface)
2	(image, data, set, signal, configured)	(data, signal, image, memory, display)
3	(data, memory, circuit, configured, element)	(signal, data, layer, user, based)
4	(data, group, control, signal, member)	(control, memory, cell, signal, acid)
5	(apparatus, member, end, image, configured)	(data, image, computer, based, apparatus)
6	(layer, control, group, light, apparatus)	(data, light, surface, apparatus, layer)
7	(layer, data, user, configured, image)	(data, configured, computer, network, user)
8	(power, side, control, group, signal)	(end, power, member, configured, surface)
9	(body, material, value, apparatus, end)	(layer, surface, material, region, element)
10	(acid, surface, signal, layer, material)	(data, signal, user, value, network)

LDA) and sentiment analysis and then apply a separate type of algorithm, CNN, to code the facial expressions in order to determine non-verbal cues of the CEO's behavior.

Management information systems as a field has focused more than other branches of management research on how to integrate machine learning techniques to automation of hiring; something that companies such as HireVue and SHL are already doing in practice. The combination of speech-to-text algorithms to obtain vectors out of candidate responses, speech features such as intonations and stresses, and emotions from facial expressions and their appropriateness in the context of the interview question, are already tools in existence in practice; though some are exploring questions of fairness in applying such methods to hiring, invoking issues such as training data biases, lack of an agreement on ground truth, and issues of user trust (Martin, 2019; van den Broek et al., 2021; Berente et al., 2021; Morse et al., 2022; Teodorescu et al., 2022c; Figueroa-Armijos et al., 2022).

Word Embeddings: Word2Vec and Related Techniques

The techniques described above do not measure the semantic relationship between words; rather, they are measures of statistical co-occurrence and linguistic patterns. These techniques are useful in situations where vocabulary analyses are relevant, such as determining the first time a term appeared in a patent application, or the closest competitors of a firm based on the text in a firm's financial statements, websites, or news posts. However, these techniques do not work well across different languages, where meaning may not be translated properly using standard machine translation tools (Carlson, 2023). Further, words that are not quite synonymous may be closely related; for instance, jobs within the same industry sector, or relationships between country names and capital cities (Mikolov et al., 2013), or taste preferences in foods (Howell et al., 2016). Using a large matrix of stemmed dictionary words and applying cosine similarity might result in a problemspace of much higher com-

putational dimensionality, which could become highly expensive to run and beyond regular consumer computing hardware, often requiring either substantial cloud computing expenses or owning specialized hardware such as a compute cluster (for an example of very extensive work with cosine similarity on the entire patent corpus, see Younge and Kuhn, 2016). Similarity-based techniques for certain specialized corpora such as patent texts have been validated with experts (such as in Arts et al., 2018) and have extensive applications in the literature, for instance in the next section discussing applications to the innovation literature in management and economics.

However, a method that reduces the dimensionality of the problem and allows for cross-language comparisons, as well as allowing representations of idioms in a vector space format (Mikolov et al., 2013) is useful provided that training on a sufficiently large corpus can be performed using a softmax output layer neural network approach.⁸ Pretrained word2vec approaches are available, such as Google's word2vec tool,⁹ on corpora such as Google News or Wikipedia articles. If run on specialized texts such as legal documents, financial statements, and so on, a separate corpus representative of documents in those categories should be sampled and the word2vec trained on that specialized corpus. Given that the technique represents words in vectors, cosine similarity can be applied here as well, and we can compare the semantic similarity of pairs of words: for instance the pairs (Paris, France); (Berlin, Germany) would reveal the close semantic relationships, similar to any pairs of capitals and respective countries.

In a recent paper by Guzman and Li, competitive strategic advantage at founding of a startup is measured using the website text of the firm at founding, and finding its five closest incumbents using word2vec, and a substantial effort in scraping of historical websites to find startup marketing statements via the Wayback Machine, and drawing financial data such as funding from Crunchbase (Guzman and Li, 2023). As another example in the strategy literature, in a recent article in *Strategic Management Journal* (Carlson, 2023), sentence embeddings (related method) are derived from descriptions of microenterprises. Since microenterprises lack the formal funding mechanisms of venture-backed startups yet are a significant source of employment in some countries, this study captures the relationship between differentiation and firm performance (a core question in strategy) in a novel context (10,000 microenterprises in eight developing countries). The sentence embedding technique (Reimers and Gurevych, 2020) allows for better comparison of texts across languages than a typical translation and cosine similarity approach.

However, word2vec and related techniques are not a panacea: apart from the training time required, and potentially higher computation time than more straightforward techniques, it can also suffer from bias as the training of the neural network in word2vec is based on existing texts, which could capture historical biases such as gender bias or bias against certain socio-economic statuses, or racial bias (Bolukbasi et al., 2016). Differences between word2vec generated vectors do tend to represent relationships between the pair of words and are similar in size to pairs of words semantically related. This can result in differences between terms such as "man" and "woman," revealing unfortunate historical gender stereotypes (Bolukbasi et

al., 2016). Some of the management literature has been researching this issue—that machine learning tools may not always be beneficial to apply—and forms the sub-field of machine learning fairness, which combines work in management information systems, computer science, and ethics (more references are in the next section of this chapter).

Applications to the Innovation Literature

Natural language processing techniques have proven essential in determining characteristics of firm strategy in structured language settings such as patents, trademarks, and court filings. Researchers in economics of innovation, strategy, finance, and more broadly management research, have utilized various aspects of patent text and different approaches to generate variables from textual data in regression models. Early work focused on characterizing relationships between patents, from the highly impactful NBER citation dataset (Hall et al., 2001), measurements of patent claim scope (Marco et al., 2019), similarity measures between pairs of patents (Younge and Kuhn, 2016; Arts et al., 2018; Whalen et al., 2020), ethnicity (Breschi et al., 2017) and gender attribution (Martinez et al., 2016; Toole et al., 2021) of inventors, name and location disambiguation (Monath et al., 2021), testing of firm strategy (Ruckman and McCarthy, 2017; Kuhn and Thompson, 2019; Kuhn and Teodorescu, 2021), and patent office policy (deGrazia et al., 2021; Pairolero et al., 2022) using text-based measures.

A variety of similarity techniques are applied to patents; for example, Arts et al. (2018) use Jaccard similarity (which is faster to compute than cosine similarity, but does not capture the same degree of nuances) to match patents that are technologically similar and test past literature on knowledge spillovers; importantly, this paper validates the applicability of the measure to patent texts using titles and abstracts of issued patents. Kuhn and Teodorescu (2021) employ Jaccard similarity to determine which firms end up accelerating their patent examination and what strategic benefits they might gain from reduced patent uncertainty. Technology similarity may be useful as a measure of knowledge flows within multinationals in a gravity model approach (Teodorescu et al., 2022a).

A very extensive text analysis effort is undertaken in Younge and Kuhn's (2016) working paper which applies pairwise cosine similarity to specifications of patents (a much larger body of text than abstracts and titles), and in Thompson and Kuhn (2020) which identifies patent races between competitors using this technique. DeGrazia et al. (2021) use several text-based techniques to show a novel negotiation mechanism in the patent examination process that accounts for a decrease in patent pendency without affecting patent quality (the examiner's amendment technique). The paper applies TF-IDF cosine similarity to determine patent examiner specialization similarity. Ruckman and McCarthy (2017) use firm and patent-level data to determine technology similarity on vectors of topics (using abstract text) in order to learn what drives licensors to certain patents in the biopharma space. Numerous other papers in the innovation literature employ text analysis techniques; this section

is just a sample of a variety of papers in the field. Newer techniques for information retrieval, such as those based on kernel methods for relation extraction and deep learning methods, should be considered when researchers have suitably large datasets to train the models on.

DISCUSSION AND CONCLUSIONS

This chapter surveyed part of the field of natural language processing for methods relevant to recently accepted practices in management research. Next, the chapter provided examples of toolkits and programming languages which the interested reader could use to implement some of the methods discussed. Current and future research directions were mentioned, and where applicable, extensive references were provided for the interested researcher. Several types of ML methods, including neural networks, graph-based techniques, decision maps, and Bayesian networks, among others, had to be omitted to attain some depth in the discussion of the other methods, especially those based on natural language processing and classification.

While this chapter provides a number of use cases and toolkits for data analysis, machine learning and NLP-based algorithms should not be viewed as a panacea for deriving measures of relevance to management research, or seen as risk-free. In the field of machine learning fairness, which has been growing in the scholarship of management information systems, computer science, and business ethics research, concerns abound regarding the use of machine learning for outcomes of socio-economic importance, such as lending, housing, hiring, and justice (Martin, 2019; Tarafdar et al., 2020; Berente et al., 2021; Teodorescu et al., 2022b; Morse et al., 2022). Current standard machine learning toolkits in common programming languages optimize on performance metrics such as accuracy, F-1 score, AUC, and so on. This selection of criteria ignores the issues of disparate outcomes by protected attributes such as gender, race, ethnicity, veteran status, age, and others. The optimization should include checking fairness criteria (Verma and Rubin, 2018; Teodorescu et al., 2022b; Morse et al., 2022), which sometimes indicate bias even in the case of algorithms that achieve overall high accuracy (Teodorescu and Yao, 2021). This may be due to biases existing in training data due to historical reasons such as biases in word embeddings (Bolukbasi et al., 2016), or due to culturally different preferences in context-dependent prediction problems such as determining who is hireable, where ground truth is difficult to ascertain (van den Broek et al., 2021; Teodorescu et al., 2022c). This relatively new research area is fertile ground for new scholars to begin work in, as well as a good opportunity for interdisciplinary collaborations. For example, algorithms have created a new space where employers and employees on one end enforce, and on the other resist, algorithm use in the workspace (Kellogg et al., 2020). A leading journal in information systems, *MIS Quarterly*, has dedicated an entire issue to the topic of managing machine learning (ML) fairly in the Special Issue on Managing AI of September 2021 (Berente et al., 2021). Additional perspectives from management scholars on fairness in ML would be welcome.

The technical complexity of the methods employed does not necessarily imply interpretability of the measures derived, and there are numerous opportunities to apply fairly straightforward text analysis techniques such as bag of words and calculations of similarity to core questions in management, as exemplified in the innovation research literature or in recent strategic management articles employing text analysis techniques. The extensive set of references provided in this chapter are a helpful resource to researchers interested in starting with text analysis or machine learning in management, as well as to provide additional references for scholars already applying these techniques.

ACKNOWLEDGMENTS

The author thanks the editors of this *Handbook* as well as the peer reviewers for the very helpful feedback provided, which much improved the chapter. Much earlier draft versions of the chapter were shared, as is customary in our field, in the form of working papers, as HBS Working Paper Series #18-011, SSRN working paper site, and an earlier version was presented at the INFORMS annual meeting conference. This chapter was revised twice, following a presentation at the Copenhagen Business School Contributors' Meeting in 2022, and peer and editorial feedback, for which the author is grateful. The author also thanks Alfonso Gambardella, Tarun Khanna, Shane Greenstein, John Deighton, William Kerr, Neil Thompson, Michael Toffel, Frank Nagle, Andy Wu, Stephen Hansen, Andrew Tole, Asrat Tesfayesus, Raffaella Sadun, Info Mierswa, Knut Makowski, Aaron Yoon, Andrei Hagiu, Yo-Jud Cheng, and Daniel Brown for feedback as well as suggestions for literature references. The programming and statistical tools referenced are for illustration purposes only.

NOTES

- 1 The mentions throughout this chapter of various toolkits and software packages are not an endorsement of them. The opinions expressed are solely of the author, and are based on his experience with these toolkits, languages, and packages.
- 2 The SENTIWORDNET annotated corpus for sentiment analysis research is available at: <http://sentiwordnet.isti.cnr.it/>. Accessed December 30, 2019.
- 3 The Opinion Lexicon consists of 6800 English words annotated with positive and negative sentiment and is freely available at the University of Chicago's website: <https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html>. Accessed December 30, 2019.
- 4 <https://aylien.com/>. Accessed April 9, 2023.
- 5 For a list of current linguistic corpora included with NLTK, see http://www.nltk.org/nltk_data/. Accessed April 9, 2023.
- 6 Gensim is a topic modeling toolkit available for the Python programming language, available at <https://radimrehurek.com/gensim/>. Accessed April 9, 2023.
- 7 Patent Claims Research Dataset, United States Patent and Trademark Office: <https://www.uspto.gov/learning-and-resources/electronic-data-products/patent-claims-research-dataset>. Accessed April 9, 2023.

- 8 A readily available implementation of Word2Vec is in the Gensim package in Python; there is a similar package in R called word2vec, and a plug-in to the pre-trained word2vec model by Google in library rword2vec. Variations such as doc2vec are also easily findable in Python and R.
- 9 Google word2vec tool: <https://code.google.com/archive/p/word2vec/>. Accessed November 1, 2022.

REFERENCES

- Albright, R., Cox, J., and Daly, K. 2001. Skinning the cat: comparing alternative text mining algorithms for categorization. In *Proceedings of the 2nd Data Mining Conference of DiamondSUG*. Chicago, IL. DM Paper (Vol. 113).
- Anderson, P.L. 2004. *Business Economics and Finance with MATLAB, GIS, and Simulation Models*. Chapman & Hall/CRC.
- Arts, S., Cassiman, B., and Gomez, J.C. 2018. Text matching to measure patent similarity. *Strategic Management Journal*, 39(1): 62–84.
- Asuncion, A., Welling, M., Smyth, P., and Teh, Y.W. 2009. On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (pp. 27–34). Montreal: AUAI Press.
- Baccianella, S., Esuli, A., and Sebastiani, F. 2010. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. *Proceedings of LREC 2010*, 2200–2204.
- Balazs, J.A., and Velásquez, J.D. 2016. Opinion mining and information fusion: a survey. *Information Fusion*, 27: 95–110.
- Balsmeier, B., Li, G.C., Chesebro, T., Zang, G., Fierro, G., Johnson, K., Kaulagi, A., Lück, S., O'Reagan, D., Yeh, B., and Fleming, L. 2016. Machine learning and natural language processing on the patent corpus: data, tools, and new measures. UC Berkeley Working Paper, UC Berkeley Fung Institute, Berkeley, CA.
- Bandiera, O., Prat, A., Hansen, S., and Sadun, R. 2020. CEO behavior and firm performance. *Journal of Political Economy*, 128(4): 1325–1369.
- Berente, N., Gu, B., Recker, J., and Santhanam, R. 2021. Managing artificial intelligence. *MIS Quarterly*, 45(3): 1433–1450.
- Bird, S., Klein, E., and Loper, E. 2009. Learning to classify text. In S. Bird, E. Klein, and E. Loper (eds), *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit* (pp. 221–259). Sebastopol, CA: O'Reilly Media.
- Blei, D.M., Ng, A.Y., and Jordan, M.I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan): 993–1022.
- Bollen, J., Mao, H., and Zeng, X. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1): 1–8.
- Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., and Kalai, A.T. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, 29. Neurips 2016 Proceedings. https://proceedings.neurips.cc/paper_files/paper/2016.
- Brandimarte, P. 2006. *Numerical Methods in Finance and Economics. A MATLAB-Based Introduction*. Hoboken, NJ: John Wiley & Sons.
- Breschi, S., Lissoni, F., and Miguelez, E. 2017. Foreign-origin inventors in the USA: testing for diaspora and brain gain effects. *Journal of Economic Geography*, 17(5): 1009–1038.
- Carlson, N.A. 2023. Differentiation in Microenterprises. *Strategic Management Journal*, 44(5): 1141–1167.
- Chen, H., Chiang, R.H.L., and Storey, V.C. 2012. Business intelligence and analytics: from big data to big impact. *MIS Quarterly*, 36(4): 1165–1188.

- Choudhury, P., Wang, D., Carlson, N.A., and Khanna, T. 2019. Machine learning approaches to facial and text analysis: discovering CEO oral communication styles. *Strategic Management Journal*, 40(11): 1705–1732.
- deGrazia, C.A.W., Pairolero, N.A., and Teodorescu, M.H. 2021. Examination incentives, learning, and patent office outcomes: the use of examiner’s amendments at the USPTO. *Research Policy*, 50(10): 104360.
- Devitt, A., and Ahmad, K. 2007. Sentiment polarity identification in financial news: a cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Vol. 7. Prague, Czech Republic, 1–8.
- Dinu, L., and Iuga, I. 2012. The Naïve-Bayes classifier in opinion mining: in search of the best feature set. In *International Conference on Intelligent Text Processing and Computational Linguistic* (pp. 556–567). Berlin: Springer.
- Figueroa-Armijos, M., Clark, B.B., and da Motta Veiga, S.P. 2022. Ethical perceptions of AI in hiring and organizational trust: the role of performance expectancy and social influence. *Journal of Business Ethics*, 1–19.
- Friedman, J., Hastie, T., and Tibshirani, R. 2001. *The Elements of Statistical Learning*. Berlin: Springer.
- Gans, J.S., Goldfarb, A., Lederman, M. 2017. Exit, tweets and loyalty. NBER Working Paper No. 23046. National Bureau of Economic Research, Cambridge, MA. <http://www.nber.org/papers/w23046>.
- Ghosh, S., Vinyals, O., Strophe, B., Roy, S., Dean, T., and Heck, L. 2016. Contextual LSTM (CLSTM) models for large scale NLP tasks. arXiv preprint: 1602.06291.
- Guzman, J., and Li, A. 2023. Measuring founding strategy. *Management Science*, 69(1): 101–118.
- Hall, B.H., Jaffe, A.B., and Trajtenberg, M. 2001. The NBER patent citation data file: lessons, insights and methodological tools. NBER Working Paper No. 8498. National Bureau of Economic Research, Cambridge, MA. <http://www.nber.org/papers/w8498>.
- Hall, P., Dean, J., Kabul, I.K., and Silva, J. 2014. An overview of machine learning with SAS Enterprise Miner. In *Proceedings of the SAS Global Forum 2014 Conference*. <https://support.sas.com/resources/papers/proceedings14/SAS313-2014.pdf>.
- Hayton, J.C., Allen, D.G., and Scarpello, V. 2004. Factor retention decisions in exploratory factor analysis: a tutorial on parallel analysis. *Organizational Research Methods*, 7(2): 191–205. <https://doi.org/10.1177/1094428104263675>.
- Hoberg, G., and Phillips, G. 2010. Product market synergies and competition in mergers and acquisitions: a text-based analysis. *Review of Financial Studies*, 23(10): 3773–3811.
- Hoffman, M., Bach, F.R., and Blei, D.M. 2010. Online learning for latent Dirichlet allocation. *Advances in Neural Information Processing Systems*, 23: 856–864.
- Howell, P.D., Martin, L.D., Salehian, H., Lee, C., Eastman, K.M., and Kim, J. 2016. Analyzing taste preferences from crowdsourced food entries. In *Proceedings of the 6th International Conference on Digital Health Conference*, 131–140.
- Hu, D., Zhao, J.L., Hua, Z., and Wong, M.C.S. 2012. Network-based modeling and analysis of systemic risk in banking systems. *MIS Quarterly*, 36(4): 1269–1291.
- Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, 168–177. <http://dl.acm.org/citation.cfm?id=1014052&picked=prox>.
- Huang, A. 2008. Similarity measures for text document clustering. In *Proceedings of the 6th New Zealand Computer Science Research Student Conference (NZCSRSC2008)*, Christchurch, New Zealand, 49–56.
- Jordan, M.I., and Mitchell, T.M. 2015. Machine learning: trends, perspectives, and prospects. *Science*, 349(6245): 255–260.
- Kaplan, S., and Vakili, K. (2015). The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal*, 36(10): 1435–1457.

- Kellogg, K.C., Valentine, M.A., and Christin, A. 2020. Algorithms at work: the new contested terrain of control. *Academy of Management Annals*, 14(1): 366–410.
- Kotu, V., and Deshpande, B. 2014. *Predictive Analytics and Data Mining: Concepts and Practice with Rapidminer*. Waltham, MA: Morgan Kaufmann.
- Kučera, H., and Francis, W.N. 1967. *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press.
- Kuhn, J.M., and Teodorescu, M.H. 2021. The track one pilot program: who benefits from prioritized patent examination? *Strategic Entrepreneurship Journal*, 15(2): 185–208.
- Kuhn, J.M., and Thompson, N.C. 2019. How to measure and draw causal inferences with patent scope. *International Journal of the Economics of Business*, 26(1): 5–38.
- Lee, C., Shleifer, A., and Thaler, R. 1991. Investor sentiment and the closed-end fund puzzle. *Journal of Finance*, 46(1): 75–109.
- Leech, G. 1991. The state of the art in corpus linguistics. In J. Svartvik, K. Aijmer, and B. Altenberg (eds), *English Corpus Linguistics: Studies in Honour of Jan Svartvik* (pp. 8–29). London: Longman.
- Li, F. 2010a. Textual analysis of corporate disclosures: a survey of the literature. *Journal of Accounting Literature*, 29: 143–165.
- Li, F. 2010b. The information content of forward-looking statements in corporate filings—a naïve Bayesian machine learning approach. *Journal of Accounting Research*, 48(5): 1049–1102.
- Li, G.C., Lai, R., D’Amour, A., Doolin, D.M., Sun, Y., Torvik, V.I., Amy, Z.Y., and Fleming, L. 2014. Disambiguation and co-authorship networks of the US patent inventor database (1975–2010). *Research Policy*, 43(6): 941–955.
- Lichman, M. 2013. UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA. <http://archive.ics.uci.edu/ml>.
- Liu, L.Y., Jiang, T.J., and Zhang, L. 2016. Hashtag recommendation with topical attention-based LSTM. In *Proceedings of the 26th International Conference on Computational Linguistics*, Osaka, Japan, 943–952.
- Loughran, T., and McDonald, B. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1): 35–65.
- Loughran, T., and McDonald, B. 2014. Measuring readability in financial disclosures. *Journal of Finance*, 69(4): 1643–1671.
- Louridas, P., and Ebert, C. 2017. Machine learning. *Computing Edge*, April: 8–13.
- Lugmayr, A. 2013. Predicting the future of investor sentiment with social media in stock exchange investments: a basic framework for the DAX Performance Index. In M. Friedrichsen and W. Mühl-Benninghaus (eds), *Handbook of Social Media Management* (pp. 565–589). Berlin and Heidelberg: Springer. http://dx.doi.org/10.1007/978-3-642-28897-5_33.
- Manning, C.D., Raghavan, P., and Schütze, H. 2008. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.
- Manning, C.D., and Schütze, H. 1999. Topics in information retrieval. In C.D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing* (pp. 539–554). Cambridge, MA: MIT Press.
- Marco, A.C., Sarnoff, J.D., and deGrazia, C.A.W. 2019. Patent claims and patent scope. *Research Policy*, 48(9): 103790.
- Marcus, M.P., Marcinkiewicz, M.A., and Santorini, B. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2): 313–330.
- Martin, K. 2019. Designing ethical algorithms. *MIS Quarterly Executive*, 18(2), Article 5.
- Martinez, G.L., Raffo, J., and Saito, K. 2016. Identifying the gender of PCT inventors. Economic Research Working Paper No. 33, *WIPO Economics and Statistics Series*, World Intellectual Property Organization, November.

- Melville, P., Gryc, W., and Lawrence, R.D. 2009. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Paris, France, 1275–1284.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26. Neurips conference proceedings, https://papers.nips.cc/paper_files/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.
- Monath, N., Madhavan, S., DiPietro, C., McCallum, A., and Jones, C. 2021. Disambiguating patent inventors, assignees, and their locations in PatentsView. American Institutes for Research. www.air.org.
- Morse, L., Teodorescu, M.H.M., Awwad, Y., and Kane, G.C. 2022. Do the ends justify the means? Variation in the distributive and procedural fairness of machine learning algorithms. *Journal of Business Ethics*, 181: 1083–1095.
- Pairolero, N., Toole, A., DeGrazia, C., Teodorescu, M.H., and Pappas, P.A. 2022. Closing the gender gap in patenting: evidence from a randomized control trial at the USPTO. In *Academy of Management Best Paper Proceedings*, 2022(1): 14401.
- Piryani, R., Madhavi, D., and Singh, V.K. 2017. Analytical mapping of opinion mining and sentiment analysis research during 2000–2015. *Information Processing and Management*, 53(1): 122–150.
- Porter, M.F. 1980. An algorithm for suffix stripping. *Program*, 14(3): 130–137.
- Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., and Welling, M. 2008. Fast collapsed Gibbs sampling for latent Dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, Las Vegas, Nevada, 569–577.
- Pustejovsky, J., and Stubbs, A. 2012. Corpus analytics. In J. Pustejovsky and A. Stubbs (eds), *Natural Language Annotation for Machine Learning* (pp. 53–65). Sebastopol, CA: O'Reilly Media.
- Ravi, K., and Ravi, V. 2015. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*, 89: 14–46.
- Reimers, N., and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. arXiv preprint arXiv:2004.09813.
- Ruckman, K., and McCarthy, I. 2017. Why do some patents get licensed while others do not? *Industrial and Corporate Change*, 26(4): 667–688.
- Salton, G., and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5): 513–523.
- Sardeshmukh, S.R., and Vandenberg, R.J. 2017. Integrating moderation and mediation: a structural equation modeling approach. *Organizational Research Methods*, 20(4): 721–745. <https://doi.org/10.1177/1094428115621609>
- Savage, N. 2012. Gaining wisdom from crowds. *Communications of the ACM*, 55(3): 13–15.
- Srivastava, S.B., Goldberg, A., Manian, V.G., and Potts, C. 2017. Enculturation trajectories: language, cultural adaptation, and individual outcomes in organizations. *Management Science*. doi:10.1287/mnsc.2016.2671.
- Struhl, S. 2015. In the mood for sentiment. In Struhl, S., *Practical Text Analytics: Interpreting Text and Unstructured Data for Business Intelligence* (pp. 120–143). London: Kogan Page Publishers.
- Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2): 267–307.
- Tan, T.Z., Quek, C., and Ng, G.S. 2007. Biological brain-inspired genetic complementary learning for stock market and bank failure prediction. *Computational Intelligence*, 23(2): 236–261.

- Tarafdar, M., Teodorescu, M., Tanriverdi, H., Robert, L., and Morse, L. 2020. Seeking ethical use of AI algorithms: challenges and mitigations, *Forty-First International Conference on Information Systems*, India.
- Teh, Y., Jordan, M., Beal, M., and Blei, D. 2007. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581.
- Teodorescu, M.H. 2018. The need for speed: uncertainty reduction in patenting and effects on startups. In *Academy of Management Proceedings*, 2018(1): 10977.
- Teodorescu, M.H.M., Choudhury, P., and Khanna, T. 2022a. Role of context in knowledge flows: host country versus headquarters as sources of MNC subsidiary knowledge inheritance. *Global Strategy Journal*, 12(4): 658–678.
- Teodorescu, M.H., Morse, L., Awwad, Y., and Kane, G.C. 2022b. Do the ends justify the means? Variation in the distributive and procedural fairness of machine learning algorithms. *Journal of Business Ethics*, 181: 1083–1095.
- Teodorescu, M.H., Ordabayeva, N., Kokkodis, M., Unnam, A., and Aggarwal, V. 2022c. Determining systematic differences in human graders for machine learning-based automated hiring. Brookings Institution Center on Regulation and Markets Working Paper, June.
- Teodorescu, M.H., and Yao, X. 2021. Machine learning fairness is computationally difficult and algorithmically unsatisfactorily solved. In *2021 IEEE High Performance Extreme Computing Conference (HPEC)*, IEEE, September, 1–8.
- Tetlock, P.C. 2007. Giving content to investor sentiment: the role of media in the stock market. *Journal of Finance*, 62(3): 1139–1168.
- Thompson, N.C., and Kuhn, J.M. 2020. Does winning a patent race lead to more follow-on innovation? *Journal of Legal Analysis*, 12: 183–220.
- Tonidandel, S., King, E.B., and Cortina, J.M. (2018). Big data methods: leveraging modern data analytic techniques to build organizational science. *Organizational Research Methods*, 21(3): 525–547. <https://doi.org/10.1177/1094428116677299>.
- Toole, A., Jones, C., and Madhavan, S. 2021. PatentsView: an open data platform to advance science and technology policy. USPTO Economic Working Paper No. 2021-1, SSRN Working Paper. <https://ssrn.com/abstract=3874213>.
- Trajtenberg, M., Shiff, G., and Melamed, R. 2006. The “names game”: harnessing inventors’ patent data for economic research. NBER Working Paper No. w12479. National Bureau of Economic Research, Cambridge, MA. <http://www.nber.org/papers/w12479>.
- Van den Broek, E., Sergeeva, A., and Huysman, M. 2021. When the machine meets the expert: an ethnography of developing AI for hiring. *MIS Quarterly*, 45(3).
- Verma, S., and Rubin, J. (2018, May). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness*.
- Wang, C., Paisley, J., and Blei, D.M. 2011. Online variational inference for the Hierarchical Dirichlet Process. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, Fort Lauderdale, Florida, 752–760.
- Whalen, R., Lungeanu, A., DeChurch, L., and Contractor, N. (2020). Patent similarity data and innovation metrics. *Journal of Empirical Legal Studies*, 17(3): 615–639.
- Willett, P. 1988. Recent trends in hierarchic document clustering: a critical review. *Information Processing and Management*, 24(5): 577–597.
- Williams, C., and Lee, S.H. 2009. Resource allocations, knowledge network characteristics and entrepreneurial orientation of multinational corporations. *Research Policy*, 38(8): 1376–1387.
- Xiao, H., and Tibor, T. 2010. October. Efficient collapsed Gibbs sampling for latent Dirichlet allocation. In *Proceedings of 2nd Asian Conference on Machine Learning*, Tokyo, Japan, 63–78.
- Younge, K.A., and Kuhn, J.M. 2016. Patent-to-patent similarity: a vector space model. Available at SSRN https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2709238.

- Yu, Y., Duan, W., and Cao, Q. 2013. The impact of social and conventional media on firm equity value: a sentiment analysis approach. *Decision Support Systems*, 55(4): 919–926.

4. Captains don't navigate with a keyboard: developing AI for naturalistic decision-making

Adrian Bumann

DECISION-MAKING IN THE WILD

In November 2014, a collision occurred between two cargo ships, the chemical tanker *Kraslava* and general cargo ship *Atlantic Lady*, in the Drogden Channel, a narrow and heavily trafficked shipping lane between Denmark and Sweden. Despite no harm to human life or the environment, both ships suffered significant damage to their structures. Weather conditions on that day were not optimal, with thick fog limiting visibility to less than 100 metres and southerly winds of 30 km/h, but were not uncommon during Scandinavian winters. Both ships' bridges were manned with experienced navigators who had sailed through the Drogden Channel multiple times before. Due to the poor visibility, both bridge teams heavily relied on electronic equipment such as the electronic chart system (ECDIS) to navigate.

As *Kraslava* proceeded southbound towards the channel exit, the bridge team observed *Atlantic Lady* on their ECDIS approaching from their port side, assuming that *Atlantic Lady* would turn to starboard after passing a buoy and pass *Kraslava*'s port side. This was common navigational practice and would have been a safe maneuver. A few minutes later, *Kraslava*'s captain expressed concerns that *Atlantic Lady* still had not turned. Meanwhile on *Atlantic Lady*, the chief officer reported to the captain that they were now clear of the buoy and ready to turn. Seconds later, both ships came within visible range, and it became apparent that *Atlantic Lady* was crossing ahead of *Kraslava*'s bow. Merely 20 seconds after *Kraslava*'s captain voiced his concerns, the ships collided.

After the incident, both bridge teams were perplexed about how the collision could have transpired. Both ships' captains assumed that they were positioned on the outer perimeter of the channel and that the other ship was at fault for veering onto the wrong side of the lane. In actuality, the collision occurred in the middle of the channel. The Danish investigating body noted that the collision "was the result of several factors that coincided within a short timeframe that created a risk of collision, which was not recognized by the bridge teams on either ship until within a minute of the collision" (DMAIB, 2015, p. 27). These factors included the fog, which prevented navigators from correctly assessing their position in relation to the channel buoys; limitations of electronic instruments, where the ECDIS would have had to be zoomed in to such an extent to provide accurate position information that the display would have been

rendered useless for determining forthcoming navigation; and a generally challenging turning maneuver. While none of these factors constituted a significant risk in themselves, in conjunction they created a small margin between success and failure.

INTRODUCTION

Naturalistic decision-making refers to the process by which people make decisions in complex and dynamic real-world environments. The collision between *Kraslava* and *Atlantic Lady* provides a vivid example of some of the unique challenges that decision-makers face, including high time pressure, uncertainty, and reliance on domain knowledge and electronic support systems (Hutchins, 1995). Misjudgments in such contexts can have severe consequences. As a former seagoing nautical officer, I have witnessed multiple similar situations in which minor cognitive inconsistencies can accumulate, resulting in a potentially hazardous situation, although fortunately none as catastrophic as the aforementioned collision.

Advances in data science hold promise to increasingly help with poorly structured tasks in complex situations by incorporating multiple sources of information (Berente et al., 2021). In recent years, artificial intelligence (AI) systems have been increasingly employed to support decision-making in various fields, from human resources (HR) processes (van den Broek et al., 2021), to pizza production (Domino's, 2017), to detecting fashion trends (Shrestha et al., 2021). In many cases, the development of these systems has been guided by the assumption that advanced AI models can capture all relevant parameters within organizationally bounded processes, and that decision-making can be modeled as a rational process, with decision-makers carefully weighing all available options based on clearly defined criteria.

However, these assumptions break down in naturalistic decision-making contexts, where the decision-making process is much more complex and dynamic. In these situations, decision-makers must rely on their intuition and a variety of external cues to make decisions rapidly and under pressure. While challenges for AI development from heightened contextual complexity have been outlined conceptually and highlighted as future research avenues (Benbya et al., 2021; Berente et al., 2021), there is a paucity of research that explains how such challenges unfold in empirical settings. Further exploration of such challenges is crucial, because the increasing reliance on electronic support systems necessitates that these systems are well designed. In maritime navigation, for example, the vast majority of accidents are attributed to human factors (Wróbel, 2021), with many being caused by overreliance on or misinterpretation of information provided by electronic equipment. Thus, it is essential to understand the unique characteristics of naturalistic environments, and the need for operators to develop AI systems that are technically robust and supportive of decision-making tasks.

Hence, the aim of the chapter is to investigate challenges in developing AI systems for naturalistic decision-making, and how developers mitigate these challenges. To help with that endeavor, this chapter reports on a two-year longitudinal study of

the development of an AI-based decision support system for maritime navigation, including trials with experienced nautical officers in simulated and real-life environments. The study contributes both to information systems literature following recent calls for phenomenon-based examination of emerging challenges when expanding the use domains of AI (Berente et al., 2021), and to practitioners developing AI decision-support for naturalistic environments.

BACKGROUND

Naturalistic Decision-Making

The decision-making process of humans has been a focus of research for many years (Weick, 1993). That research has outlined two contrasting decision-making paradigms, classical and naturalistic. Classical decision-making is concerned with providing prescriptive guidelines for clearly defined processes. This paradigm suggests that humans should analyze available information, identify and assess a range of alternative actions, and choose the most desirable outcome (Klein et al., 1997). However, such decision-making is difficult to apply in naturalistic environments, such as maritime navigation, aviation, firefighting, military operations, or medical surgery. These settings are generally characterized by unstable conditions, high uncertainty, vague goals, limited time, and high stakes (Lipshitz et al., 2006). As shown in the example of *Kraslava* and *Atlantic Lady*, classical decision-making is complicated by limited and ambiguous information and short time windows.

How do operators in naturalistic environment then make decisions? Put simply, they typically do not try to find the most optimal course of action, but rather a satisfactory one. Naturalistic environments exist in real time, meaning that the situation will change even if no action is taken. For instance, Klein et al. (1997) showed that firefighters on average only consider two alternative options before making a decision. There is no point in weighing all available strategies how to extinguish a fire if the fire will continue to spread in the meantime. Thus, naturalistic decision-making relies heavily on the operator's expertise and intuition to determine what constitutes a viable and appropriate action. Experts in a particular domain develop mental models, which are cognitive schemata that represent their knowledge and experience in that domain. These mental models allow experts to quickly recognize patterns and make decisions based on their experience, intuition, and contextual information. The decision-making process is often automatic and based on the recognition of similarities between the current situation and previously encountered situations (Schraagen, 2008).

Naturalistic decision-making takes place in complex sociotechnical systems (Hutchins, 1995) and is increasingly supported with electronic aids. Persons stepping onto a ship's bridge for the first time often remark that "this looks just like a spaceship," due to the large variety of computer screens and sensor indicators that provide information about the ship's dynamic state and surroundings. Generally, such aids

are useful in improving situational awareness, as they allow operators to triangulate real-world information (observing a ship visually) with electronically generated information (the corresponding radar signal) (Aylward, 2022). However, this can also have negative effects. Overreliance on electronic systems can lead to complacency and a lack of situational awareness (Zhang et al., 2021). From my experience, it is particularly easy for inexperienced navigators to stare obsessively at the radar screen to make sense of their situation, while forgetting to simply look out of the window. If electronic information is presented in a confusing or uncommon format, operators may be overwhelmed and find it difficult to prioritize and make sense of all the data. Finally, electronic systems can sometimes fail or provide incorrect information, which can lead to poor decision-making if operators do not recognize the problem (Serman and Sweeney, 2004).

Therefore, it is essential that support systems for naturalistic decision-making are reliable, easily interpretable, and useful to the specific situation. In the case of poorly designed systems, understanding their consequences and the development process leading up to them can help in adapting suitable organizational structures and processes (Lipshitz et al., 2006). For instance, certain models of the Boeing 737 MAX included an electronic flight stabilization feature that sometimes incorrectly assumed a stall condition, and consequentially pushed the plane's nose down. Investigations showed that pilots were neither fully aware of that feature, nor able to counteract it, leading to two fatal plane crashes and subsequent organizational reforms within Boeing (NTSB, 2019).

Developing AI for Naturalistic Decision-Making

Recent advances in AI and the generation of large amounts of data have led to the evolution of decision-support systems from rule-based expert systems to those that can guide ambiguous and unstructured processes (Gupta et al., 2022). Empirical research has primarily focused on the development of AI applications within the boundaries of organizational processes or virtual settings (Seidel et al., 2020; van den Broek et al., 2021). While studies exploring AI development for naturalistic settings are still scarce (Hodges et al., 2022), two general challenges can be derived from pertinent literature.

First, developers of AI decision-support systems face limitations in accounting for all possible external factors that may influence decision-making in naturalistic contexts, due to their unbounded and dynamic nature (Spurgin and Stupples, 2017). This is a challenge that even non-digital decision-support tools like checklists face, as evidenced by Chesley Sullenberger's widely praised decision to rely on his intuition, rather than follow an engine failure checklist designed for high-altitude situations, when he successfully landed a powerless airplane on the Hudson River (NTSB, 2010). For developers of AI systems, it is thus essential to consider both the system's level of envisioned self-sufficiency, that is, the range of conditions the system can handle, and self-directedness, that is, the range of conditions which the system can autonomously manage (van den Broek et al., 2020).

The challenge of responding to contextual complexity is exemplified by Ashby's Law of Requisite Variety, a central principle in cybernetics which states that a regulator, such as an AI system, must possess a level of variety exceeding that of the external system being controlled (Cybulski and Scheepers, 2021). For example, an autonomous car may perform safely in a test environment with limited inputs, but it may struggle when faced with a greater variety of inputs, such as those encountered in heavy rain or when avoiding stray animals. Similarly, a ship moves during its voyage through various operational states with differing levels of complexity. For instance, while navigating narrow fairways or berthing the ship, navigators need to account for a higher number of parameters compared to when sailing on open seas. Given that high variation in operational, technical, and environmental complexity, a large variation of autonomy states is possible with varying degrees of human support and/or AI support (van den Broek et al., 2020). To address this challenge, increasingly advanced AI methods such as reinforcement learning hold promise in allowing decision-support systems to interact with their environment and account for dynamic or uncertain information, thus incorporating greater external variety (Grantner et al., 2016). For instance, the AI system presented later in this chapter could account for dynamic changes in wind conditions to tune future predictions.

A second challenge for developing AI decision-support in naturalistic environments lies in heightened human-machine interface (HMI) requirements. Given their complexity, most decision processes in naturalistic environments are more likely to be augmented, rather than automated, by algorithmic decision systems and thus still require a human-in-the-loop (Benbya et al., 2021). Although algorithms can outperform human decision-making in certain situations—for example, routine tasks or tasks subject to human bias (Kahneman and Riepe, 1998)—expert operators excel in intuitive perception skills, such as automatically recognizing patterns and anomalies in chaotic environments (Shrestha et al., 2021). However, they might not be able to apply their expertise when new factors, such as electronic support systems, are introduced (Simkute et al., 2021). For instance, a nautical officer might intuitively detect a change in wind speed by observing wave height, but might fail to recognize that when the wind speed is displayed on an ECDIS if he is unfamiliar with the interface.

Therefore, it is important for decision-makers to be able to scrutinize AI-produced information, particularly in high-stakes situations (Tacker and Silvia, 1991). However, many AI-based systems are complex and opaque, which can make it difficult for users without extensive technical knowledge to understand how the system works and why it is making certain recommendations. This black box nature of AI-based models is commonly identified as a threat to user trust and decision accountability (Zhang et al., 2021). Subsequently, developers must consider how they design AI tools that integrate seamlessly in complex sociotechnical environments (Martelaro and Ju, 2018). To tackle this challenge, researchers are exploring methods to develop transparent and interpretable AI systems, such as by providing visualizations or explanations of the decision-making process. The design principle of explainable AI (XAI) has gained significant attention lately, advocating for AI systems to offer users explanations of how outputs are derived. For instance, AI models used to interpret

medical imaging such as X-rays can provide explanations by visually highlighting those image features that contributed to a diagnosis (Lebovitz et al., 2021). However, the explainability of an AI system is not solely determined by its design characteristics; it also emerges from how users interact with the system and how well the system aligns with users' cognitive schemata (Waardenburg et al., 2022), making it thus important to investigate both developers' technical design choices and how they interact with the intended environment and operators within.

The use of AI in maritime navigation has mainly focused on two areas: automating simple tasks, such as using an autopilot to maintain a steady heading, and supporting strategic decision-making, such as optimizing route planning. However, development of AI systems that can augment or automate complex navigation tasks is still in its nascent stage, for several reasons (Munim et al., 2020). For one, it is much more complex to model a ship's behavior in a predictive algorithm than, for example, that of a car, due to environmental conditions, hydrodynamic effects, and interaction with other ships (Perera, 2017). Moreover, traffic regulation at sea is less constrained than that on land, and includes concepts such as "good seamanship," which cannot easily be codified in algorithms (MacKinnon et al., 2020). Finally, most AI applications are tested in ship simulators due to cost considerations. However, this approach makes it harder to account for real-world issues such as erroneous onboard sensors, which can lead to performance discrepancies between simulation and actual deployment (Perera, 2017).

METHOD

This chapter draws upon a longitudinal, interpretative case study (Walsham, 1995) of Neptune, a publicly funded, two-year innovation consortium formed to develop an AI-based ship predictor for maritime navigation. Neptune consisted of five Northern European public and private organizations from maritime, global satellite navigation systems (GNSS) and telecommunications industries, including a multinational corporation with 15 000+ employees.

Ship predictors are crucial for navigators as they provide a future trajectory of a ship's movement on their electronic chart system (ECDIS), usually for time spans of 1–12 minutes (Figure 4.1). Current ship predictors rely on dead reckoning, which uses past information on course, speed, and rate of turn. Such predictions can be misinterpreted, and require an experienced human operator to triangulate with other information. More advanced predictors that account for a ship's individual maneuvering characteristics or loading conditions exist, but are costly and far from widespread adoption. Neptune's envisioned AI predictor would use reinforcement learning to increase prediction accuracy by: (1) integrating additional shipboard sensors and individual ship parameters in the predictive model; and (2) assessing accuracy of past predictions to train the model for future predictions. Compared to existing approaches, the AI capability would lower development costs and allow for using the same algorithm across different ships as it would self-train to their

respective characteristics. Neptune was selected as a suitable case for this research because of its objective to develop commercially viable AI applications, which led to collaboration of diverse experts, adherence to high-quality standards, and a strong focus on user needs.

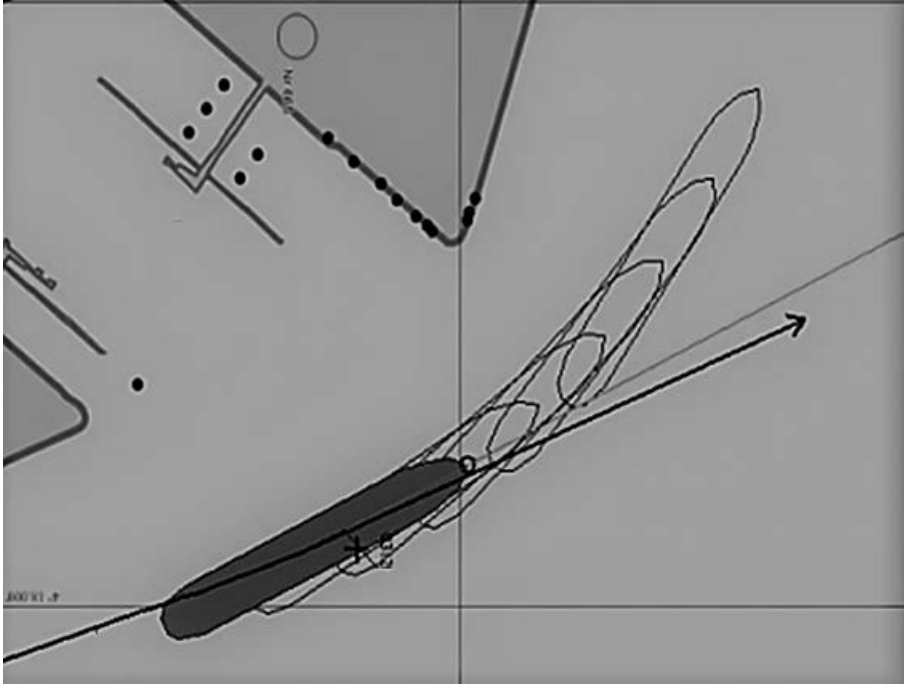


Figure 4.1 Example display of ship predictor

Starting in October 2020, I studied Neptune’s development process over the course of 20 months, collecting various data sources to allow triangulation. First, participant observation included attending 65 project meetings, three sea trials, and a two-day user study in a professional bridge simulator. During the sea trials, I spent a total of ten days on board the test ships, observing the practical activities related to installation, testing, and evaluation of the AI predictor. During the user study, I observed how experienced captains executed various maneuvers with and without the AI predictor, and discussed their experience with developers. My background as a nautical officer allowed me to embed myself as a “native” in the study context (Brannick and Coghlan, 2007), being able to speak the language of maritime experts, understand domain-specific processes, and assist in minor tasks such as navigation. Second, I conducted 17 semi-structured interviews with 14 respondents across all Neptune work packages. This included data scientists, maritime experts, GNSS experts, telecommunication experts, and two project managers. Finally, extensive project

documentation, such as internal reports, presentations, and newsletters, provided technical insights and allowed comparisons of envisioned designs against final outcomes. The collected data were analyzed in an iterative and systematic manner using data-driven thematic analysis. First, a broad overview of the development process was constructed using the software Aeon Timeline to understand causal linkages between different observations. Second, field notes, interview transcripts, and relevant documentation were read to identify and corroborate challenges and mitigating actions. Various themes were developed, organized into first-order codes supported by various data sources, and then mapped across the three stages of AI development outlined by Shrestha et al. (2021).

CHALLENGES AND MITIGATION STRATEGIES FOR DEVELOPING AI FOR NATURALISTIC DECISION-MAKING

Ensuring Reliable Input Data

High-quality input data is an essential element in developing any AI application, as illustrated by the adage “garbage in, garbage out.” Thus, the Neptune team faced the initial challenge of determining which data sources to use and how to ensure their quality. Existing dead reckoning ship predictors extrapolate a ship’s future trajectory based on a limited number of past data points. Therefore, one of Neptune’s primary objectives was to include more data points to enhance prediction accuracy, such as wind speed, propulsion, and rudder sensors. This often presented a challenge between the envisioned complexity and feasibility.

An example of this is when the team discussed whether the AI predictor should incorporate a complex hydrodynamic effect that occurs when a ship navigates in narrow waterways, known as the “bank effect.” This phenomenon is notoriously difficult to detect and counteract, and can contribute to maritime accidents, as seen in the grounding of the *Ever Given* in 2021 (George, 2021). Given its complexity and importance for maritime safety, modelling this effect in Neptune’s predictor was initially deemed a suitable goal. However, developers soon realized that the necessary data were not easily accessible and would require complex integration of sea chart information in the AI model. Although such integration would have been technically feasible, it was unclear whether data quality could be always maintained. A data scientist reflected that: “we could calculate hydrodynamic effects against fixed objects shown in the ECDIS, it would require the predictor to have information about the chart contours around it ... And it was very clear by [ECDIS manufacturer] that such integration would be a big issue.” As a result, the team chose to focus on incorporating data points that were considered essential or available in the common maritime National Marine Electronics Association (NMEA) data standard.

In that regard, key activities were three sea trials where Neptune developers could test technical integration onboard three different ships. However, the variety and complexity of onboard legacy systems presented various challenges to ensure

reliable input data. For instance, developers initially considered including various data points related to the rudder movement, to improve prediction accuracy. On large ships, a rudder is typically controlled by hydraulic pumps. Depending on the condition and number of pumps, it can take almost a minute from the time the navigator sets a new rudder angle until the rudder reaches its desired position. Including both “set rudder angle” and “current rudder angle” would thus enable the AI predictor to be more adaptive to the navigator’s intentions. However, developers eventually decided to omit the “set rudder angle” because more simulated processes might result in more potential error sources if they did not match with physical reality. A data scientist noted that:

the steering pumps might not be digitized. Or you assume it takes 40s from hard-port to hard-starboard, but in reality, it takes 60s. That’s yet another coefficient to keep track of. The possibility of being more accurate, it’s there. But then the complexity to reach that accuracy is higher ... You don’t want too many or too few coefficients.

Similarly, many sensors were found to not be perfectly calibrated, meaning for instance that while the physical rudder was set midships, the rudder sensor would indicate a few degrees to starboard. Such poor calibration is not uncommon on merchant ships and does not pose problems for traditional dead reckoning predictors. For Neptune’s AI predictor, however, this resulted in the predictor indicating a constant drift to starboard, even if the ship actually sailed in a straight line. Therefore, an early adaptation was to implement functionalities in the AI predictor to automatically detect and self-correct such value biases.

Finally, the sea trials highlighted the beneficial role of maritime expertise in detecting technical issues that were not immediately apparent to other actors. For instance, the predictor value indicating the ship’s direction was calculated using an algorithm developed for cars. The algorithm assumed that a car’s heading was equal to the orientation of its axle, which neglected the possibility of sideways drift experienced by ships. A GNSS expert commented that this “was an easy error to fix. But it was [maritime expert] who detected it visually on the ECDIS display, we did not detect it in the code. It was really his ‘gut instinct’.” Similarly, during one sea trial, developers experienced periodic outages while receiving Global Positioning System (GPS) data on a dedicated radio frequency. Initially, GNSS experts attributed the loss of signals to ships sailing too close to islands, which would obstruct radio waves. However, the problem persisted even when sailing in open waters. Eventually, a maritime expert discovered that these outages coincided with local coastal weather forecasts transmitted on the same frequency, which helped to identify and solve the issue. Reflecting on the importance of testing in physical onboard environments, Neptune’s project manager noted:

I’m very happy about these sea trials, as they get us closer to the actual usage and the environment. There are so many things that you don’t really understand until you actually experience them or get a feeling for something that could be a problem that we didn’t think about.

Refining AI System Functionality

As Neptune tested and refined its AI predictor, considerable effort went into ensuring that the functionality matched user needs and provided support in situations where existing systems were found lacking. The wide range of operational states in a ship's voyage meant many potential scenarios where an AI predictor could provide useful knowledge output. Developers noted that navigation in confined waters would benefit the most from increased prediction accuracy, since open waters typically provide enough room to maneuver, and existing ship predictors provided sufficient information. "[In open waters], you won't notice even a deviation of 2m. Because you wouldn't want to be in a situation where being off by 1m puts you in a dangerous situation. It's when you're close to obstacles or dock the ship, then [improved prediction] becomes useful" (maritime expert). Instead, during development and sea trials, it became clear that the predictor functionality was particularly suited for situations where short-term predictions of less than three minutes and high levels of accuracy were beneficial, allowing Neptune to highlight two suitable situations: "We identified two main things [the predictor] can help [operators with]: to avoid close-quarter [collision] situations and handling the ship in maneuvering in narrow waters" (maritime expert). Subsequently, Neptune focused its test activities on situations where a ship would dock, or where two ships would navigate close to each other, to evaluate its performance.

Data scientists noted the added challenge of developing a predictor designed for real-life use. "I developed similar predictors for [simulated environments] where you have control over all variables. That's not the case at sea. You cannot tweak the ship's behavior in the code" (data scientist). To facilitate testing technical functionalities, data scientists developed a browser-based simulation software dubbed EcdiSim. EcdiSim resembled an actual ECDIS and included a small ship symbol that could be steered with rudimentary input controls. Through this software, data scientists could test how the predictor would detect and deal with artificially generated sensor data disturbances, such as high fluctuations or values outside pre-defined ranges. In addition, this allowed maritime experts to play with a prototype and provide input on design choices or detect technical issues while testing. A data scientist recalled that while using EcdiSim, Neptune actors with:

experience in ship handling were like "Yeah, this doesn't behave as expected in this situation, this is weird, etc." And then we went into the code and tried to figure out what happened. And then you find like a wrong minus sign that was causing an issue, or weird ship behavior at low velocities, stuff like that.

The usefulness of EcdiSim was demonstrated in addressing an issue known as the "windshield wiper effect," where the AI predictor was excessively sensitive to input fluctuations. This effect occurred when sudden wind gusts or minor rudder adjustments caused significant, sudden changes in the predicted ship trajectory, leading to the displayed ship symbol moving erratically from side to side. While

high sensitivity enabled the early prediction of changes in ship movements, excessive fluctuation was deemed a limitation for readability and a potential risk for user acceptance. “If you’ll show [a captain] a predictor moving like a wiper, they’ll say, ‘This is bloody stupid. Why is it showing that we’re moving this way?’ Because they don’t ask how it’s actually working under the surface. And that makes them lose trust” (maritime expert). Consequentially, data scientists implemented filters in the algorithm to reduce input sensitivity and increase input latency, making the predictor move more sluggishly. This highlighted the challenge of finding the right balance between desired sensibility and interpretability; that is, determining what constituted a “responsive enough” prediction:

We involved [maritime experts] to get a suitable [filter balance]. How many seconds in the past do we average? After testing several iterations and different ships [in EcdiSim], we found approximately 7 seconds a reasonable value ... It’s tricky. You want to avoid the wiper effect, but you still want to be able to set rudder hard to port and see a change in the prediction now, not with 20 seconds delay (data scientist).

Achieving Contextual Fit

The final stage in Neptune’s development process involved deploying the AI predictor to experienced captains and assessing whether it actually supported their decision-making as intended. These captains had not been involved in prior development stages, and were deemed unbiased in their assessment. Notably, this highlighted the challenge in achieving contextual fit between the AI system and user needs.

Initially, developers had planned to conduct these tests in EcdiSim, but developers noted that the software failed to properly capture the complexity of the captains’ work environment. Subsequently, Neptune booked two days in a professional bridge simulator that included a realistic mock-up of a ship’s bridge and control consoles. A data scientist reflected that while their self-developed:

EcdiSim isn’t bad, it’s a desktop simulator [without] all this other information that people use when they’re actually navigating. These captains are not used to drive a ship with mouse and keyboard ... So we decided “nope, didn’t work well, let’s do full bridge simulators” so the captains can handle the ship in a comfortable manner as they do in reality.

During those two days, Neptune conducted A/B tests, where captains would do specific maneuvers such as collision avoidance or docking with both traditional and Neptune’s predictor to compare differences.

Overall, captains reacted positively and found that the AI predictor helped their situational awareness in most situations. Notably, they appreciated the function to customize when ECDIS should automatically switch to the AI predictor, for instance when going below a certain speed or in the vicinity of another ship. It is already common practice during voyage planning to define such values for different voyage segments, and typically reflect a change of navigational requirements. In the open sea, navigators prefer to maintain a large distance from obstacles or other ships.

However, in river passages or heavy-traffic areas, such a strategy is not always practical, and navigators must reassess what safety margins are feasible. While the AI predictor could be also manually toggled, allowing to set custom settings with the same parameters used in voyage planning provided adaptability to different cognitive schemata (Simkute et al., 2021). The responsible HMI designer highlighted simplicity and adaptability as essential in his design philosophy, noting that:

the trickiest, but also one of the most fun things, is the simplification. A lot of systems have these engineer-designed HMIs that overwhelm the user with technical complexity. But for the user, it needs to be approachable, identifiable. [My approach is] to design an exploration path for users where they can use [the predictor] bare-bones first, and discover more advanced functionality, as they get more experienced with the tool.

However, captains also expressed concerns about understanding how the AI predictor generated its output, and when to trust it. A captain commented he did not “have to understand the data science details, but when I’m close to the pier and the predictor shows I’ll collide with it, I want to know if that’s correct or not.” While existing non-AI ship predictors too can potentially produce erroneous predictions, for example due to faulty sensors, their functionality is simple enough that navigators can detect these errors with proper training. In contrast, Neptune developers noted that the technical complexity of the AI predictor would make it more difficult to scrutinize its output. Their AI predictor was designed to continuously assess its own performance and display an accuracy estimate, indicating potential inaccuracies if past predictions were unreliable due to factors such as poor sensor input or strong wind. The user’s ability to intuitively detect potentially erroneous information output strongly depends on how HMI is designed. Since traditional predictors did not have this functionality, much discussion revolved around how to present such accuracy estimates to the navigators to avoid confusion, for instance by displaying a traffic light or error bars on screen. The HMI designer noted: “Captains just want to have a quick look and know what’s going on ... There is a lot of information on the ECDIS. So we want to absolutely avoid information overload.”

Finally, captains emphasized the importance of proper training to prevent exaggerated and risky assumptions about the AI predictor’s capabilities. While captains were able to ask developers questions during the testing phase, and felt that they had a good understanding of how to interpret the presented information, less experienced navigators might “hear ‘AI’ and think of the prediction as something absolute.” In collision avoidance situations, the use of the AI predictor’s key feature, in which two ships could exchange their predictors for better situational awareness, could be misunderstood. One captain noted that another navigator might see his shared predictor turning towards them and assume that a collision is imminent, even though the prediction was based only on the input data related to the ship’s state (wind, propulsion, rudder, and so on) and not on the captain’s intentions. Such misunderstandings could lead to unnecessary avoidance maneuvers and confusion.

Feedback from captains and Neptune's maritime experts helped towards an HMI design presenting the predictor output in an interpretable and useful manner. Both the HMI designer and data scientists reflected positively on the involvement of user perspectives instead of evaluating the AI system's usefulness from a purely technical perspective. "We could have cherry picked the top situations and showed really good [prediction] performance ... [But] we felt that with a quantitative approach, we could just twist the numbers in any direction. It was more valuable building on the [captains'] experience to answer, 'is this useful for [navigators]?''" (data scientist).

WHAT CAN BE LEARNED FROM THIS RESEARCH?

While algorithmic decision-support is becoming increasingly ubiquitous, most research has focused on applications in bounded organizational processes (Cybulski and Scheepers, 2021). However, Neptune's case demonstrates the importance of developers understanding the distinctive features of naturalistic environments, and the needs of operators to create AI systems that are both technically resilient and helpful in decision-making. Three noteworthy insights emerge from this study.

First, it is not necessary to account for all environmental parameters in AI systems in order to support naturalistic decision-making. Advances in data science allow AI systems to fully account for the variety, and thus automate processes within organizational confines, such as product quality control (Cybulski and Scheepers, 2021). In the words of Ashby's Law, these AI-based regulator systems possess enough variety to control the input variety of the environment for which they are designed. This chapter, however, illustrates the challenges in developing a regulator system, an AI decision-support system for maritime navigation, which can account for a variety of dynamic inputs in a naturalistic environment. Ultimately, the aggregate variety of Neptune's AI predictor remained below the requisite variety to fully control the phenomena, that is, to capture all parameters that influenced a ship's future trajectory. However, the aggregated input data and developed algorithm still provided output that significantly improved navigators' decision-making in specific situations.

Achieving high prediction accuracy was hindered by situating the predictor within a complex sociotechnical system, which consists of the ship as a highly modularized technical system, and human operators with limited cognitive capacity. Highlighting these constraints, MacKinnon et al. (2020) note that the "AI paradigm is difficult to apply when considering naturalist decision-making processes and largely ignores the challenges of spatial and temporal aspects typical of navigation in complex situations" (p. 430). Hence, contextual understanding of the data sources is crucial to mitigate constraints from the sociotechnical system, and to strike an appropriate balance between incorporating a wide range of environmental factors for improved prediction accuracy and ensuring the system's robustness against flawed data inputs (Cybulski and Scheepers, 2021). For instance, Neptune developers omitted certain onboard sensor data from their predictor. Although including more sensors would have been technically feasible, such as including values for both set and current rudder angle,

this could have produced erroneous predictions if the physical rudder movements did not match the coefficients modeled in the algorithm. Similarly, the responsiveness of the predictor was reduced to prevent confusion for navigators when presented with a highly fluctuating prediction, highlighting the heightened difficulty to scrutinize AI-produced information output (Lebovitz et al., 2021). While a fully autonomous ship might be able to account for such high responsiveness, and in fact might benefit from it, human operators in cognitively demanding environments require easily interpretable information. This suggests that particularly for high-risk situations, an “imperfect” but reliable AI system is better suited than a system that can manage more complexity but is more prone to errors.

Second, common design principles to make AI-produced information explainable only apply to a certain extent in the context of naturalistic decision-making. With the increasing prevalence of AI systems and growing awareness of the risks associated with unexplainable AI, AI developers often aim to enable users to understand the reasoning behind produced outputs and to make AI a “glass box, not a black box” (Rai, 2020). This is useful when users have time to scrutinize all available information; for example, when analyzing medical images. However, this may be counterproductive in situations with high time pressure, such as in the collision of *Kraslava* and *Atlantic Lady*, where the time frame to act was merely 20 seconds. Even if the AI prediction is technically more accurate than traditional systems, it may distract or lead to erroneous decisions if the output is not easily interpretable.

Thus, appropriate HMI design is crucial to avoid potential information overload. Rather than explaining underlying AI processes to the user, developers may aim for “unremarkable AI” (Yang et al., 2019), that is, providing reliable, easily interpretable information that integrates smoothly into the operators’ cognitive environment. In the case of potentially ambiguous output, it is better to allow switching automatically to robust non-AI systems. For instance, in the case of Neptune, nautical officers appreciated the function to toggle custom speed or collision risk thresholds for when the AI predictor should be enabled, because it allowed them to adapt the system to their own cognitive schemata (Simkute et al., 2021). In addition, good interpretability is important to build trust and good first impressions to ensure future adoption (Zhang et al., 2021).

Third, this chapter highlights the benefits of employing testing methods with different degrees of contextual immersion to integrate tacit domain knowledge in the AI development process. Prior research has highlighted the beneficial role of boundary objects, such as prototypes or mockups, to allow experts from different domains to translate knowledge (Perschina et al., 2019). This too was the case in Neptune’s desktop simulation EcdiSim, which allowed maritime experts to assess the general feel of the AI predictor. Particularly in the early development stages, this was useful to make key design choices, for instance determining easy-to-read display options, and a suitable balance between predictor latency and sensitivity.

However, this study also shows that it is equally important to test AI systems in environments that mirror the sociotechnical context of the end users. Since naturalistic decision-making relies heavily on the operators’ intuitive skills, testing in

a high-fidelity simulator, such as the full-mission bridge simulator used in Neptune's field tests, allowed for a more accurate assessment of how the AI predictor would perform in real-world situations. The simulator replicated the ship's physical environment and provided an immersive experience for the nautical officers, which allowed them to use their tacit knowledge to assess the system's performance methodically in different scenarios. This allowed for more effective integration of domain knowledge into the AI development process, which ultimately led to a better-performing system. While such physical full-mission simulators are most commonly available in aviation and maritime navigation, recent advances in virtual reality (VR) technology promise similar testing grounds for other naturalistic domains, such as firefighting or medical surgery (Hodges et al., 2022). Similarly, although Neptune's sea trials were primarily conducted to test technical integration, I noticed several occasions where the intuition of maritime experts aided in refining the AI system design; for instance, identifying faulty output from an algorithm that was initially developed for cars.

Moreover, testing in a realistic context also highlighted the need for effective human–AI interaction. In the field tests, the nautical officers were able to use the AI predictor in conjunction with their own judgment, and the system provided additional information to help them make informed decisions. However, the study also identified potential issues with overreliance on the AI predictor, and the need for clear communication between the system and the user.

In conclusion, employing testing methods with different degrees of contextual immersion proved crucial for developing effective AI systems that integrate tacit domain knowledge and work seamlessly with end users. Ultimately, the most crucial aspect of a new decision-support system is not whether it incorporates AI or other innovative technologies, but whether it “disappears into the hand” of its user and becomes an intuitive part of their decision-making process.

REFERENCES

- Aylward, K. (2022). *Towards an Understanding of the Consequences of Technology-Driven Decision Support for Maritime Navigation*. Doctoral Thesis, Chalmers University of Technology.
- Benbya, H., Pachidi, S., and Jarvenpaa, S.L. (2021). Special Issue Editorial: Artificial Intelligence in Organizations: Implications for Information Systems Research. *Journal of the Association for Information Systems*, 22(2), 281–303.
- Berente, N., Gu, B., Recker, J., and Santhanam, R. (2021). Managing Artificial Intelligence. *MIS Quarterly*, 45(3), 1433–1450.
- Brannick, T., and Coghlan, D. (2007). In Defense of Being “Native”: The Case for Insider Academic Research. *Organizational Research Methods*, 10(1), 59–74. <https://doi.org/10.1177/1094428106289253>
- Cybulski, J.L., and Scheepers, R. (2021). Data Science in Organizations: Conceptualizing its Breakthroughs and Blind Spots. *Journal of Information Technology*, 36(2), 154–175. <https://doi.org/10.1177/0268396220988539>.
- DMAIB (2015). *ATLANTIC LADY and KRASLAVA. Collision on 1 November 2014*. Danish Maritime Accident Investigation Board.

- Domino's (2017). *DOM Pizza Checker*. DOM Pizza Checker. <https://dom.dominos.co.nz>.
- George, R. (2021, April 3). Wind ... Or Worse: Was Pilot Error to Blame for the Suez Blockage? *The Guardian*. <https://www.theguardian.com/environment/2021/apr/03/wind-or-worse-was-pilot-error-to-blame-for-the-suez-blockage>.
- Grantner, J.L., Fuller, S.T., and Dombi, J. (2016). Fuzzy Automaton Model with Adaptive Inference Mechanism for Intelligent Decision Support Systems. *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2377–2384. <https://doi.org/10.1109/FUZZ-IEEE.2016.7737991>.
- Gupta, S., Modgil, S., Bhattacharyya, S., and Bose, I. (2022). Artificial Intelligence for Decision Support Systems in the Field of Operations Research: Review and Future Scope of Research. *Annals of Operations Research*, 308(1/2), 215–274. <https://doi.org/10.1007/s10479-020-03856-6>.
- Hodges, J.L., Lattimer, B.Y., and Champlin, V.L. (2022). The Role of Artificial Intelligence in Firefighting. In M. Naser and G. Corbett (eds), *Handbook of Cognitive and Autonomous Systems for Fire Resilient Infrastructures* (pp. 177–203). Springer International Publishing. https://doi.org/10.1007/978-3-030-98685-8_8.
- Hutchins, E. (1995). *Cognition in the Wild*. MIT Press.
- Kahneman, D., and Riepe, M.W. (1998). *Aspects of Investor Psychology*. 14.
- Klein, G.A., Calderwood, R., and Clinton-Cirocco, A. (1997). Rapid Decision Making on the Fire Ground. *Proceedings of the 30th Annual Human Factors Society Meeting*.
- Lebovitz, S., Levina, N., and Lifshitz-Assaf, H. (2021). Is AI Ground Truth really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What. *MIS Quarterly*, 45(3), 1501–1525.
- Lipshitz, R., Klein, G., and Carroll, J. S. (2006). Introduction to the Special Issue. Naturalistic Decision Making and Organizational Decision Making: Exploring the Intersections. *Organization Studies*, 27(7), 917–923. <https://doi.org/10.1177/0170840606065711>
- MacKinnon, S.N., Weber, R., Olinderson, F., and Lundh, M. (2020). Artificial Intelligence in Maritime Navigation: A Human Factors Perspective. In: N. Stanton (ed.), *Advances in Human Aspects of Transportation. AHFE 2020*. Advances in Intelligent Systems and Computing, vol. 1212. Springer. https://doi.org/10.1007/978-3-030-50943-9_54.
- Martelaro, N., and Ju, W. (2018). Cybernetics and the design of the user experience of AI systems. *Interactions*, 25(6), 38.
- Munim, Z.H., Dushenko, M., Jimenez, V.J., Shakil, M.H., and Imset, M. (2020). Big Data and Artificial Intelligence in the Maritime Industry: A Bibliometric Review and Future Research Directions. *Maritime Policy and Management*, 47(5), 577–597. <https://doi.org/10.1080/03088839.2020.1788731>.
- NTSB (2010). *Loss of Thrust in Both Engines After Encountering a Flock of Birds and Subsequent Ditching on the Hudson River, US Airways Flight 1549, Airbus A320-214, N106US, Weehawken, New Jersey, January 15*.
- NTSB (2019). *Assumptions Used in the Safety Assessment Process and the Effects of Multiple Alerts and Indications on Pilot Performance*.
- Perera, L.P. (2017). Navigation Vector Based Ship Maneuvering Prediction. *Ocean Engineering*, 138, 151–160. <https://doi.org/10.1016/j.oceaneng.2017.04.017>.
- Pershina, R., Soppe, B., and Thune, T.M. (2019). Bridging analog and digital expertise: Cross-domain collaboration and boundary-spanning tools in the creation of digital innovation. *Research Policy*, 48(9), 103819. <https://doi.org/10.1016/j.respol.2019.103819>
- Rai, A. (2020). Explainable AI: From Black Box to Glass Box. *Journal of the Academy of Marketing Science*, 48(1), 137–141. <https://doi.org/10.1007/s11747-019-00710-5>.
- Schraagen, J.M. (Ed.). (2008). *Naturalistic Decision Making and Macrocognition*. Ashgate.
- Seidel, S., Berente, N., Lindberg, A., Lyytinen, K., Martinez, B., and Nickerson, J. V. (2020). *Artificial Intelligence and Video Game Creation: A Framework for the New Logic of Autonomous Design*, 2(3), 32.

- Shrestha, Y.R., Krishna, V., and von Krogh, G. (2021). Augmenting Organizational Decision-Making with Deep Learning Algorithms: Principles, Promises, and Challenges. *Journal of Business Research*, 123, 588–603. <https://doi.org/10.1016/j.jbusres.2020.09.068>.
- Simkute, A., Luger, E., Jones, B., Evans, M., and Jones, R. (2021). Explainability for Experts: A Design Framework for Making Algorithms Supporting Expert Decisions more Explainable. *Journal of Responsible Technology*, 7–8, 100017. <https://doi.org/10.1016/j.jrt.2021.100017>.
- Spurgin, A.J., and Stupples, D.W. (2017). *Decision-Making In High Risk Organizations Under Stress Conditions*, 255.
- Sterman, J.D., and Sweeney, L.B. (2004). Managing Complex Dynamic Systems: Challenge and Opportunity for Naturalistic Decision-Making. In Henry Montgomery, Raanan Lipshitz and Berndt Brehmer (eds), *How Professionals Make Decisions*. CRC Press.
- Tacker, E.C., and Silvia, M.T. (1991). Decision making in complex environments under conditions of high cognitive loading: A personal expert systems approach. *Expert Systems with Applications*, 2(2), 121–127. [https://doi.org/10.1016/0957-4174\(91\)90109-R](https://doi.org/10.1016/0957-4174(91)90109-R).
- van den Broek, Griffioen, J.R. (Jaco), and van der Drift, M. (Monique). (2020). Meaningful Human Control in Autonomous Shipping: An Overview. *IOP Conference Series: Materials Science and Engineering*, 929(1), 012008. <https://doi.org/10.1088/1757-899X/929/1/012008>.
- van den Broek, Sergeeva, A., and Huysman, M. (2021). When the Machine Meets the Expert: An Ethnography of Developing AI for Hiring. *MIS Quarterly*, 45(3), 1557–1580.
- Waardenburg, L., Huysman, M., and Sergeeva, A.V. (2022). In the Land of the Blind, the One-Eyed Man Is King: Knowledge Brokerage in the Age of Learning Algorithms. *Organization Science*, 33(1), 59–82. <https://doi.org/10.1287/orsc.2021.1544>.
- Walsham, G. (1995). Interpretive Case Studies in IS Research: Nature and Method. *European Journal of Information Systems*, 4(2), 74–81. <https://doi.org/10.1057/ejis.1995.9>.
- Weick, K.E. (1993). The Collapse of Sensemaking in Organizations: The Mann Gulch Disaster. *Administrative Science Quarterly*, 38(4), 628. <https://doi.org/10.2307/2393339>.
- Wróbel, K. (2021). Searching for the Origins of the Myth: 80% Human Error Impact on Maritime Safety. *Reliability Engineering and System Safety*, 216, 107942. <https://doi.org/10.1016/j.ress.2021.107942>.
- Yang, Q., Steinfeld, A., and Zimmerman, J. (2019). Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–11. <https://doi.org/10.1145/3290605.3300468>.
- Zhang, Z.T., Liu, Y., and Hußmann, H. (2021). Pilot Attitudes Toward AI in the Cockpit: Implications for Design. *2021 IEEE 2nd International Conference on Human–Machine Systems (ICHMS)*, 1–6. <https://doi.org/10.1109/ICHMS53169.2021.9582448>.

5. Reconfiguring human–AI collaboration: integrating chatbots in welfare services

Elena Parmiggiani, Polyxeni Vassilakopoulou, and Ilias Pappas

INTRODUCTION

Public service delivery is changing. Digital communication channels between citizens and public agencies are increasingly adopted alongside front desks and call centers (Mehr, 2017). One prominent example in this scenario is chatbots, or conversational agents (CAs), a technology that is today widely adopted as a way for organizations to provide customer service (Brandtzaeg and Følstad, 2018; Seeger et al., 2021). Chatbots have been classified in different ways based on their design, the breadth and depth of the tasks they are supposed to carry out, and the duration of their exchanges with humans (Grudin and Jacques, 2019). Recently, advancements in artificial intelligence (AI) applications have boosted the availability of AI-based chatbots that can respond to queries formulated in natural language (Diederich et al., 2022). Public agencies often deploy chatbot-based communication channels to allow citizens to send inquiries and receive real-time answers (Androutsopoulou et al., 2019; Hickok, 2022). Due to the nature of AI with which they are programmed, chatbots take an active role in shaping communication (Diederich et al., 2022). However, we wonder if their impact might go beyond communication. The purpose of this chapter is to provide reflections on how decisions made about AI chatbots in welfare services might have consequences on the work of taking decisions about a case.

Welfare services matter in terms of AI adoption in citizen–public agency communication because they consist of frontline work (also called “street-level bureaucracy”) performed by caseworkers (or case handlers) whose task is delivering services to citizens that are accessible and non-discriminatory (Busch and Henriksen, 2018; EESC, 2015). The vision of national and international governments is that AI-based tools help in achieving this aim while making service delivery more efficient (European Commission, 2018). With the introduction of AI-based chatbots, caseworkers continue to have an important—but arguably new—role because the law in the European Union mandates that decisions about citizens can be aided by AI, but be ultimately taken by a human: Article 22 of the General Data Protection Regulation (GDPR) states that data subjects have the right not to be subject to a decision that produces legal or similarly significant effects based solely on automated processing (Art. 22 GDPR, 2016). In terms of improved efficiency, AI-based chatbots make it possible to scale up communication between citizens and public agencies in periods of very high

demand, sometimes being the only available channel due to high pressure on welfare services, such as at the start of the COVID-19 pandemic (NASCIO, 2022).

As a result, welfare services in particular, and public service delivery in general, deserve attention because this overall promise of improved efficiency that comes with AI-based systems is not a goal in itself, as one would expect in a private organization, but is a means toward the end of providing fair services to all citizens regardless of their background (Hickok, 2022). There are technical and practical challenges to this, however, to balance algorithmic performance and data management (Hutchinson et al., 2021) with public agencies' social function of preserving fairness towards citizens (Schmager et al., 2023). While AI tools have a degree of autonomy that traditional systems do not have, European public agencies face strict limitations in their possibility of training their AI-based systems with data due to privacy regulations such as the GDPR.

In sum, while there is consensus that chatbots become an active part of citizen–agency communication, it remains unclear what this will entail for: (1) how case handlers and citizens interact around an AI-based chatbot; and (2) how case handlers make decisions about a case with AI.

We investigate these aspects based on a two-year study of the deployment of Selma by the Labor and Welfare Administration (LWA) of a Scandinavian country.¹ Selma is an AI-based text-based virtual assistant that answers written inquiries by citizens.²

LWA is a public agency resulting from the merger of welfare, labor, and municipal welfare services in the mid-2000s. It is responsible for a third of the state budget and manages programs related to pensions, unemployment benefits and follow-up, and child benefits, among other tasks. Approximately 850 full-time case handlers are employed at LWA contact centers across the country.

Selma started as a student project in 2017 and was released in 2018. The initial motivation for adopting Selma was the need for support for citizens who struggled to navigate and find information relevant to their case on LWA's very large website. First tested in connection with child benefits, it can in principle handle all domains today. If Selma is not able to answer a query, it redirects the conversation to a human case handler (during working hours). At the beginning of the COVID-19 pandemic in 2020, it was considered essential by LWA's management to handle the steep rise of requests for information by citizens.

This study is part of our longitudinal engagement with the digitalization of welfare services in Scandinavia. We became increasingly curious about the adoption of AI to enhance citizen-oriented decision-making. Data we report in the following section were collected by our research team, including the three authors of this chapter and six master students at our universities during autumn 2020 to spring 2021. Data consisted of video-based interviews and workshops with information technology (IT) experts and caseworkers at LWA. Most caseworkers had been enrolled as "chatbot trainers" to train and improve Selma. In addition, we carried out a large-scale survey with users of chatbots in public services in the country, and collected and analyzed several public and internal documents about the digitalization of welfare services.

We adopt the concept of delegation (Ribes et al., 2013) from actor-network theory (ANT).³ According to ANT, human and technical (including digital) actors are considered equal on the analytical level (Latour, 2005). Delegation is used to describe a process “in which organizational work and agency are passed back and forth across the shifting line between ‘social’ and ‘technical’ elements” (ibid., p. 1). As we shall discuss, we find that frontline communication with citizens—traditionally happening via a call center; see Bratteteig and Verne (2012)—is not replaced by Selma but is gradually delegated to a shifting network of human and digital agents. With this lens, we find that the deployment of Selma emerges as an act of reconfiguring work, that is, the weaving together of the features of technology, materials, discourses, roles, and power structures (Androutsopoulou et al., 2019; Twizeyimana and Andersson, 2019) involving both citizens and LWA case handlers. In this process, on the one hand, citizens learn to interact in novel ways with the chatbot as a new means of communication, something which is new compared to traditional situations such as telephone-based communication. At the same time, case handlers see their work transformed, now involving tasks of aligning with the chatbot and working to develop it further (cf. Vassilakopoulou et al., 2022). We find that the decisions taken by LWA’s caseworkers involved in Selma’s development of its capabilities and appearance (decisions about AI) have consequences for the way the chatbot is both adopted and adapted by citizens and caseworkers. As Selma develops to achieve a more natural conversation flow, citizens learn to adapt their attitude and language to Selma. Caseworkers take on the role of “chatbot trainers”: they learn not only to train and use Selma to perform routine tasks, but also leverage it to perform light case processing and handle citizens’ inquiries (decisions with AI).

BACKGROUND: AI IN CHATBOT-BASED COMMUNICATION IN THE PUBLIC SECTOR

AI is not a new field. It is a broad area of computer science concerned with developing systems that can perceive and infer information in ways that are typically associated with humans. Artificial intelligence refers to machines performing the cognitive functions typically associated with humans, including perceiving, reasoning, and learning (Rai et al., 2019). For a technology to be characterized as “AI” it is important that it demonstrates an ability to learn (McCarthy et al., 2006). A distinctive trait of current AI technologies is that they are frequently based on models that learn autonomously based on input datasets and adjust parameters in the process. The ongoing availability of large datasets and cheaper computational technologies has sparked a boost in AI-based applications in several domains with a significant impact on work and organizing (Waardenburg and Huysman, 2022). As a result, workers are increasingly confronted with AI tools in the workplace, ranging from human resources management to healthcare, and from policework to education (Faraj et al., 2018; Lebovitz et al., 2021; van den Broek et al., 2021; Waardenburg et al., 2018). The literature in infor-

mation systems (IS) and organization studies, for example, has vividly illustrated how AI impacts decision-making processes (Wilson and van der Velden, 2022). Making decisions with AI, scholars have found, involves a redistribution of expertise to improve algorithms' accuracy (Grønsund and Aanestad, 2020). In this process, scholars have found that AI developers and domain practitioners tend to rely on hybrid combinations of AI and domain-based expertise, as opposed to solely one or the other (Rai et al., 2019; van den Broek et al., 2021).

An important, but often trivialized, aspect of AI-based tools is that AI models learn to interpret input data and therefore change their behavior through training and use. As a result, there are no boundaries between the implementation and use of AI in practice (see also Waardenburg and Huysman, 2022). This feature opens up new possibilities, but also new challenges, for how decisions with AI are made. Recent research has pointed to the risks associated with using AI in an opaque way that might lead to deferring responsibility for decision-making to black-boxed systems instead of humans (Vassilakopoulou, 2020). Scholars investigating the impact of technology on decision-making in public services have found that technology tends to either diminish frontline workers' discretionary power, or enhance their work and, simultaneously, better inform citizens (Boulus-Rødje, 2018). Against this backdrop, however, the uptake of AI-based tools and their impact on decision-making about citizens in welfare services has remained under-researched. This is partly due to the difficulty to identify interesting or successful cases as a result of the challenges specific to the public sector that we mentioned above. AI-based chatbots are interesting in this scenario because they are an example of an AI-based tool that is currently deployed or tested by several public agencies despite all challenges.⁴ For this reason, it might be useful to look at extant research to understand what has been written so far about chatbot-based communication. A bird's-eye view of the literature points to two main strands of research on the study of chatbots, which relate mostly to how decisions about the AI embedded in chatbots are made, and how these decisions impact communication.⁵

The first strand of research looks specifically at the chatbot's look-and-feel role within chatbot-human interactions, consequently taking what we call a "dyadic perspective" highlighting design features to improve user experience (Benke et al., 2020), particularly for end users. This research has for instance highlighted that chatbots can increase awareness and communication efficiency, for example in teams (ibid.), or can act as mediators of trust between different stakeholders, such as users and developers (Lee et al., 2021).

A second strand of research, also taking a dyadic perspective on chatbots as an interface between the organization and the human, has focused on the conversation act and people's perceptions. For example, Følstad and Skjuve (2019) investigate the design features of chatbots and suggest that the look and feel (persona) of the chatbot might impact the user experience, while users today have realistic expectations about the chatbot's abilities to handle simple queries. Sannon et al. (2018) find that users tend to disclose sensitive topics if the chatbot is more personified. Seeger et al.

(2021) identify factors that stimulate and impact perceived humanlike qualities in the chatbot.

These two strands of research are important because they highlight what features might increase the usability of chatbots, something which impacts the interaction between citizens and case handlers. These findings point to a need to further understand how such decisions taken about AI-based chatbots impact how decisions are made with AI by case handlers, and with what impact for citizens in the context of the public sector (Bratteteig et al., 2020; Makasi et al., 2020; Simonsen et al., 2020).

We propose to extend dyadic perspectives and consider the broader network of human and technical actors that chatbot adoption triggers. We follow Baird and Maruping (2021), who propose the concept of “delegation” to unpack the relationship between a human and an AI artifact. We are in particular inspired by Ribes et al. (2013), who are explicit about the roots of the concept in the ANT tradition:

Delegation draws on a seemingly obvious insight regarding the interchangeability of human and technical work: put simply, that actors pursuing goals in the world may do so through technical or social means—or both. For instance, hotel managers seeking the return of room keys may remind, cajole, and threaten their patrons, or they may simply attach a weight that makes the key hard to forget and unpleasant to carry. (Ribes et al., 2013, p. 2; see also Latour, 1992)

In short, through delegation humans can decide to transfer the performance of specific tasks to one or more non-human agents. Classical examples are speed bumps replacing policemen, or heavy keyrings to help remind hotel guests to return keys upon checkout (Latour, 1992). Given the self-learning nature of AI, technical agents such as AI-based chatbots can be given the task of controlling a conversation with a human that was previously carried out by another human. While it might be tempting to assume a one-to-one replacement of humans with technical artifacts in the delegation process, Ribes et al. (2013) find that delegation, instead, “should be thought of as a *redistribution* of human work and social ties rather than a complete supplanting of them.” Through this approach, we follow Ribes and colleagues in drawing particular attention to who/what does the work, including emerging new human or non-human actors to sustain collaboration and, consequently, the redistribution of responsibilities and authority for everyday decision-making and upkeep (Baird and Maruping, 2021).

SELMA FROM CHATBOT TO SERVICE AGENT: RECONFIGURATIONS OF HUMAN–AI COLLABORATION

In this section, we present key moments in the evolution of Selma, showing how it has gradually transitioned from an interface between citizens and LWA’s databases to a service agent coordinating with its human “co-workers” and citizens. In the process, Selma has become part of a network of stakeholders in and around LWA, not only replacing human chat employees but also supporting them in traditionally

human-to-human interactions. Chat employees, initially recruited among case handlers to leverage their expertise, also act as chatbot trainers.

We illustrate the work performed and decisions made about the chatbot's features and the organization of the chatbot trainers' work.

From Repetitive Looping to a More Natural Conversation Flow

The very first task that Selma is delegated is answering written queries by citizens, a task usually performed by human caseworkers on the phone or meeting citizens at LWA offices. The more Selma is used, the more it learns to adjust its reactions to incoming queries. LWA's chatbot trainers have focused specifically on training Selma to ensure a more natural conversation flow and avoid repetitive looping. One of the functions introduced in Selma's AI algorithm to achieve this is called "action trigger":

The chatbot can be giving the exact same answer over and over again, say that you ask for "When will I get my parental benefit?". And it gives you an answer, and then maybe you try to rephrase yourself. ("But when will I get my parental benefit?") To get a more specific answer, the chatbot would try to give the same answer, but then because this answer was already given, we have another answer that will say something like, "I think I have given you the only answer I have to this, it might be better to rephrase, or talk to a human?" (IS7)

It is at this stage that human case handlers are involved in the conversation: "If you ask the same question two times in a row then you have the same action twice ... then the reply contains the offer to talk to a human" (IS3).

A second function being added to Selma's AI algorithm to create a feeling of a more natural conversation flow for the citizen is the so-called automatic semantic understanding (ASU) to help Selma analyze the meaning of the words used by the citizen. When ASU is triggered, Selma involves the citizen actively by asking to help in navigating the possible interpretations they have come up with:

Instead of "I don't give an answer", or saying "I don't understand", or if she is uncertain, we have three different reactions there, she thinks of one option, but she is not sure if it has to do with [the percentage of the sick leave], she thinks you are asking about the payment date, but she is not sure. Instead of printing the answer "I don't know", She would say, "I can help you with this, but I'm not sure if it related to", and if you said payment date, instead of salary, she would say ("I'm not sure if it's related to salary but I think I can help you with payment day"). She can give multiple options, if she tied between "I'm not sure if it is sickness benefits, or parental benefits, you are asking about since you mention both", she might say "I'm not sure what you are asking, but maybe some of these options may help?" (IS7)

Overall, the chatbot trainers focused on training Selma to ensure a more natural conversation flow, and to achieve this, two features are key: action triggering and semantic understanding. Both features reconfigure human-AI collaboration. Action triggers provide the possibility to involve human case handlers in the conversation.

Semantic understanding gives the possibility of getting help for the chatbot from the citizen. So, it is not only the chatbot providing help to the citizen, but also the citizen providing help to the chatbot, engaging in a conversation that becomes more natural and closer to an actual peer-to-peer exchange.

From a Conversation to Understanding the Broader Context

Learning to interpret citizens' intended meanings is important for Selma to improve. To do so, however, its AI algorithm needs to develop sufficient awareness of the context in which the citizen finds themselves when submitting a query. Citizens focus on getting the information they need based on the situation they find themselves in, which might be very stressful. Sometimes the chatbot is the only choice they have to communicate efficiently with LWA, especially in periods of high workload such as the beginning of the COVID-19 pandemic.

Selma's look-and-feel plays an important role in letting citizens feel taken care of, without forgetting that they are, after all, talking to an algorithm. The assumption made by LWA was that Selma is thus supposed to meet the citizen with a personified look. Selma's name and look were chosen, maybe not surprisingly, to communicate a warm, empathic feeling. Selma's icon changes color when citizens talk to a human case handler.

However, some users think that they are talking to a human from the start. This generates different expectations about what the chatbot can handle, and thus its competence. Assuming that the chatbot is a human, citizens might write very long sentences that are difficult to process for an AI algorithm: "Many people know very little about how they can communicate with a chatbot. And for this reason, they often start the conversation as they would with a human being. They have a story to tell, right, so there is some background information before they get to the point" (IS8).

However, this makes it difficult for the chatbot to identify keywords and provide the citizen with a relevant answer in the given context that the chatbot might not get.

As a result, it is not only Selma which learns to adjust its replies. Citizens too learn to adapt their language to the chatbot. For example, users have learned to be suspicious about whether an answer is actually relevant. Several informants indicated that they are unsure whether the answer is actually relevant in their current situation, and thus feel safer talking to a human agent who can confirm it:

Not least on behalf of [the state] that people gain confidence in this type of technology, and then we have to deliver. We are still on a very small scale of what is possible to achieve, so we have to slowly but surely score better on the expectations of the user and create more value. Offering Selma on the logged-in page, that she can retrieve payment information about you as a person, then Selma can take the step further. (IS9)

Some citizens even learn to bypass Selma to get hold of a human case handler, for example by writing "human" in the chat. These examples show the importance of understanding the broader context in which citizens interact with chatbots. Citizens

often expect the chatbot to behave like a human and may provide too much background information, making it difficult for the chatbot to provide relevant responses. In such a situation, citizens can easily be disappointed. Continuously improving chatbot capabilities must go hand in hand with efforts to ensure that unrealistic expectations about the chatbot's capabilities are not created.

From Self-Service for Citizens to Case Workers' Personal Assistant

Case handlers have learned to let Selma undertake some light case processing, thus assisting them in performing some mundane tasks and thereby relieving stress:

The biggest change is actually similar to the changes one sees on the phone and more self-service solutions. You filter out the simple questions. You do not sit down to find the form, or where you see the case processing time, or where you log in. So what you are left with are longer and more complex conversations to a greater extent. But this also applies to other channels. The general trend is that human beings are left with more complex [tasks] and routine answers are filtered out very quickly. (IS7)

It is in a way Selma's task to take over the simple conversations [so that] chat employees can use their professional competence to a greater extent than to answer the payment date for sickness benefits, where can I find an application for child benefit. (IS4)

Having Selma as a "right hand" for the case handler/chatbot trainer is quite useful, as they can monitor the live chat, and come up with suggestions about possible answers from the "knowledge base" (that is, the information the chat employee uses, and other relevant information that is stored there), where they can find the information and copy it:

Selma is actually used for quite a lot internally in LWA ... Sometimes the chat employees at the LWA offices, for example, sit in meetings with the user and have something they are unsure of. So instead of starting to search on [the agency's website], as they sometimes have to do ... they can use Selma and get a response. So [Selma] is used internally, but we have not systematized it so much in relation to our employees. (IS20)

I find it very useful. I know the chat employees use Selma when they are in conversation with a user. Also, what is good instead of trying to look for a special question, you can just write a message to Selma, and she will find it right away. And it is very good to know that what is written there is checked by a subject coordinator, so you know the information that is there is correct. (IS16)

However, since Selma is getting new functions, more service areas, and new ways of asking and answering a question, those who train Selma see their workload increase. The chatbot trainers' job has become more complex as the areas Selma has been operating in have expanded over the years:

We started with parental benefits. It is a very narrow area, and it was easy to build and get Selma to predict correctly, but as we expanded we included the rest of the family-related benefits. Then unemployment benefits, sickness benefits, pensions, disability, and so on.

We see that it is another level of complexity because many other words and many keywords are used for many different benefits. So, we have competing answers because what Selma does when she predicts based on what we call keywords that are essential words like “what is unemployment benefit?” (IS10)

These examples show that chatbots can act as personal assistants for case workers not only by filtering out simple questions, but also by providing quick access to information that the case workers themselves can use, and even providing suggestions about possible answers from the knowledge base. So, chatbots are not merely automating information provision, but also becoming part of a new network of human and technology actors that collaborate for service provision.

Visions for the Future: From Simple Information Provision to Light Case Processing and Relationship Management

Working with Selma, caseworkers have familiarized themselves with it and begun to envision future tasks that the chatbot could take on. Some interviewees expect Selma to be able to initiate contact with citizens soon:

When approaching the time when kindergartens open, we have an enormous pressure on the family area, to simply receive those who call and send a message to say that they have been given a place in a kindergarten, and the cash support must be stopped from this and that date, that it may typically be an example where Selma instead of saying “you can send us a written message by logging in and sending here” which then a person should receive, that one then looks at what more it is possible to develop Selma [so that Selma can reply] “thank you for letting me know that you have been given a place in a kindergarten. Can you answer this and this point” and that it never goes via the contact center, but that it is then if they are logged in directly into a case processing solution on cash support and then it stops and catches up if something is missing, then it triggers a new action ... It will be a great way to relieve stress, without taking away any particular service or risking so much. (IS7)

Other caseworkers hope to see Selma proactively contacting citizens through social media:

If we go even further in time, I would have liked Selma to be able to be a little more proactive, that it might see that the unemployment benefits case is about to go out or that you are approaching the maximum amount of sickness benefits, it may be a better example, then the chatbot can contact you via, well, whatever it is we use. Facebook, Messenger, WhatsApp, yes, there are many channels, and ask “What are your plans next?” “Should you have a meeting with a caseworker?”. That Selma can seek you out, on certain conditions, and help you get started so you do not miss out on your rights. (IS19)

As caseworkers familiarize themselves with chatbots they generate ideas on future more proactive tasks that chatbots could take on. Instead of just responding to citizens’ requests, they could initiate contact with citizens, having a more central role in relationship management.

DISCUSSION

Perspectives on the consequences of AI for work often revolve around the idea of human agents being replaced by a chatbot. Chatbots have been described as a tool to release knowledge workers, such as caseworkers, of dull or routine tasks and let them focus on more complex problems (Wang et al., 2023). From this point of view, a chatbot such as Selma is useful to reduce caseworkers' burden associated with human–human communication by acting as a middleman that filters citizens' queries, answers some (arguably, the simpler ones) and, if needed, forwards them to caseworkers.

On a deeper level, however, chatbots are part of a very interesting phenomenon in which the work associated with communication between citizens and caseworkers is reshuffled. Thus, our findings overcome dyadic perspectives of chatbots as interfaces whose look-and-feel impact their adoption or usability (Benke et al., 2020; Følstad and Skjuve, 2019), toward a more nuanced perspective on how AI use triggers a redistribution of work tasks. Understanding the use of AI communication channels as delegation (cf. Baird and Maruping, 2021; Ribes et al., 2013) is helpful to embrace a more dynamic account of how work in public service provision evolves. We showed how citizens see their work changed, where “work” here is intended as the efforts they must make to get in touch with LWA: they learn to write in a way that could be more understandable for the chatbot and, if necessary, to lure it into being forwarded to a human agent. The chatbot adapts its behavior the more it is trained. As Verne et al. (2022) vividly illustrated, a chatbot's active role is also evident when they provide false information if they misunderstand a query. Caseworkers' jobs evolved too, from traditional case processing toward including tasks to train the chatbot based on their experience. From this latter perspective, they take on the new role of “chatbot trainers” and devise new ways of using Selma, such as navigating information in LWA's knowledge base more quickly during conversations with citizens.

The label “chatbot trainers,” assigned by LWA to caseworkers who are involved in the continuous improvement of the chatbot, deserves closer scrutiny. The term “trainer” typically indicates a person tasked with keeping other people or animals ready for an activity or a job. It reminds us of Ribes et al.'s (2013) observation that an important part of the reconfiguration of work due to delegation is system maintenance: every system must be maintained to avoid degradation. While Ribes and colleagues' work was not specifically about AI, we find that this point is even more crucial in the case of AI-based systems, which by nature must be fed with new input data and trained to reach better decisions. In the case of Selma, the caseworker's role is accordingly revised to include system maintenance tasks.

Such maintenance tasks are performative of social relations. Ribes et al. (2013) observed that technologies have an active role in scaffolding social orders, according to the ANT lens of delegation: “technologies ... are not a mediator of communication between distant humans, but ... instead work to sustain a particular social order” (p. 10). What does it mean that the work to sustain the social order in public service provision changes in LWA's case? As mentioned earlier, AI chatbots such as Selma

have the potential to free caseworkers of menial tasks and direct them toward more knowledge-based ones (including, but not limited to, becoming a chatbot trainer), although the consequences for their further education are yet to be explored and would require more long-term studies. This is the classical, instrumental perspective on sociotechnical systems front-staging instrumental objectives of work and economic efficiency (cf. Sarker et al., 2019). More in general, however, AI chatbots sustain social order in the sense that automating (even trivial) communication with citizens has consequences for how a state agency cares for those in need. Front-staging chatbot-based communication might end up further marginalizing vulnerable citizens who are not in the position to use chatbots effectively, or at all. This might end up minimizing the human dignity of those more in need by reducing their possibilities for state support.

Caseworkers' job in welfare is open-ended (cases often stay open for a long time, or are recurrent) and relies on a significant degree of empathic connection with citizens who usually interact with welfare services in vulnerable moments of their lives. Several extant studies warn of the dehumanizing nature of AI which, as a technology, might not be able to provide the warm hands required to handle citizens in need (Kane et al., 2021; Mirbabaie et al., 2022; Rhee, 2018). A trend toward a cold AI that struggles to understand citizens' inquiries might lead to tensions between citizens and public agencies (and governments). It is early to say in what ways we are heading. Our application of the concept of delegation as a reconfiguration of work might help in sensitizing researchers and practitioners to important aspects, tasks, and connections which follow from the use of chatbots that might be leveraged in a constructive way to achieve feedback-driven systems that can empower both citizens and caseworkers in open-ended, empathic conversations (Sharma et al., 2023).

Finally, the role of caseworkers/chatbot trainers as maintainers of the AI system implies that the work they perform on Selma impacts its capabilities (along with other factors). This points to a more general theme: decisions made about Selma as an AI system have consequences for how decisions are made with it. Let us think of the crucial task of caseworkers, namely that of making decisions about a case involving a citizen, involving for example sick leave support, parental support, or child support. Selma is not allowed to make decisions by itself (it would be illegal according to European and national legislation), so the human agent must be in the loop by design. However, as caseworkers adopt Selma both to respond to citizens and to search for relevant information, they use it also as support to take decisions with AI. For example, the way Selma learns to adapt to a more or less natural conversation flow might lead citizens to give different types of information that are then forwarded to human caseworkers. More in general, the more queries that are added to Selma's knowledge base, the more information caseworkers can find when interacting with citizens. This point contributes to the ongoing conversation in IS and organization studies on the role of AI in decision-making (see, e.g., Lebovitz et al., 2021): there is a need for further research on how AI's implementation and training (decisions about AI) impact the AI's aptitude and inner workings (decisions with AI). This is

particularly important in contexts in which humans are in the loop, but take on novel or revised roles to collaborate with the AI.

CONCLUSIONS

Our study of Selma as a new, digital communication chatbot-based channel between citizens and caseworkers at LWA provides a critical analysis of work (re)configurations that chatbot adoption might trigger. Our research is at an early stage, but we hope for future research to investigate these themes further and to contribute with longitudinal studies on how such (re)configurations evolve, what new jobs or tasks emerge or disappear, and what challenges to regulations are surfaced. The adoption of CAs and chatbots in particular in the public sector is still in its infancy, although widespread. Future research will tell whether AI-based chatbots will be a successful communication channel or be supplanted by different approaches, although we observe that chatbot use is soaring in both public and private contexts.

Finally, Baird and Maruping (2021) define delegation to agentic IS systems, such as chatbots, as the process of “transferring of rights and responsibilities for task execution to another” (p. 317). Our analysis contributes to this by inviting future research to analyze such processes of transferring rights and responsibilities, with a nuanced analysis of how work tasks in welfare evolve. Such an analysis sensitizes researchers to trace how new rights or responsibilities emerge and are (re)distributed across networks of human and technical agents.

NOTES

- 1 All names have been anonymized.
- 2 As of March 2023, it is only available in the national language.
- 3 A complete review of ANT and its influences on theories of technology in Information Systems and other fields is beyond the scope of this chapter. For the purpose of this chapter, it is sufficient to remind the reader that ANT is a theoretical and epistemological approach that understands phenomena as evolving networks of human and non-human elements (actors) whose relationships shift over time. For more details about ANT, see Latour (2005).
- 4 See, for example, <https://www.boost.ai/case-studies/public-sector-nordics-conversational-ai-case-study> in the Nordics.
- 5 See Wang et al. (2023) for a more comprehensive overview of the literature on chatbots.

REFERENCES

- Androutsopoulou, A., Karacapilidis, N., Loukis, E., and Charalabidis, Y. (2019). Transforming the Communication between Citizens and Government through AI-Guided Chatbots. *Government Information Quarterly*, 36(2), 358–367. <https://doi.org/10.1016/j.giq.2018.10.001>.

- Baird, A., and Maruping, L.M. (2021). The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts. *MIS Quarterly*, 45(1), 315–341. <https://doi.org/10.25300/MISQ/2021/15882>.
- Benke, I., Knierim, M.T., and Maedche, A. (2020). Chatbot-based Emotion Management for Distributed Teams: A Participatory Design Study. *Proceedings of the ACM on Human–Computer Interaction*, 4(CSCW2), 118, 30. <https://doi.org/10.1145/3415189>.
- Boulus-Rødje, N. (2018). In Search for the Perfect Pathway: Supporting Knowledge Work of Welfare Workers. *Computer Supported Cooperative Work (CSCW)*, 27(3), 841–874. <https://doi.org/10.1007/s10606-018-9318-0>.
- Brandtzaeg, P.B., and Følstad, A. (2018). Chatbots: Changing User Needs and Motivations. *Interactions*, 25(5), 38–43. <https://doi.org/10.1145/3236669>.
- Bratteteig, T., Saplacan, D., Soma, R., and Svanes Oskarsen, J. (2020). Strengthening Human Autonomy in the Era of Autonomous Technology: Contemporary Perspectives on Interaction with Autonomous Things. *Proceedings of the 11th Nordic Conference on Human–Computer Interaction: Shaping Experiences, Shaping Society*, 1–3. <https://doi.org/10.1145/3419249.3420097>.
- Bratteteig, T., and Verne, G. (2012). Creating a Space For Change Within Sociomaterial Entanglements. *Scandinavian Journal of Information Systems*, 24(2). <http://aisel.aisnet.org/sjis/vol24/iss2/7>.
- Busch, P.A., and Henriksen, H.Z. (2018). Digital Discretion: A Systematic Literature Review of ICT and Street-level Discretion. *Information Polity*, 23(1), 3–28. <https://doi.org/10.3233/IP-170050>.
- Diederich, S., Brendel, A.B., Morana, S., and Kolbe, L. (2022). On the Design of and Interaction with Conversational Agents: An Organizing and Assessing Review of Human–Computer Interaction Research. *Journal of the Association for Information Systems*, 23(1), 96–138. <https://doi.org/10.17705/1jais.00724>.
- EESC (2015). *Principles for Effective and Reliable Welfare Provision Systems*. European Economic Social Committee. <https://www.eesc.europa.eu/en/our-work/opinions-information-reports/opinions/principles-effective-and-reliable-welfare-provision-systems>.
- European Commission (2018). *Artificial Intelligence for Europe*. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN>.
- European Council (2016). Art. 22 GDPR, no. 2016/679, <https://gdpr-info.eu/art-22-gdpr/>.
- Faraj, S., Pachidi, S., and Sayegh, K. (2018). Working and Organizing in the Age of the Learning Algorithm. *Information and Organization*, 28(1), 62–70. <https://doi.org/10.1016/j.infoandorg.2018.02.005>.
- Følstad, A., and Skjuve, M. (2019). Chatbots for Customer Service: User Experience and Motivation. *Proceedings of the 1st International Conference on Conversational User Interfaces*, 1–9. <https://doi.org/10.1145/3342775.3342784>.
- Grønsund, T., and Aanestad, M. (2020). Augmenting the Algorithm: Emerging Human-in-the-Loop Work Configurations. *Journal of Strategic Information Systems*, 29(2), 101614. <https://doi.org/10.1016/j.jsis.2020.101614>.
- Grudin, J., and Jacques, R. (2019). Chatbots, Humbots, and the Quest for Artificial General Intelligence. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–11. <https://doi.org/10.1145/3290605.3300439>.
- Hickok, M. (2022). Public Procurement of Artificial Intelligence Systems: New Risks and Future Proofing. *AI and SOCIETY*. <https://doi.org/10.1007/s00146-022-01572-2>.
- Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., and Mitchell, M. (2021). Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 560–575. <https://doi.org/10.1145/3442188.3445918>.

- Kane, G.C., Young, A.G., Majchrzak, A., and Ransbotham, S. (2021). Avoiding an Oppressive Future of Machine Learning: A Design Theory for Emancipatory Assistants. *MIS Quarterly*, 45(1), 371–396. <https://doi.org/10.25300/MISQ/2021/1578>.
- Latour, B. (1992). Where are the missing masses? In W.E. Bijker and J. Law (eds), *Shaping Technology/Building Society: Studies in Sociotechnical Changes* (pp. 225–258). MIT Press.
- Latour, B. (2005). *Reassembling the Social: An Introduction to Actor–Network Theory*. Oxford University Press.
- Lebovitz, S., Levina, N., and Lifshitz-Assa, H. (2021). Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What. *MIS Quarterly*, 45(3b), 1501–1526. <https://doi.org/10.25300/MISQ/2021/16564>.
- Lee, M., Frank, L., and IJsselstein, W. (2021). Brokerbot: A Cryptocurrency Chatbot in the Social-technical Gap of Trust. *Computer Supported Cooperative Work (CSCW)*, 30(1), 79–117. <https://doi.org/10.1007/s10606-021-09392-6>.
- Makasi, T., Nili, A., Desouza, K., and Tate, M. (2020). Chatbot-Mediated Public Service Delivery: A Public Service Value-Based Framework. *First Monday*. <https://doi.org/10.5210/fm.v25i12.10598>.
- McCarthy, J., Minsky, M.L., Rochester, N., and Shannon, C.E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine*, 27(4), Article 4. <https://doi.org/10.1609/aimag.v27i4.1904>.
- Mehr, H. (2017). *Artificial Intelligence for Citizen Services and Government*. Harvard Ash Center Technology and Democracy.
- Mirbabaie, M., Brendel, A., and Hofeditz, L. (2022). Ethics and AI in Information Systems Research. *Communications of the Association for Information Systems*, 50(1). <https://doi.org/10.17705/1CAIS.05034>.
- NASCIO (2022). *Chat with Us: How States are Using Chatbots to Respond to the Demands of COVID-19*. National Association of State Chief Information Officers (NASCIO). <https://www.nascio.org/resource-center/resources/chat-with-us-how-states-are-using-chatbots-to-respond-to-the-demands-of-covid-19/>.
- Rai, A., Constantinides, P., and Sarker, S. (2019). Next-Generation Digital Platforms: Toward Human–AI Hybrids. *MIS Quarterly*, 43(1), iii–ix.
- Rhee, J. (2018). *The Robotic Imaginary*. University of Minnesota Press. <https://www.upress.umn.edu/book-division/books/the-robotic-imaginary>.
- Ribes, D., Jackson, S.J., Geiger, S., Burton, M., and Finholt, T. (2013). Artifacts that organize: Delegation in the distributed organization. *Information and Organization*, 23, 1–14.
- Sannon, S., Stoll, B., DiFranzo, D., Jung, M., and Bazarova, N.N. (2018). How Personification and Interactivity Influence Stress-Related Disclosures to Conversational Agents. *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 285–288. <https://doi.org/10.1145/3272973.3274076>.
- Sarker, S., Chatterjee, S., Xiao, X., and Elbanna, A. (2019). The Sociotechnical Axis of Cohesion for the IS Discipline: Its Historical Legacy and its Continued Relevance. *MIS Quarterly*, 43(3), 695–719. <https://doi.org/10.25300/MISQ/2019/13747>.
- Schmager, S., Gröder, C.H., Parmiggiani, E., Pappas, I., and Vassilakopoulou, P. (2023, January 3). What Do Citizens Think of AI Adoption in Public Services? Exploratory Research on Citizen Attitudes through a Social Contract Lens. *HICSS Proceedings*. HICSS, Hawaii. <https://hdl.handle.net/10125/103176>.
- Seeger, A.-M., Pfeiffer, J., and Heinzl, A. (2021). Texting with Humanlike Conversational Agents: Designing for Anthropomorphism. *Journal of the Association for Information Systems*, 22(4), 931–967. <https://doi.org/10.17705/1jais.00685>.
- Sharma, A., Lin, I.W., Miner, A.S., Atkins, D.C., and Althoff, T. (2023). Human–AI Collaboration Enables more Empathic Conversations in Text-Based Peer-to-Peer Mental Health Support. *Nature Machine Intelligence*, 5(1), Article 1. <https://doi.org/10.1038/s42256-022-00593-2>.

- Simonsen, L., Steinstø, T., Verne, G., and Bratteteig, T. (2020). “I’m Disabled and Married to a Foreign Single Mother”. Public Service Chatbot’s Advice on Citizens’ Complex Lives. In S. Hofmann, C. Csáki, N. Edelmann, T. Lampoltshammer, U. Melin, P. Parycek, G. Schwabe, and E. Tambouris (eds), *Electronic Participation* (pp. 133–146). Springer International Publishing. https://doi.org/10.1007/978-3-030-58141-1_11.
- Twizeyimana, J.D., and Andersson, A. (2019). The Public Value of E-Government—A Literature Review. *Government Information Quarterly*, 36(2), 167–178. <https://doi.org/10.1016/j.giq.2019.01.001>.
- van den Broek, E., Sergeeva, A., and Huysman, M. (2021). When the Machine Meets the Expert: An Ethnography of Developing AI for Hiring. *MIS Quarterly*, 45(3), 1557–1580. <https://doi.org/10.25300/MISQ/2021/16559>.
- Vassilakopoulou, P. (2020). Sociotechnical Approach for Accountability by Design in AI Systems. *ECIS 2020 Research-in-Progress Papers*. https://aisel.aisnet.org/ecis2020_rip/12.
- Vassilakopoulou, P., Parmiggiani, E., Shollo, A., and Grisot, M. (2022). Responsible AI: Concepts, Critical Perspectives and an Information Systems Research Agenda. *Scandinavian Journal of Information Systems*, 34(2). <https://aisel.aisnet.org/sjis/vol34/iss2/3>.
- Verne, G., Steinstø, T., Simonsen, L., and Bratteteig, T. (2022). How Can I Help You? A Chatbot’s Answers to Citizens’ Information Needs. *Scandinavian Journal of Information Systems*, 34(2). <https://aisel.aisnet.org/sjis/vol34/iss2/7>.
- Waardenburg, L., and Huysman, M. (2022). From Coexistence to Co-Creation: Blurring Boundaries in the Age of AI. *Information and Organization*, 32(4), 100432. <https://doi.org/10.1016/j.infoandorg.2022.100432>.
- Waardenburg, L., Sergeeva, A., and Huysman, M. (2018). Hotspots and Blind Spots. In U. Schultze, M. Aanestad, M. Mähring, C. Østerlund, and K. Riemer (eds), *Living with Monsters? Social Implications of Algorithmic Phenomena, Hybrid Agency, and the Performativity of Technology* (pp. 96–109). Springer International Publishing.
- Wang, X., Lin, X., and Shao, B. (2023). Artificial Intelligence Changes the Way we Work: A Close Look at Innovating with Chatbots. *Journal of the Association for Information Science and Technology*, 74(3), 339–353. <https://doi.org/10.1002/asi.24621>.
- Wilson, C., and van der Velden, M. (2022). Sustainable AI: An Integrated Model to Guide Public Sector Decision-Making. *Technology in Society*, 68, 101926. <https://doi.org/10.1016/j.techsoc.2022.101926>.

6. Circumspection as a process of responsible appropriation of AI

Margunn Aanestad

INTRODUCTION

This chapter discusses the process through which technologies that are labelled “artificial intelligence” (AI) become introduced into an organization and embedded into its processes and routines. Through a qualitative case study, we examine what is going on in the period following the initial decision to adopt AI. We have mapped the bi-directional interactions between AI and the organization; both the decisions that were made regarding the AI technology, data, and usage, and how the organization’s existing decision-making practices were impacted by the new technology. During the fieldwork, we were particularly interested in how the organization’s own mandate, goals, and values shaped the process when seeking to appropriate AI technology responsibly.

IS researchers have addressed not just the potential benefits of AI, but also the risks and challenges of adopting and appropriating AI. Unintended consequences have been documented (Pachidi et al., 2021; van den Broek et al., 2021), and practical approaches to control AI have been suggested (Asatiani et al., 2021). The consequential call for responsible AI has resulted in a vibrant research activity across several research fields. For instance, Vassilakopoulou et al. (2022) identified three different streams of research on responsible AI: one based in computer science (addressing AI characteristics such as explainability, transparency, fairness, reliability, robustness); one based in human–computer interaction and human-centered design (discussing how to design “human-centered AI”); and one drawing from ethics and philosophy (discussing how to define “ethical AI” and decide whether AI adheres to fundamental human values). Information systems (IS) researchers’ concerns with AI are typically connected to its learning capabilities, and often focused on machine learning approaches, where the technology is shaped by its use of data rather than fully defined by its underlying software code (Faraj et al., 2018; Benbya et al., 2021; Teodorescu et al., 2021). This results in new forms of inscrutability and autonomy that pose significant challenges for organizations and managers (Berente et al., 2021). Some of the existing works concern adoption of ready-made, productified AI solutions where the computational agency is black-boxed and hidden from its users. In contrast to this, the current empirical case discussed in this chapter is not about such a wholesale adoption process. Rather, it concerns a process of stepwise and careful trials of various analytic technologies, including different machine learning algorithms. This was done in-house by the organizational employees and with a high

degree of deliberation and critical assessment. The longitudinal empirical study therefore offers an opportunity to investigate such an organizational decision-making process in a more detailed way. This process of tentative exploration of a novel technology I choose to name “circumspection.”

ENCOUNTERING TECHNOLOGY THROUGH CIRCUMSPECTION

The influential IS researcher Claudio Ciborra (1951–2005) offered profound, phenomenologically inspired accounts of core IS phenomena (see e.g., Ciborra 2002). His contributions were often critical of the too streamlined and smooth stories of technology implementation and use that abounded in the management IS field. In his polemic style he criticized the assumptions of rationality, the neglect of material complexity, and the unproblematic belief in normative “recipes” associated with much IS research. Based on empirical phenomenology-based studies of managers’ “lived experiences” with complex information technology (IT) projects, he promoted different interpretations and understandings (Ciborra et al., 2000). Among other things, he introduced a different vocabulary that might help to facilitate observation of the “messiness” of real implementation processes. If we look to the lived experiences, he wrote, we will see activities characterized by care, hospitality, and cultivation (Ciborra, 1997). We will see:

a great amount of care taking performed by the various actors involved in the design, implementation and use of IT applications ... continuous commitment from the initial needs analysis throughout constructing the system, training the users, introducing the system into practice, modifying it as new practice emerges, and so on. (ibid., p. 73)

Similarly, the technology that is adopted is initially perceived as unstable and ambiguous (like a stranger), and it requires organizational members’ “being there amidst ambiguity, intimacy, sporting hospitality as well as tamed hostility towards what the new and the unknown is disclosing” (ibid., p. 75). The cultivation metaphor, contrasted with the rational, planning-oriented “construction” metaphor, indicates an approach characterized by “interference with and support for a material that is in itself dynamic and possesses its own logic of growth” (ibid., p. 75), acknowledging that the technology may have an agency of its own.

Care, Ciborra claims, proceeds through three stages: first, as “intentional perception” where the organization considers the technology at a distance; then “circumspection,” getting engaged and wrestling with the technology in situated implementation; and finally, arriving at a place of proper, deep understanding. The notion of circumspection is particularly apt for our purpose here, as it indicates the stage where the organization engages with the technology:

We encounter circumspection as the form of concern that consists in practical problem solving and learning ... While systems are in use their handiness is put to test, their friend-

liness is assessed, their fit with the workflow is monitored, their limits explored, etc. We cope with deficiencies and breakdowns, surprises and shifting effects. We learn how the organization reacts and evolves, how it improvises solutions in an opportunistic fashion. (Ciborra, 1997, pp. 73–74)

The steps that an organization goes through, from considering AI at a distance, via exploring and experimenting, towards achieving familiarity and productive use, are crucial for any manager to understand. This phase may consist of questioning what type of AI will be useful, for what purposes it will be useful, and what it takes for AI to be implemented in a productive way. This chapter zooms in on this process of making decisions about AI. It describes the steps of hypothesizing and articulating expectations, following by trialing, and realizing the huge amount of preparatory work required in its information infrastructure, towards building experience and evaluating the value and benefit of the technology. Through this circumspection process the AI technologies are shaped and adapted to the organization; however, significant changes also occur in the organization and its practices. Thus, the focus here is on both the decision-making about AI, and how the AI implementation changes the organizational decision-making, albeit in an indirect manner.

EMPIRICAL CASE STUDY

A qualitative case study was conducted in a public sector organization, hereafter called “the agency.” The agency was responsible for making payments from the national insurance scheme to healthcare providers and suppliers, as well as to refund some of the health-related expenses incurred by citizens. Traditionally, the main activity of the agency had been oriented towards facilitating the reimbursement of expenditures, based on submitted claims from the service provider. In addition, the agency was responsible to ensure compliance with rules and regulations. As part of this, it performed internal audits where some of the reimbursement claims were re-examined after payment had occurred, to detect erroneous and/or deliberately fraudulent claims. Through several projects the agency has sought to build stronger capacity to conduct these post-payment controls more efficiently and to support its clients in submitting correct claims.

Case Background

One of the main driving forces for the decision to explore AI technologies was a long-standing need to strengthen the agency’s capacity for conducting audits. Previously, the reimbursements claims coming from its clients were controlled though a mix of automated and manual checks. When a service provider submitted a reimbursement claim, it would be screened for errors by an automated validation engine. Around 2000 rules were built into this engine, which rejected erroneous claims automatically. If the claim passed these validation checks, payment was

automatically issued. In addition to these automated checks, the agency conducted manual post-payment audits. These audits were risk-based and targeted individual suspicious cases. Which case to investigate was selected based on a multi-step screening process where composite risk indicators were calculated. Earlier, this involved considerable amounts of data work to produce the necessary insights to select audit candidates. Also, conducting the individual audits was demanding, as they required highly specialized staff and might involve lengthy legal processes. Most of these post-payment audits were successful in the sense that they uncovered actual irregularities. However, the agency did not have capacity to audit more than a small fraction of the potentially suspicious cases. Thus, the automated submission checks and post-payment audits were insufficient to catch all errors. The agency wished to strengthen the capacity to do such audits, and therefore sought to leverage recent advantages in business intelligence and data analytics. In 2021 an innovation project funded by the national research council was initiated. The aim was to utilize digital technology and the already available data better to improve the audit work. It was expected that there would be gains related both to better decision-making in the selection of candidates for audits, and better support during the conduct of audits. One of the plans was to trial the use of algorithms based on machine learning (ML) to understand the already available data. Specifically, an ML-based prediction model for identifying candidates for audits was envisioned.

Data Collection and Analysis

A qualitative case study was conducted between June 2021 and April 2023. The author was a member of the innovation project with a responsibility to conduct research. As such, access was provided to project members, project activities, and internal documentation relating to the implementation process. The account in this chapter is based on participant observation in five project meetings and four workshops, in addition to 13 formal interviews and a large number of informal conversations with organizational members, including with the project leader on project progress, challenges, and achievements.

The analysis strategy was mainly inductive. Initially a chronological timeline of activities and events was created. In the first step of analysis the information provided in the interviews, project documents, and observation notes was related to temporal evolution of activities in the project. In the second step, the overall empirical account was split up into different activity streams within the project. The narrative in the next section is structured to show these activity streams. Each of them discloses emerging strategies to support decision-making in the organization along with the AI implementation process.

FINDINGS FROM THE EMPIRICAL STUDY: CIRCUMSPECTION AS RESPONSIBLE APPROPRIATION

An initial starting point for the organization was to ensure access to data, to analytics competence, and to the tools required. Much of this work took place in the early phase of the project. Legal clearances had to be established before any use (even exploratory use) of the data could commence. These clearances had to state that the project's desired data use did not introduce any mission creep when compared with the original legal mandates under which the data had been collected and were processed, and that the data processing plans, including the combination of data from various sources, were within the legal mandate.

Establishing a Data Infrastructure and Working on Data Availability

Among the core tasks for the audit team was, firstly, to select candidates for audit, and secondly, to conduct the audit on the selected candidates, by collecting, processing, and analyzing data about the case. While the other activities in the agency had dedicated IT systems that supported their work, the auditors relied on manual processes and used several non-integrated IT systems. Two databases owned and operated by the “mother” agency contained the relevant information about the claims, claimants, and payments. Previously, the audit teams would import data from precompiled reports from these two databases into Excel. They had to manually merge data across spreadsheets whenever information was scattered across various reports. To conduct these analyses was time-consuming and could be error-prone. Often people would have to work after hours to ensure that the data extraction processes did not abort or that the computations did not stall. The aim of the analysis work was to create an overview of the population (that is, all recipients of reimbursements) and identify outliers based on certain parameters. The resulting so called “risk lists” would be crucial for determining on which cases an audit should be started. In addition, during an audit, there might arise a need for ad hoc analysis, for example to compare the profile of the service provider to the whole population of providers.

To improve the digital infrastructure, the agency established two new data warehouses and an analytics sandbox with RStudio and Power BI. This enabled the agency's data scientists to develop, test, and store algorithms, as well as the results of queries. The data warehouses contained copies of the most relevant subsets of the data from the two original databases. Thus, they were significantly smaller and the time to conduct queries was reduced. In sum, the new analysis facility allowed much more flexible and quicker analysis. Together, this enabled an increase in capacity to run both standard and ad hoc analyses. Thus, it was perceived to increase the operational efficiency of the agency, even before any machine learning algorithms were deployed.

The initial plan of the innovation project was to employ predictive models that would indicate probability of risk of errors in claims. This was, however, not straightforward. While the two databases contained the reimbursement claims—that is,

central data element for the analysis—they did not contain all data about the service providers and their interactions with the agency. Other relevant information existed in other internal information systems. For instance, historical data from previous audits was stored in the archive system. These were mainly textual reports about the audit process and its outcome. These reports had not been designed or formatted with a future reuse in mind. If they should be used, this required manual information processing. To become a valuable data source for a data-driven organization, the registration practices of the various audit case handlers would have to be standardized and oriented also to documenting in a way that contributed to building up the overall insight of the agency beyond the case handling itself.

While manual information processing of the historical audit reports was possible, there was also another problem. Given that the agency had not had the capacity to investigate all the candidate cases where errors or fraud may have been present, there was not a proper estimate of the amount of the problem on the ground, and there were just too few audited cases to learn from. In the population as a whole, many of the non-audited cases might contain undetected errors, both unintentional mistakes and deliberate fraud. Therefore, the set of historical audit cases could not serve as the “ground truth” about the prevalence of non-compliance in the population. Also, there was a concern that the risk indicators that had been used historically could potentially systematically discriminate against specific groups of service providers, leading other groups to escape scrutiny and detection. Furthermore, there had been rather frequent changes in the reimbursement rules and other regulations during the previous years, and a significant change in the service providers’ behavior during the coronavirus pandemic (for example, a shift from physical to digital consultations). This dynamism in the empirical field further reduced the learning opportunity from historical data. Taken together, this meant that supervised machine learning methods that are trained on labelled data (that is, need to be fed linked input–outcome data sets) to predict non-compliance, could not be implemented.

Strengthening the Organizational Decision-Making Ability

While risk prediction might be out of reach for the time being, risk detection was still a feasible strategy. The agency oriented the efforts towards strengthening its general knowledge base and the detection capabilities. This was done through building better support for the existing processes for calculation of risk indicators based on deviant cases. Several strategies were followed to build more knowledge about the situation on the ground, as a basis for making decisions on what types of analyses would be useful to strengthen detection capabilities.

Harnessing experts’ insights about risk factors and patterns

While there was insufficient historical information on the actual amount and nature of errors within the information systems, there were many experienced employees in the agency. The project sought to harness the experience-based insights from employees such as case handlers, personnel at the help desk, and investigators, who would have

experience from interacting with the users in these different roles. The project organized several workshops with employee experts, each focusing on a specific domain. Here the project team elicited knowledge on typical errors, the patterns in which they would occur, and what the employees considered good indicators of these errors. The knowledge of members of the audit department was enriched with the insights and understandings from other departments. These other employees encountered the clients in other situations, such as in training situations (which revealed what aspects of the reimbursement rules were challenging to understand for beginners) or helpdesk (what questions people have on practical, complex cases), or monitoring (what the error logs reveal about which errors people typically make). Considering the risk of introducing bias based on employees' stereotypes, these insights were not directly implemented as red flags. Instead, this information, combined with the archival data of audit reports and previously known risk indicators, generated several non-compliance scenarios that informed the construction of risk indicators that were selected to be semi-automated and precalculated as input into the candidate selection practice, which continued to be manual.

Later, an organization-wide channel in the Sharepoint infrastructure was established. This should be used by employees who received tips about suspicious behavior, detected misunderstandings through for example their work on the helpdesk or in training sessions, or had knowledge of fraudulent actions from previous cases or ongoing investigations. Previously, such inputs from employees used to be reported, for example, by sending emails to their leaders or in occasional meeting discussions. Now the employees were required to register this into the compliance channel. New procedures ensured that these incoming hints, tips, and indicators from this channel would be systematically considered and integrated into the regular risk assessment exercises. This replaced earlier ad hoc procedures and ensured both broader capturing and more systematic use of the employees' relevant insights.

Supporting the work of detecting outliers by more comprehensive analysis

To identify candidates for audits, the auditors used to depend on the detection of outliers. A risk indicator called estimated economic risk (EER), which indicated how much one actor's use of the available tariffs deviates from other comparable actors' usage, was calculated. This was combined with other risk indicators, such as incoming tips about irregularities or historical irregularities, into a composite "risk list." This risk list was then discussed and edited in several rounds by auditors, and managers discussed the prioritizations before deciding on which of the cases an audit process would be initiated. These lists had earlier been generated once a year, but were now generated each quarter. This was made possible by the improved data infrastructure. Also, this meant better support for the audit work itself. The data scientists in the project supported the auditors' work by precalculating time series of various risk indicators; informed, among other things, by the risk scenarios derived from the employee workshops. Standardized and easily updated statistics reports now complemented the data reports and simplified the analysis. Also, the pre-existing practices used in specific areas (that is, a specific category of service providers) were

expanded into also other areas (other service provider categories), and a joint risk list replacing the earlier category-specific lists was generated. This was done to enable a joined-up view on the risk, helping to allocate resources to audits based on a bigger picture than the previous area-by-area approach.

Exploring the potential value of other data sources

All claims that were submitted had to pass the automatic submission checks. The validation engine had implemented more than 2000 explicit rules and would reject claims which had clear errors. For instance, duplicate claims, claims which contained multiple tariffs that could not be combined, claims where the time period did not match the service, and so on, would be rejected. The rejected claims were either returned to the sender together with a textual note (an error notification report), or routed into manual claims handling. The log of error reports was considered to have potential value to learn about errors made. An informant said:

this is useful information that should be followed up and potentially can be used as red flags ... In the Compliance project we are going to build a risk database, and a good place to start is then to look at earlier sanctions and how the error notes look, how often the rules are triggered, and how they co-vary with each other. (Interview with project member, auditor)

The data scientists produced several correlation matrices based on the feedback (the error notifications) generated by the automatic rejection of claims in order to understand such co-variation. One catch-all risk prediction model has not yet been developed, and may never be. Instead the data scientists, through exploratory analysis of the data and in dialogue with the audit experts, have developed and validated several specific, more targeted predictive models.

The data from the analysis of error logs were also interesting beyond offering overall insights on patterns. In the words of a project member:

In the Audit department, the rejected claims is something we look at, the various reasons for rejection ... when the tariffs are not possible to be combined in the same bill or within a short time period, and then they just move the time period a bit and then they get the payment. But we look at a small group, on those who have a high motivation to extract as much money as possible, I don't think this holds for the majority.

In other words, these data were considered a source of insights about the service providers' behavior, for example, disclosing which acts would be attempted to make a claim pass the automatic checks. This indicates the significance of understanding how the providers behave and act.

Establish adaptive capacity through experiments

Impacting behavior through early intervention to prevent errors in reporting was considered important. Different activities tried different means for establishing a feedback loop and monitoring change in behavior. The agency designed and implemented experiments that it was hoped would generate both improved compliance and new

insights about the prevalence of errors and providers' behavior. As a starting point, the experiments consisted of tweaking the error notification reports that were sent back to the providers with erroneous and automatically rejected claims. After a pilot test that verified the technical feasibility of the experimental setup, a group of new service providers were targeted. They would be supplied with a series of feedback reports on how their usage of tariffs compared with the rest of the population. Also, the regular, general feedback reports sent every three months to the service providers were improved in clarity. These interventions both provided objective and factual information, and demonstrated that the agency was actively monitoring the claim practices. The aim was to "inform, influence and build trust" (project report). The plan is to be able to implement such experiments in an adaptive fashion, evaluating the effects along the way and thus learning about and impacting behavior at the same time, leading to a better knowledge base and better compliance.

In the future, continuous change is to be expected: there will be changes in the regulation (for example, of reimbursement tariffs), legislation, and in medical procedures, and the service providers' reporting practices will adapt to these changes. It is also expected that the error patterns will change, along with changing opportunities for misreporting and perceived detection risks. The agency therefore needs to build capacity for ongoing learning and adaptation. Currently, the audit activities happen post-hoc. There is a significant time delay between the act and the audit, and the audit will therefore have limited effect on preventing future errors. The adaptive capacity that is being built through these experiments may counter this tendency, and connect the act and the audit response more tightly. An ideal solution to this also requires enrollment of other actors; also the service providers' own information systems need to be upgraded to indicate errors or issue warnings as the information is being recorded. This would increase the likelihood of reducing the errors at the source.

DISCUSSION

Circumspection as Evolution of the Information Infrastructure

During the "circumspection" phase, where the agency engaged with the new technology, several changes happened both in the information infrastructure and in the agency's epistemic capabilities. The organizational knowledge relevant to the performance of its core tasks was not only represented in codifiable forms and embedded in IT systems, but it was also distributed across and embedded in a "hybrid assemblage" of tools, systems, people, routines: the agency's information infrastructure. The pre-existing information infrastructure had been developed to support the pre-existing modes of working, which did not employ AI technologies.

Thus, it had to evolve to accommodate the planned implementation of AI in the form of predictive models. From the empirical study, we saw that the actual process of exploratory and tentative implementation triggered a set of changes in the agency's information infrastructure. For one, it triggered changes that emerged from a

“preparatory” mode: to deploy AI, the data scientists recognized that swifter and more flexible access to data had to be established. The data warehouses were created along with the analytics resources (tools, sandbox), and data science skills were developed through training and recruitment. Following this, the increased speed and ease with which both routine and ad hoc analyses could be conducted was perceived as a significant benefit, even before any predictive models had been explored. Other changes emerged in a more “responsive” mode, following realizations about the potential and limitations of both the existing information infrastructure and the new technology. The alternative, compensatory strategies were pursued when it was realized that supervised machine learning was not feasible to employ. Moreover, the project triggered expansions of the infrastructure and work practices, from well-known service domains (general practitioners and specialist doctors) to less well-known domains where adjustments were necessary. In the long run, the project contributed to strengthening the agency’s monitoring capability. Its analysis capacity was boosted, a more comprehensive risk picture systematically built up, and targeted prediction models developed.

Circumspection as a Process of Strengthening Decision-Making

The AI implementation was from the outset driven by need to increase processing capacity, not by any “AI hype” which expected that specific tools would be generic silver bullets. Still, there has been an adjustment of the expectations relating to the AI. The process of preparing to implement AI was beneficial to the agency even though very little AI was deployed at the time of writing in 2022. It did lead to a strengthened data infrastructure, the building of novel skills (both specific data science skills and broader organizational awareness of the potential and requirements of AI), and it instigated new connections and relations across pre-existing organizational silos. Even before it was deployed, the AI contributed to disciplining the organization, making it clear how new demands needed to be met.

There have been changes of work tasks through automating certain tasks. Replacing the previous manual processes of generating analyses with a quicker generation of more standardized pre-calculated analyses was perceived as a desirable change. Not only did it remove the need to work after hours to monitor the analysis generation, but also, overall, it left the agency better prepared to use available data better. In the process of AI finding its place in the organization’s decision-making processes, there was a shift from the humans conducting the tedious work towards monitoring and governing the automatic execution of these tasks.

More interestingly, the implementation aligned with ongoing organizational work to establish more solid understandings of the phenomenon of compliance and non-compliance on the ground. The project made even more visible that the agency had insufficient and fragmented knowledge about the number of errors, both unintended and deliberate, and about their prevalence and occurrences. These processes, which at the time of writing in 2022 have merely started, will contribute to a reworked “record” of the various insight-generating activities of the agency; the

agency's "memory" will also be strengthened, partly triggered by the AI implementation. Thus, a central learning point from our case study is that we should not solely focus on the AI technology and its capabilities, but we should also pay attention to how the organizational capabilities, both pre-existing and novel, evolve and are impacted when the organization engages with AI technologies.

Circumspection as a Process of Responsible Appropriation

The case shows a deployment of AI technologies which was careful, and not rushed by any external concerns. The various steps were assessed from many dimensions, such as whether the planned data processing was within the legal scope, or whether the possible risk models might contain bias from either human stereotypes or historical selection bias. A report states that: "This project shall not just train machine learning models, it is in equal part about ensuring traceability in the registration and use of findings, tips and using the sanctions, in order to be able to use these data in analyses in a safe way." The project members ensured that the data used for developing the analysis models were anonymized (with respect to patient information) and pseudonymized (with respect to service provider identity). The work to strengthen the knowledge base for decision-making was systematic: they would develop various risk indicators and compare them, including their co-variation. Also, they developed various prediction models for different risk indicators and assessed the outliers' characteristics. These new insights were also to be used to critically assess their current risk indicators and models, as to whether they systematically discriminated against various groups. Finally, the models were intended to be continuously updated; this was also considered to improve the precision and lessen the impact of audits on the compliant service providers who occasionally might commit errors.

Also, considering its societal role, the agency maintained a dialogue with the professional organizations of the service providers, to maintain legitimacy and ensure that the intended actions would not trigger resistance. They were also conscious of the risk of "mission creep" that could emerge if they prioritized easily achievable gains. For instance, in a workshop, one project member said that: "if we just choose number of audits as our KPI [key performance indicator], we would of course go for the easier cases so that we could maximize number of hits. But that would not be a right priority when we look at the larger picture." The fundamental values, goals, and societal responsibility of the agency thus had a definite impact on the process through which the new technology was adopted and appropriated. The agency thus pursues the various strategies that Teodorescu et al. (2021) propose to ensure fairness: namely, reactive oversight, proactive oversight, informed reliance, and supervised reliance.

CONCLUSION

The empirical story has described the initial phase of an AI adoption and appropriation process. The work is ongoing, and it is to be expected that the next phases will bring new demands and achievements. So far, the introduction of AI technologies has necessitated work on the data infrastructure, introduction of novel information systems applications, and specialized data science competence. It has also effected changes in information use practices and subsequently in data generation, and triggered the establishment of novel relations across organizational silos to “know what we know that we know,” that is, to realize the organization’s latent epistemic capabilities. This story thus shines the spotlight not just on the novel actor—the AI technology—but even more so on the pre-existing organizational practices and resources, and on the need to evolve to accommodate the new technology.

The agency sought to deploy predictive models, based on machine learning algorithms. It is important to note that such technologies would often be classified as “just” business analytics and not as “real AI.” One reason for the often fuzzy content of the notion of AI is that AI is a so-called “horizon concept,” a concept that points to the frontier of what is emerging. Berente et al. (2021) define AI as “*the frontier of computational advancements that references human intelligence in addressing ever more complex decision-making problems*.” In short, AI is whatever we are doing next in computing” (p. 1435). As such, it denotes those technological capabilities which are not yet realized. This has the effect that specific technologies where the imagined capabilities have become manifest are somehow “dethroned” and lose their category membership; they acquire their own given names and are no longer considered “real AI.” For instance, the current AI wave is oriented to automation of cognitive types of tasks. Initially, robotic process automation (RPA) and chatbots figured centrally in mappings of organizational use of AI, but these technologies are no longer considered to be at the frontier. Similarly, AI researchers often seek to distinguish “real AI” (involving machine learning, and preferably deep learning) from “just” business intelligence and analytics, even while these may employ machine learning algorithms. I thus have used AI in a broader way, because it is not that important whether this “really” was AI or not. To the organization this was a novel technology, and it triggered the circumspection process which is in focus here: “the match to be achieved in vivo between the new systems and the unfolding work situation” (Ciborra, 1997, p. 74). Through circumspection, the novel technology, or the unknown stranger that Ciborra also talks about, is becoming known. This process may imply a loss of category membership, so that it is no longer thought about as being AI.

The chapter is written from an augmentation rather than an automation perspective. This implies that the text does not cast AI as a radical game changer. While much popular discourse may claim that AI will rapidly and completely replace existing ways of working, this is not often the case on the ground, at least not in well-established organizations. Rather, through the process of circumspection, AI technologies’ promises are evaluated and their ability to deliver value are tested before they are assigned their tasks and roles in the organization (Grønsund and

Aanestad, 2020). Focusing on the process of circumspection and evolution of the information infrastructure helps to emphasize continuity over disruption, and gradual rather than revolutionary change. This is of course a matter of perspective: if we choose tasks, rather than jobs, as our unit of analysis, we will see automation happening. If we choose a temporal frame for our study that spans beyond a couple of years, we may see that smaller changes may add up to become fundamental and radical, and that second- and third-order effects become more visible than they were within a few months or years (Baptista et al., 2020).

Also, the processes will not look similar in all settings. Our case concerns a public sector organization with a well-defined mandate and task which sought to strengthen its capabilities to fulfill its core tasks. In other situations—for example, where industry incumbents or entrants compete to create and exploit new markets—the process of deploying AI may look very different. In our case, some tasks, such as pre-calculations, were automated to create efficiency, while other tasks were maintained as human tasks. The technology was intended to be used in an assistive mode, providing data-generated insights that would support tactical decision-making. This account is an example of how AI may be used as “enhancing human agency, without removing human responsibility” (Floridi et al., 2018, p. 692), and thus illustrates a process of responsible appropriation of AI.

ACKNOWLEDGMENTS

The research was made possible through financial support from the Research Council of Norway (grants 321044 and 341289), and through the informants’ willingness to share their experiences. This is gratefully acknowledged.

REFERENCES

- Asatiani, A., Malo, P., Nagbøl, P.R., Penttinen, E., Rinta-Kahila, T., and Salovaara, A. (2021). Sociotechnical envelopment of artificial intelligence: An approach to organizational deployment of inscrutable artificial intelligence systems. *Journal of the Association for Information Systems*, 22(2), 325–352.
- Baptista, J., Stein, M.K., Klein, S., Watson-Manheim, M.B., and Lee, J. (2020). Digital work and organisational transformation: Emergent Digital/Human work configurations in modern organisations. *Journal of Strategic Information Systems*, 29(2), 101618.
- Benbya, H., Pachidi, S., and Jarvenpaa, S. (2021). Special issue editorial: Artificial intelligence in organizations: Implications for information systems research. *Journal of the Association for Information Systems*, 22(2), 281–303.
- Berente, N., Gu, B., Recker, J., and Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, 45(3), 1433–1450.
- Ciborra, C. (1997). De Profundis? Deconstructing the concept of strategic alignment. *Scandinavian Journal of Information Systems*, 9(1), 67–82.
- Ciborra, C. (2002). *The labyrinths of information: Challenging the wisdom of systems*. Oxford University Press.

- Ciborra, C., Braa, K., Cordella, A., Dahlbom, B., Hepsø, V., Failla, A., and Hanseth, O. (2000). *From control to drift: The dynamics of corporate information infrastructures*. Oxford University Press.
- Faraj, S., Pachidi, S., and Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. *Information and Organization*, 28(1), 62–70.
- Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., and Rossi, F. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707.
- Grønsund, T., and Aanestad, M. (2020). Augmenting the algorithm: Emerging human-in-the-loop work configurations. *Journal of Strategic Information Systems*, 29(2), 101614.
- Pachidi, S., Berends, H., Faraj, S., and Huysman, M. (2021). Make way for the algorithms: Symbolic actions and change in a regime of knowing. *Organization Science*, 32(1), 18–41.
- Teodorescu, M.H., Morse, L., Awwad, Y., and Kane, G.C. (2021). Failures of fairness in automation require a deeper understanding of human–ML augmentation. *MIS Quarterly*, 45(3), 1483–1500.
- van den Broek, E., Sergeeva, A., and Huysman, M. (2021). When the machine meets the expert: An ethnography of developing AI for hiring. *MIS Quarterly*, 45(3), 1557–1580.
- Vassilakopoulou, P., Parmiggiani, E., Shollo, A. and Grisot, M. (2022). Responsible AI: Concepts, critical perspectives and an information systems research agenda. *Scandinavian Journal of Information Systems*, 34(2), article 3.

7. Responsible AI governance: from ideation to implementation

Patrick Mikalef

INTRODUCTION

The growing interest in artificial intelligence (AI) has been accompanied by a plethora of cases where negative or unintended consequences have emerged (Wang and Siau, 2019). Such occurrences have prompted policymakers, researchers, and practitioners to think about how the development and use of AI can follow “responsible” principles (Floridi, 2019; Mikalef et al., 2022). Such principles concern aspects such as transparency, explainability, auditability, and security of AI, among others (Arrieta et al., 2020; Bucker et al., 2022). While there is considerable development of knowledge around responsible AI principles and what they entail (Clarke, 2019), there is to date limited understanding of how organizations should implement them in practice (Schiff et al., 2020b). In other words, we still lack knowledge regarding how to convert often high-level principles into implementable governance practices (Wang et al., 2020). This issue is becoming increasingly complex when considering the different stakeholders that are relevant in the design, deployment, and use of AI technologies in contemporary organizations. Furthermore, ethics and norms are both context-specific and evolving over time (Emerson and Conroy, 2004). These characteristics of ethics and norms make it increasingly complex to provide a universally applicable framework for responsible AI governance (van Rijmenam and Schweitzer, 2018).

To explore how AI applications are developed, deployed, and managed in a responsible way, the notion of responsible AI is analyzed from an information systems (IS) perspective, and specifically from the vantage point of governing such technologies. Adopting such a perspective allows us to move from an often high-level, abstract, and descriptive understanding of principles, to one that regards them as actionable and embedded in the development of AI. In essence, responsible AI governance entails the design, resource orchestration, and value generation of AI technologies in adherence to responsible AI principles. As such, it differs from conventional notions of information technology (IT) governance by placing an emphasis on the responsible and ethical use of AI throughout its lifecycle (Theodorou and Dignum, 2020). Through a synthesis of latest scientific research and reports on the principles that underlie responsible AI, this chapter then presents an approach through which organizations can manage their AI projects throughout their lifecycles and ensure that responsible AI principles are met. While there is extensive discussion on the principles that should underlie responsible AI, there is to date very limited

understanding of how such principles should be incorporated in the management of AI technologies. This lack of knowledge significantly hinders the embedding of responsible principles in AI technologies, which has resulted in several prominent cases where negative or unintended consequences were observed (Acemoglu and Restrepo, 2018).

To explore this gap in knowledge, this chapter presents responsible AI governance in the form of a framework that can be used as the basis for planning AI initiatives. In addition, a process-view is incorporated in the framework which helps in illustrating the key phases of responsible AI governance deployment and the critical points to consider at different phases. The framework is grounded on an evolutionary perspective which assumes that responsible AI governance comprises a set of practices that are influenced by, and influence, the environment in which they operate. As a result, organizations need to be able to identify external stimuli which in turn shape their responsible AI governance practices. At the same time, how organizations decide to govern their AI technologies is argued to influence organizational performance outcomes, both directly and indirectly. As a result, it is important to identify the mechanisms through which responsible AI governance can exert an effect on organizational operations, as well as how it influences external perceptions of the organization and social norms.

In closing, the chapter develops a discussion around the research, practical, and policymaking implications of governing AI through a responsible lens. The framework and findings from the case examples also serve as a means to problematize the ongoing research discourse, as well as to critically assess the current practices that are used within organizations. Through this analysis, the chapter proposes a set of approaches that can be implemented in order to ensure that AI applications follow responsible principles, and discusses the implications from an organizational and value-generation perspective. Finally, some prominent themes for future research that build on a multi-disciplinary approach are highlighted.

FROM RESPONSIBLE AI PRINCIPLES TO GOVERNANCE

The idea of governing AI in an ethical and responsible way is not new, and builds on a long history of considering the ethical consequences of introducing new technologies (Anderson and Anderson, 2007). Since the mid-1900s, primarily through science fiction novels, the implications of designing sentient technologies that exhibit features of artificial intelligence have been explored. During the last decade, fiction has turned into reality, as the field of artificial intelligence has been witnessing a long-awaited “AI spring” (Shin, 2019). The rapid proliferation of AI techniques, combined with the availability of vast amounts of data and increasing processing power, has facilitated the emergence of AI applications in practice (Enholm et al., 2022). As a result, there has been a renewed interest in identifying how AI should be managed to minimize any negative or united consequences (Mayer et al., 2020). An outcome of such efforts has been a multitude of responsible, trustworthy, and ethical

Table 7.1 Principles of responsible AI

Dimension	Description	References
Accountability	AI systems should provide appropriate opportunities for feedback, relevant explanations, and appeal	EC (2019); Liu et al. (2019)
Transparency	AI systems should commit to transparency and responsible disclosure regarding AI systems	Arrieta et al. (2020); EC (2019)
Fairness	AI systems, and the underlying algorithms and datasets, should not promote, reinforce, or increase unfair biases	EC (2019); Hughes et al. (2019)
Human agency	AI systems should support human autonomy and decision-making	Ågerfalk (2020); EC (2019); Jarrahi (2018)
Technical safety	AI systems should prevent risks, minimize unintentional and unexpected harm, and prevent unacceptable harm	Dobbe et al. (2021); EC (2019)
Privacy	AI systems should give the opportunity for notice and consent, encourage architectures with privacy safeguards, and provide appropriate transparency and control over the use of data	EC (2019); Manheim and Kaplan (2019)
Societal well-being	AI systems should ensure the prevention of harm to the broader society and other sentient beings and the environment	EC (2019); Schiff et al. (2020b)

guidelines which have been proposed by international organizations, governments, and private enterprises, among others (Fjeld et al., 2020). A growing consensus within this suggests that responsible AI follows a certain set of principles, which include aspects of accountability, transparency, fairness, human agency, technical safety, privacy, and societal well-being (Smuha, 2019). Specifically, these principles entail the aspects described in Table 7.1.

The overarching logic of responsible AI principles is to provide stakeholders with a set of dimensions to gauge how AI systems should be developed and used. Over recent years, this idea has gained widespread popularity, such that a review conducted in 2020 at the Berkman Klein Center for Internet and Society at Harvard University identified 36 documents that mentioned different dimensions of responsible AI (Fjeld et al., 2020). To date, this number has increased significantly; however, there is a growing consensus that responsible AI is a multi-dimensional concept that spans different levels within and outside organizational boundaries, and involves different stages of AI development and deployment. In an attempt to contextualize the practices that underpin the management of AI within organizations, the concept of responsible AI governance has emerged. Responsible AI governance builds on a long stream of research on governance of technology (De Haes and Van Grembergen, 2004; Weill and Ross, 2005), which distinguishes between the necessary practices to effectively leverage IT resources to achieve organizational objectives. The convergence of the two notions lies in the fact that responsible AI principles dictate what needs to be attained, while governance mechanisms define how that is achieved (Graham et al., 2003). In other words, governance entails the means and mechanisms by which principles are set in action. When it comes to responsible AI governance, the principles that have been extensively discussed in prior literature provide the foundation upon which governance mechanisms must be designed. Effectively this

means that responsible AI principles shape the structures, roles, and processes associated with managing AI (Tallon et al., 2013).

As illustrated in Figure 7.1, governance is conceptualized by encompassing structures, processes, and relational mechanisms (Van Grembergen et al., 2004). In line with prior definitions, governance entails the transformation of IT artifacts to meet present and future demands of the business and the business customers. As such, the notion of responsible AI governance includes the structural, procedural, and relational mechanisms that enable organizations to leverage AI in an ethical and responsible way. Structural mechanisms involve the practices of assigning responsibilities for supervising, directing, and planning responsible AI governance. Procedural mechanisms involve the practices of shaping resource allocation processes, defining established approaches of performing tasks, and conducting evaluation of developed applications. Relational mechanisms include the practices that shape involvement on knowledge sharing, idea exchange, and communication approaches (Tallon et al., 2013). Through this dimensionalization it is then possible to map different aspects related to responsible AI principles into actionable and implementable responsible AI governance practices. The direction of the arrows indicates that responsible AI governance principles form the basis upon which governance practices are shaped, while the evaluation of these practices informs the refinement and reconsideration of responsible AI principles. Thus, the relationship between principles and governance practice is a dynamic one that evolves over time.

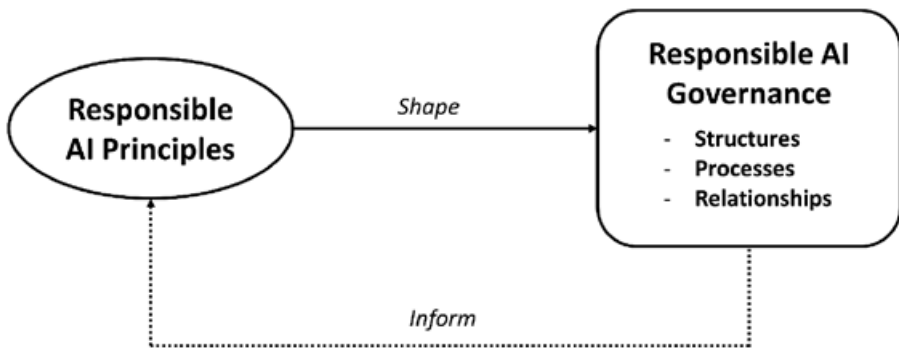


Figure 7.1 Relationship between responsible AI principles and governance

Several recent studies have started exploring the notion of responsible AI governance in empirical settings by identifying challenges of implementation as well as effects on internal and external performance. A recent study by Eitel-Porter (2021) examined the implementation of ethical AI and found that organizations which implement strong governance frameworks, which are overseen by an ethics board, and have established appropriate training, reduce the risks associated with unintended or negative effects of AI. Furthermore, having such governance practices makes it easier for businesses to scale their AI applications. Responsible AI governance frameworks

therefore have been associated with changes in structures and processes within organizations, as well as the design of AI systems. Indicative of this is a recent report of the Boston Consulting Group which highlights several elements that should be included in such practices, including the design and use of tools and methods (Mills et al., 2020). Specifically, the report mentions the need to create toolkits comprising of tutorials, modular code samples, and standardized approaches for addressing common issues such as data bias and biased outcomes. In addition, the relationships and roles in such settings need to be redefined, which also entails empowering leadership of responsible AI through new roles such as chief AI ethics officer, among others.

While there has been significantly more work on important elements of responsible AI governance and how to implement such principles in practice, there is a smaller but growing stream of research that looks into the effects on organizational outcomes. Much of this research is grounded on prior studies of corporate social responsibility (CSR), which has a long tradition of examining mechanisms through which such practices influence key organizational outcomes. While a lot of the effects of responsible AI governance have been grounded on assumptions or theorizing, there is a developing consensus that such effects can be identified internally and externally. Internal effects primarily involve improved employee productivity, knowledge sharing, and efficiency (Papagiannidis et al., 2022). On the other hand, external effects concern perceptions of customers on ethical and responsible behavior of the organization, relationship with key stakeholders and government agencies, as well as financial and non-financial performance effects (Deshpande and Sharp, 2022; Wang et al., 2020).

RESPONSIBLE AI GOVERNANCE IN THE ORGANIZATIONAL ECOSYSTEM

While the previous section has outlined some key aspects relating to responsible AI governance, it is important to note that there is a high degree of contextuality when designing and implementing such practices. In addition, the way such practices are implemented in organizations is dynamic and constantly changes as ethics and societal norms evolve over time, so it is important that organizations adopt an approach that incorporates such an evolutionary logic. In the following sub-sections there is further elaboration on aspects of contextuality and other contingency elements that influence responsible AI governance formulation and implementation. Figure 7.2 provides a sketch of the context within which responsible AI governance practices are shaped and have an effect. This figure is developed to explain the evolutionary nature of responsible AI governance, which is seen as being in a state of constant flux, shaped and shaping societal values and norms, and influencing organizational directions. In addition it underscores the contextual nature of responsible AI governance, and how different contingency elements influence how it is adopted. The solid

lines depict direct effects, while the dotted lines represent feedback loops that lead to the refinement of principles and governance practices.

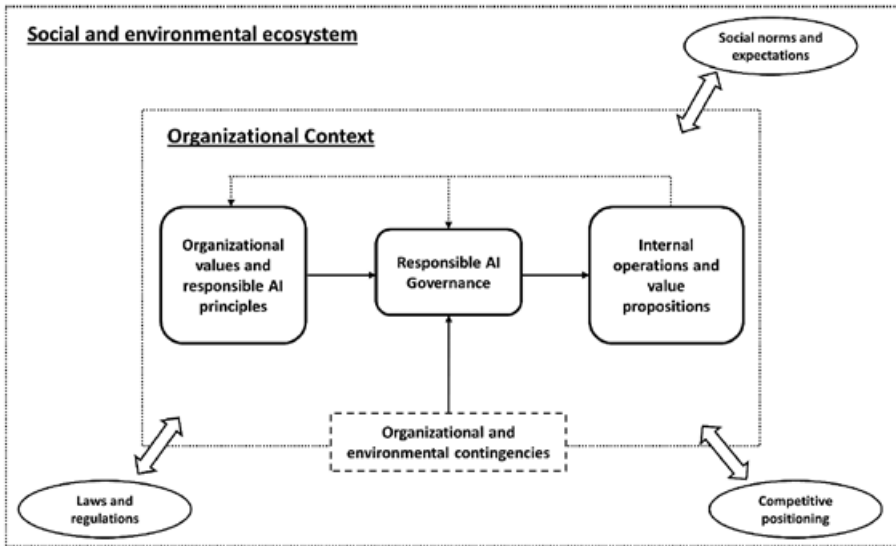


Figure 7.2 Conceptual overview of responsible AI governance in broader ecosystem

Contextuality of Responsible AI Governance

One key assumption which is made by the majority of scientific studies and reports on responsible AI principles is that the principles are implemented in a uniform manner. This assumption is based on two separate issues that appear when implementing responsible AI principles in practice. The first has to do with the fact that the conversion from responsible AI principles into actionable governance practices does not happen directly. Prior studies have shown that organizations typically have separate phases for ideation of governance practices and their subsequent implementation (Taeiagh, 2021). During the ideation phase the high-level practices and supporting documentation of approaches are formulated, which is then followed by the implementation phase where these are used as input from the relevant stakeholders in order to direct their actions (Cihon et al., 2020). Nevertheless, research has shown that there is often a gap between stated governance practices and those that are implemented (Baan et al., 2017; Sturm et al., 2021). Such gaps often result in mismanaged AI initiatives which may potentially lead to negative outcomes (Lam et al., 2021).

The second issue when implementing responsible AI governance practices is that principles are often contextualized and subject to different contingencies that underpin focal organizations. Among these issues, for instance, are differences in cultural and societal norms among organizations operating in different countries. These

differences essentially entail a focus on different aspects of responsible AI principles that may be of higher importance, or a different interpretation of how they should be implemented in practice (Yatsenko, 2021). In fact, the review of responsible AI frameworks of Fjeld et al. (2020) shows that depending on the country of origin, some principles are emphasized over others. Studies have also shown that trust-building, information consumption, and adoption of novel technologies are subject to strong national and cultural influences (Ashraf et al., 2014); thus, it is highly probable that the way responsible AI governance practices are implemented will have an important bearing on their success.

In addition to the cultural or national differences, there is likely to be variation on how responsible AI governance is implemented based on other contingency elements that characterize organizations. Among these, size class is likely to play an important role in determining the extent to which responsible AI principles are followed in practice. Organizations belonging to a large size class are more likely to have slack resources and prior experience for developing responsible AI governance, compared to smaller organizations. In addition, many of the tasks and activities associated with responsible AI governance, such as ensuring auditability, investing in resources to minimize bias at different stages of AI development, and developing explainable AI solutions for different stakeholders, entail large financial investments which smaller organizations may not be able to sustain (Ghallab, 2019; Minkinen et al., 2021). On the other hand, smaller organizations have shorter power-distances from top management to implementation teams, and are therefore more agile in rolling out new practices associated with responsible AI governance (Meske et al., 2022). Another distinction which has been noted is between new and older organizations, with the argument being that older organizations have more rigid routines and processes, so they are less likely to be reactive to new ways of governing technology, and particularly to responsible AI governance (Shneiderman, 2021). As a result, there are several contingent elements that are likely to influence the extent and the way in which responsible AI governance is implemented within organizations.

Evolution of Ethics and Norms

Another key assumption which many responsible AI frameworks build on is that societal norms and ethics are unchanged over time. Nevertheless, the expectations from society concerning technology management, and particularly AI applications, have shifted drastically over the past few years, particularly after the first real-world applications were rolled out. For instance, early applications of supervised machine learning for decision-making have showcased how strong the presence of bias can be in AI applications (Mehrabi et al., 2021). Such cases have opened the discussion on how AI can potentially propagate existing sources of bias, and ways through which organizations can mitigate their occurrence. As a consequence of this phenomenon, there is an increased focus on individuals and society concerning potential sources of bias when it comes to algorithmic decision-making, and ways in which it has been controlled for. Such concerns have been widespread, and include application

domains such as healthcare, public services, consumer services, and recruitment, among many others (Leavy, 2018). Furthermore, concerns about data ownership on digital platforms, as well as transparency of content forwarding, have been key concerns after the Cambridge Analytica case (Isaak and Hanna, 2018). Users of AI applications are therefore more aware of and concerned about how their personal data is handled, as well as for what purposes it is used (Pangrazio and Selwyn, 2019). Such concerns have also translated into revised guidelines, or even laws and regulations on organizations (Nemitz, 2018).

Apart from the above, several significant social events have redefined the agenda on environmental and social norms, which also have important implications for the management of technology, and particularly AI. For instance, environmental protection awareness has been set as a top priority over the last decade, particularly after the Paris Agreement (2015), which has also opened up the discussion about how AI should be governed in order to minimize resource use and contribute to important goals (Cortès et al., 2000). Likewise, focus on civil liberties of historically marginalized groups has redefined how AI technologies are designed and developed to be more inclusive (Mohammed and Nell'Watson, 2019; Shu and Liu, 2022).

The underlying theme in the foregoing discussion is that there is a concurrent push in constantly re-evaluating what responsible AI governance is and how it is deployed. On the one hand, the introduction of novel technologies creates phenomena that were previously unknown to us, introducing new opportunities but also new potential dangers and threats (Ashok et al., 2022). On the other hand, society and the corresponding norms and ethics that underpin it are in constant flux, which means that any corresponding practices of governing AI that are meant to align with them need to be redefined on a continuous basis (Shih, 2023). These forces in combination therefore influence what responsible AI governance encapsulates, and how it is enacted in organizations.

Organizational and Societal Effects

When debating about why organizations should deploy responsible AI governance practices, there is an implicit assumption that there is need to minimize any negative or unintended consequences of AI use, as well as to increase important organizational performance indicators (Minkkinen et al., 2021). While the former aspect of effects has been the primary focus of discussion over the last few years, the latter has received significantly less attention. We now know that responsible AI governance can lead to improved transparency on how AI systems are developed and operate, provide better explainability when it comes to high-stakes decision-making, reduce bias and unfair treatment of individuals and groups, as well as enhance auditability and accountability when required (Falco, 2019; van der Veer et al., 2021; Werder et al., 2022). As a result, it becomes clear that responsible AI governance facilitates the mitigation of the potentially negative effects that AI itself can generate (Mikalef et al., 2022).

Contrarily, although there are many anecdotal claims concerning the effects of responsible AI governance on key organizational processes and outcomes, there is substantially less research exploring them. Such effects are argued to be perceived both internally within organizations, as well as externally in their broader operational environment (Rakova et al., 2021). Internally, it is suggested that responsible AI governance can improve productivity in organizations by enhancing the perceptions of employees on the value that their organization delivers to society (Panch et al., 2018; West and Allen, 2020). This idea has been popularized based on prior findings in the corporate social responsibility literature, which has documented that adopting such practices significantly improves employee commitment to their organization, as well as their overall productivity (Saha et al., 2020). Furthermore, a recent study conducted by the Economist Intelligence Unit¹ indicates that approximately 80 percent of businesses believe that implementing responsible AI governance practices is critically important to them for talent acquisition and retention, thus providing an important leverage to foster human capital.

In addition, by promoting transparency of data collection, processing, and use throughout project lifecycles, responsible AI governance is argued to improve knowledge flows among organizational units, thus enhancing inter-departmental collaboration (Rantanen et al., 2021). Another suggested mechanism of value generation internally by responsible AI governance has to do with improved operational efficiency. According to Shekhar (2022), companies that adopt responsible AI governance experience higher returns on their AI investment, due to the fact that they have more rigorous practices on mitigating risks through training and testing data, measuring model bias and accuracy, and establishing model documentation. Doing so reduces the risk of AI applications producing predictions that are inaccurate due to oversight during development. In addition, there are early findings which show that companies that incorporate responsible AI practices throughout the product development lifecycle will build a competitive advantage through enhanced product quality (Papagiannidis et al., 2022).

When considering external effects from deployment of responsible AI governance, there are also several proposed paths through which value can be realized. One of the most discussed has to do with enhanced corporate reputation, which is suggested to improve customer engagement. The logic suggests that customers will gravitate towards organizations that have established responsible AI governance practices to ensure that no unintended or purposeful negative effects emerge during AI use (Gupta et al., 2023; Kumar et al., 2023). Furthermore, documented practices that ensure environmental and societal well-being through use of AI applications have been hinted to improve overall corporate reputation and customer attraction (Clarke, 2019). These effects, however, are not limited to customers of organizations, but are posited to influence partnership formation and supplier relationship management (Burkhardt et al., 2019). Finally, responsible AI governance has been associated with the emergence of sustainable business models and the creation of non-financial performance improvements (Mikalef et al., 2022). By redefining the key objectives and goals of organizations, responsible AI governance places a focus on values that are

seen as secondary for many organizations, such as environmental protection, societal coherency, and ethical conduct (Anagnostou et al., 2022).

FUTURE CHALLENGES AND OPPORTUNITIES

While there is growing interest from the academic and professional communities on designing and deploying responsible AI governance, there are still several topics that remain unexplored. Based on the proposed conceptual overview presented in Figure 7.2, some important areas with high relevance can already be identified. These discussed streams are indicative of some high-priority themes for organizations utilizing AI in their operations, as well as the broader ecosystem which is affected by such deployments. Nevertheless, it is by no means an exhaustive description of issues, and it is likely that additional research questions will emerge as we see more sophisticated forms of AI techniques and applications. A summary of some key themes and corresponding research questions that can guide future research can be found in Table 7.2.

Starting from the core concept of responsible AI governance, we still lack knowledge on how organizations transition from the ideation to the implementation process. In other words, there is still limited empirical knowledge about the process of designing a responsible AI governance scheme that follows corporate strategy. This is an area of inquiry that is also of high importance for practitioners, as they often navigate in uncharted waters when it comes to translating high-level responsible AI principles into actionable practices. Within this direction, there is also a lack of reference or maturity models which can pinpoint specific actions that can be taken by different stakeholders in order to ensure that responsible AI principles are followed in all relevant areas. The combination of understanding the process of how ideation moves to implementation, and how to construct maturity models to gauge the degree to which responsible AI governance is implemented, can provide organizations with practical tools to ensure that their AI systems adhere to responsible AI principles.

Adding to the above, the organizational environment in which responsible AI governance practices are developed and deployed is likely to strongly influence the different forms that governance schemes will take. Prior IS research has shown that contingencies of the environment have a strong impact on the types of governance approaches that are used, as well as how they are developed (Weber et al., 2009). Furthermore, organizations are characterized by a history which shapes values, norms, and their culture. Such deep-rooted elements are likely to exert a strong effect on the way responsible AI governance is deployed. In addition, path dependencies of older and more established firms may lead to issues of adapting to new ways of governing AI and implementing responsible principles. It is probable that newer and more agile organizations will create disruptions in their industries by building on responsible AI governance practices in the pursuit of digitally enabled sustainable business models. At the same time, we are likely to witness the emergence of new types of business models, such as AI-infused clean distributed energy grids, or smart

Table 7.2 Indicative themes and research questions

Theme	Research questions	Supporting references
Designing and deploying responsible AI governance	How is responsible AI governance strategically planned in organizations? What are the challenges of moving from ideation to implementation of responsible AI governance? How can maturity models support the deployment of responsible AI governance practices in organizations?	Cheng et al. (2021); Ghallab (2019); Jantunen et al. (2021); Schiff et al. (2020a)
Contingencies of responsible AI governance	How does resource allocation influence the adoption of responsible AI governance? How do industry pressures influence responsible AI governance adoption from organizations? What is the effect of path dependencies on adopting responsible AI governance? What configurations of contingencies facilitate higher diffusion of responsible AI governance?	Abedin (2021); Dignum (2019); Laut et al. (2021); Li et al. (2022); Rana et al. (2022); Trocin et al. (2021)
Adapting to external stimuli	What mechanisms should organizations establish to quickly identify important external stimuli? What structures are needed to rapidly adapt to changes necessitated by the external environment? How do laws and regulations influence the responsible AI governance practices that are adopted and used by organizations?	Arthaud-Day (2005); Askeff et al. (2019); Rakova et al. (2021); Sidorenko et al. (2020)
Emergent business models and forms of operating	How can organizations leverage their responsible AI governance practices in order to introduce novel business models? How does the adoption of responsible AI governance change the priorities of organizations? What new key performance indicators are important in the age of responsible AI governance, and how can we measure them?	Di Vaio et al. (2020); Langenbucher (2020); Shekhar (2022); Zhu et al. (2022)

decentralized energy systems. Such business models both depend and build on well-defined responsible AI governance schemes.

An external view of the organization also necessitates an understanding of how organizations identify emerging signals and rapidly adapt to them. For instance, when there are new trending topics that concern social and ethical issues, it is important that organizations quickly flag them and understand how they might influence their use of technology. Headlines over Amazon’s sexist hiring algorithms meant that any organizations using AI in order to hire individuals, or using algorithms to aid decision-making over individuals, were automatically subjected to more scrutiny (Lavanchy, 2018). Thus, it becomes critical to develop mechanisms to rapidly identify, make sense of, and adapt to emerging social issues. Furthermore, changes in laws and regulations require organizations to adapt the way they manage their AI systems, which may render certain practices incompatible with new directives. An interesting area for future research to explore is how such laws and regulations

influence the uptake of responsible AI governance, or the specific practices that organizations actually implement.

A final area of research which is likely to provide rich insights concerns the study of emerging business models and alternative forms of strategizing based on responsible AI governance practices. Many of the principles upon which responsible AI governance is grounded emphasize social and environmental well-being over economic gains. These two goals have until recently been seen as contrarian, so it is becoming critical that organizations plan for ways of bridging them in order to remain competitive in the long run. Recent cases of unethical corporate actions have prompted consumers to seek other alternatives, and with the prevalence of social media negative publicity has managed to spread like wildfire. Thus, responsible use of AI at all levels within organizations will be a cornerstone of organizational business models. This shift also entails a need to better understand the value of corporate relationships with customers, and how these may be mediated through the use of AI. A challenge is therefore to understand how the perceptions of customers on the organizational use of AI influences their attitudes and actions. Finally, in the coming years we will likely see agencies that conduct audits and provide certifications to companies concerning the responsible use of AI. Doing so will facilitate greater transparency for the black box of organizational use of AI, and provide direct and indirect users of these systems with more power over how organizations design and deploy AI systems.

NOTE

- 1 <https://www.eiu.com/n/staying-ahead-of-the-curve-the-business-case-for-responsible-ai/>.

REFERENCES

- Abedin, B. (2021). Managing the tension between opposing effects of explainability of artificial intelligence: a contingency theory perspective. *Internet Research*, 32(2), 425–453.
- Acemoglu, D., and Restrepo, P. (2018). Artificial intelligence, automation, and work. NBER Working Paper No. 24196, National Bureau of Economic Research, Cambridge, MA.
- Ågerfalk, P.J. (2020). Artificial intelligence as digital agency. *European Journal of Information Systems*, 29(1), 1–8.
- Anagnostou, M., Karvounidou, O., Katritzidaki, C., Kechagia, C., Melidou, K., Mpeza, E., Konstantinidis, I., Kapantai, E., Berberidis, C., and Magnisalis, I. (2022). Characteristics and challenges in the industries towards responsible AI: a systematic literature review. *Ethics and Information Technology*, 24(3), 1–18.
- Anderson, M., and Anderson, S.L. (2007). The status of machine ethics: a report from the AAAI Symposium. *Minds and Machines*, 17(1), 1–10.
- Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-López, S., Molina, D., and Benjamins, R. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115.

- Arthaud-Day, M.L. (2005). Transnational corporate social responsibility: a tri-dimensional approach to international CSR Research. *Business Ethics Quarterly*, 15(1), 1–22.
- Ashok, M., Madan, R., Joha, A., and Sivarajah, U. (2022). Ethical framework for artificial intelligence and digital technologies. *International Journal of Information Management*, 62, 102433.
- Ashraf, A.R., Thongpapanl, N., and Auh, S. (2014). The application of the technology acceptance model under different cultural contexts: the case of online shopping adoption. *Journal of International Marketing*, 22(3), 68–93.
- Askell, A., Brundage, M., and Hadfield, G. (2019). The role of cooperation in responsible AI development. *arXiv preprint arXiv:1907.04534*.
- Baan, W., Thomas, C., and Chang, J. (2017). *How advanced industrial companies should approach artificial intelligence strategy*. McKinsey & Company, November.
- Bücker, M., Szepannek, G., Gosiewska, A., and Biecek, P. (2022). Transparency, auditability, and explainability of machine learning models in credit scoring. *Journal of the Operational Research Society*, 73(1), 70–90.
- Burkhardt, R., Hohn, N., and Wigley, C. (2019). *Leading your organization to responsible AI*. McKinsey Analytics.
- Cheng, L., Varshney, K.R., and Liu, H. (2021). Socially responsible AI algorithms: issues, purposes, and challenges. *Journal of Artificial Intelligence Research*, 71, 1137–1181.
- Cihon, P., Maas, M.M., and Kemp, L. (2020). Should artificial intelligence governance be centralised? Design lessons from history. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.
- Clarke, R. (2019). Principles and business processes for responsible AI. *Computer Law and Security Review*, 35(4), 410–422.
- Cortès, U., Sánchez-Marré, M., Ceccaroni, L., and Poch, M. (2000). Artificial intelligence and environmental decision support systems. *Applied Intelligence*, 13(1), 77–91.
- De Haes, S., and Van Grembergen, W. (2004). IT governance and its mechanisms. *Information Systems Control Journal*, 1, 27–33.
- Deshpande, A., and Sharp, H. (2022). Responsible AI systems: who are the stakeholders? Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society.
- Di Vaio, A., Palladino, R., Hassan, R., and Escobar, O. (2020). Artificial intelligence and business models in the sustainable development goals perspective: a systematic literature review. *Journal of Business Research*, 121, 283–314.
- Dignum, V. (2019). *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer Nature.
- Dobbe, R., Gilbert, T.K., and Mintz, Y. (2021). Hard choices in artificial intelligence. *Artificial Intelligence*, 300, 103555.
- EC (2019). *High-Level Expert Group on Artificial Intelligence: Ethics Guidelines for Trustworthy AI*. European Commission.
- Eitel-Porter, R. (2021). Beyond the promise: implementing ethical AI. *AI and Ethics*, 1(1), 73–80.
- Emerson, T.L., and Conroy, S.J. (2004). Have ethical attitudes changed? An intertemporal comparison of the ethical perceptions of college students in 1985 and 2001. *Journal of Business Ethics*, 50(2), 167–176.
- Enholm, I. M., Papagiannidis, E., Mikalef, P., and Krogstie, J. (2022). Artificial intelligence and business value: a literature review. *Information Systems Frontiers*, 24(5), 1709–1734.
- Falco, G. (2019). Participatory AI: reducing AI bias and developing socially responsible AI in smart cities. 2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC),

- Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., and Srikumar, M. (2020). Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication (2020–21).
- Floridi, L. (2019). Establishing the rules for building trustworthy AI. *Nature Machine Intelligence*, 1(6), 261–262.
- Ghallab, M. (2019). Responsible AI: requirements and challenges. *AI Perspectives*, 1(1), 1–7.
- Graham, J., Amos, B., and Plumptre, T.W. (2003). *Governance Principles for Protected Areas in the 21st Century*. Institute on Governance, Governance Principles for Protected Areas Ottawa.
- Gupta, S., Kamboj, S., and Bag, S. (2023). Role of risks in the development of responsible artificial intelligence in the digital healthcare domain. *Information Systems Frontiers*, 25(5), 2257–2274.
- Hughes, C., Robert, L., Frady, K., and Arroyos, A. (2019). Artificial intelligence, employee engagement, fairness, and job outcomes. In *Managing Technology and Middle- and Low-Skilled Employees*. The Changing Context of Managing People series. Bingley: Emerald Publishing, pp. 61–68.
- Isaak, J., and Hanna, M.J. (2018). User data privacy: Facebook, Cambridge Analytica, and privacy protection. *Computer*, 51(8), 56–59.
- Jantunen, M., Halme, E., Vakkuri, V., Kemell, K.-K., Rebekah, R., Mikkonen, T., Nguyen Duc, A., and Abrahamsson, P. (2021). Building a maturity model for developing ethically aligned AI systems. IRIS.
- Jarrahi, M.H.J.B.H. (2018). Artificial intelligence and the future of work: human–AI symbiosis in organizational decision making. *Business Horizons*, 61(4), 577–586.
- Kumar, P., Dwivedi, Y.K., and Anand, A. (2023). Responsible artificial intelligence (AI) for value formation and market performance in healthcare: the mediating role of patient's cognitive engagement. *Information Systems Frontiers*, 25(5), 2197–2220.
- Lam, K., Iqbal, F.M., Purkayastha, S., and Kinross, J.M. (2021). Investigating the ethical and data governance issues of artificial intelligence in surgery: protocol for a Delphi study. *JMIR Research Protocols*, 10(2), e26552.
- Langenbucher, K. (2020). Responsible AI-based credit scoring—a legal framework. *European Business Law Review*, 31(4), 527–572.
- Laut, P., Dumbach, P., and Eskofier, B.M. (2021). Integration of artificial intelligence in the organizational adoption—a configurational perspective. ICIS 2021 Proceedings.
- Lavanchy, M. (2018). Amazon's sexist hiring algorithm could still be better than a human. *Phys.org*, November 1. <https://phys.org/news/2018-11-amazon-sexist-hiring-algorithm-human.html>.
- Leavy, S. (2018). Gender bias in artificial intelligence: the need for diversity and gender theory in machine learning. Proceedings of the 1st International Workshop on Gender Equality in Software Engineering.
- Li, M., Wan, Y., and Gao, J. (2022). What drives the ethical acceptance of deep synthesis applications? A fuzzy set qualitative comparative analysis. *Computers in Human Behavior*, 133, 107286.
- Liu, H.-W., Lin, C.-F., and Chen, Y.-J. (2019). Beyond State v Loomis: artificial intelligence, government algorithmization and accountability. *International Journal of Law and Information Technology*, 27(2), 122–141.
- Manheim, K., and Kaplan, L. (2019). Artificial intelligence: risks to privacy and democracy. *Yale JL and Tech.*, 21, 106.
- Mayer, A.-S., Strich, F., and Fiedler, M. (2020). Unintended consequences of introducing AI systems for decision making. *MIS Quarterly Executive*, 19(4), 239–257.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6), 1–35.

- Meske, C., Bunde, E., Schneider, J., and Gersch, M. (2022). Explainable artificial intelligence: objectives, stakeholders, and future research opportunities. *Information Systems Management*, 39(1), 53–63.
- Mikalef, P., Conboy, K., Lundström, J.E., and Popovič, A. (2022). Thinking responsibly about responsible AI and “the dark side” of AI. *European Journal of Information Systems*, 31(3), 257–268.
- Mills, S., Baltassis, E., Santinelli, M., Carlisi, C., Duranton, S., and Gallego, A. (2020). Six steps to bridge the responsible AI gap. <https://www.bcg.com/en-us/publications/2020/six-steps-for-socially-responsible-artificial-intelligence>.
- Minkinen, M., Zimmer, M.P., and Mäntymäki, M. (2021). Towards ecosystems for responsible AI. In: Dennehy, D., Griva, A., Pouloudi, N., Dwivedi, Y.K., Pappas, I., and Mäntymäki, M. (eds), *Responsible AI and Analytics for an Ethical and Inclusive Digitized Society*. I3E 2021. Lecture Notes in Computer Science, vol. 12896. Cham: Springer. https://doi.org/10.1007/978-3-030-85447-8_20.
- Mohammed, P.S., and Nell’Watson, E. (2019). Towards inclusive education in the age of artificial intelligence: Perspectives, challenges, and opportunities. In: Knox, J., Wang, Y., Gallagher, M. (eds), *Artificial Intelligence and Inclusive Education. Perspectives on Rethinking and Reforming Education*. Singapore: Springer. https://doi.org/10.1007/978-981-13-8161-4_2.
- Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133), 20180089.
- Panch, T., Szolovits, P., and Atun, R. (2018). Artificial intelligence, machine learning and health systems. *Journal of Global Health*, 8(2). Doi: 10.7189/jogh.08.020303.
- Pangrazio, L., and Selwyn, N. (2019). “Personal data literacies”: a critical literacies approach to enhancing understandings of personal digital data. *New Media and Society*, 21(2), 419–437.
- Papagiannidis, E., Mikalef, P., Krogstie, J., and Conboy, K. (2022). From responsible AI governance to competitive performance: the mediating role of knowledge management capabilities. Conference on e-Business, e-Services and e-Society.
- Rakova, B., Yang, J., Cramer, H., and Chowdhury, R. (2021). Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human–Computer Interaction*, 5(CSCW1), 1–23.
- Rana, N.P., Chatterjee, S., Dwivedi, Y.K., and Akter, S. (2022). Understanding dark side of artificial intelligence (AI) integrated business analytics: assessing firm’s operational inefficiency and competitiveness. *European Journal of Information Systems*, 31(3), 364–387.
- Rantanen, E.M., Lee, J.D., Darveau, K., Miller, D.B., Intriligator, J., and Sawyer, B.D. (2021). Ethics education of human factors engineers for responsible AI development. Proceedings of the Human Factors and Ergonomics Society Annual Meeting.
- Saha, R., Cerchione, R., Singh, R., and Dahiya, R. (2020). Effect of ethical leadership and corporate social responsibility on firm performance: a systematic review. *Corporate Social Responsibility and Environmental Management*, 27(2), 409–429.
- Schiff, D., Biddle, J., Borenstein, J., and Laas, K. (2020a). What’s next for AI ethics, policy, and governance? A global overview. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.
- Schiff, D., Rakova, B., Ayes, A., Fanti, A., and Lennon, M. (2020b). Principles to practices for responsible AI: closing the gap. *arXiv preprint arXiv:2006.04707*.
- Shekhar, R. (2022). Responsible artificial intelligence is good business. *LSE Business Review*. <https://blogs.lse.ac.uk/businessreview/2022/08/30/responsible-artificial-intelligence-is-good-business/>.

- Shih, T. (2023) Research funders play an important role in fostering research integrity and responsible internationalization in a multipolar world. *Accountability in Research*, 1–10. DOI: 10.1080/08989621.2023.2165917
- Shin, Y. (2019). The spring of artificial intelligence in its global winter. *IEEE Annals of the History of Computing*, 41(4), 71–82.
- Shneiderman, B. (2021). Responsible AI: bridging from ethics to practice. *Communications of the ACM*, 64(8), 32–35.
- Shu, Q., and Liu, H. (2022). Application of artificial intelligence computing in the universal design of aging and healthy housing. *Computational Intelligence and Neuroscience*, 2022. <https://doi.org/10.1155/2022/4576397>.
- Sidorenko, E., Arzumanova, L., and Amvrosova, O. (2020). Adaptability and flexibility of law in the context of digitalization. International Scientific and Practical Conference.
- Smuha, N.A. (2019). The EU approach to ethics guidelines for trustworthy artificial intelligence. *Computer Law Review International*, 20(4), 97–106.
- Sturm, T., Fecho, M., and Buxmann, P. (2021). To use or not to use artificial intelligence? A framework for the ideation and evaluation of problems to be solved with artificial intelligence. Proceedings of the 54th Hawaii International Conference on System Sciences.
- Taeiagh, A. (2021). Governance of artificial intelligence. *Policy and Society*, 40(2), 137–157.
- Tallon, P.P., Ramirez, R.V., and Short, J.E. (2013). The information artifact in IT governance: toward a theory of information governance. *Journal of Management Information Systems*, 30(3), 141–178.
- Theodorou, A., and Dignum, V. (2020). Towards ethical and socio-legal governance in AI. *Nature Machine Intelligence*, 2(1), 10–12.
- Trocin, C., Mikalef, P., Papamitsiou, Z., and Conboy, K. (2021). Responsible AI for digital health: a synthesis and a research agenda. *Information Systems Frontiers*, 25(5), 2139–2157 (2023). <https://doi.org/10.1007/s10796-021-10146-4>.
- van der Veer, S.N., Riste, L., Cheraghi-Sohi, S., Phipps, D.L., Tully, M.P., Bozentko, K., Atwood, S., Hubbard, A., Wiper, C., and Oswald, M. (2021). Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries. *Journal of the American Medical Informatics Association*, 28(10), 2128–2138.
- Van Grembergen, W., De Haes, S., and Guldentops, E. (2004). Structures, processes and relational mechanisms for IT governance. In Van Grembergen, W. (ed.), *Strategies for Information Technology Governance* (pp. 1–36). Igi Global.
- van Rijmenam, M., and Schweitzer, J. (2018). How to build responsible AI? Lessons for governance from a conversation with Tay. AOM Specialized Conference: Big Data and Managing in a Digital Economy.
- Wang, W., and Siau, K. (2019). Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity: a review and research agenda. *Journal of Database Management (JDM)*, 30(1), 61–79.
- Wang, Y., Xiong, M., and Olya, H. (2020). Toward an understanding of responsible artificial intelligence practices. Proceedings of the 53rd Hawaii International Conference on System Sciences.
- Weber, K., Otto, B., and Österle, H. (2009). One size does not fit all—a contingency approach to data governance. *Journal of Data and Information Quality (JDIQ)*, 1(1), 4.
- Weill, P., and Ross, J. (2005). A matrixed approach to designing IT governance. *MIT Sloan Management Review*, 46(2), 26.
- Werder, K., Ramesh, B., and Zhang, R. (2022). Establishing data provenance for responsible artificial intelligence systems. *ACM Transactions on Management Information Systems (TMIS)*, 13(2), 1–23.
- West, D.M., and Allen, J.R. (2020). *Turning Point: Policymaking in the Era of Artificial Intelligence*. Washington, DC: Brookings Institution Press.

- Yatsenko, O. (2021). Polysubjectivity and Contextuality of the Ethical in Contemporary Digital Culture. *Future Human Image*, 15(15), 148–158.
- Zhu, L., Xu, X., Lu, Q., Governatori, G., and Whittle, J. (2022). AI and Ethics—Operationalizing Responsible AI. In Chen, F. and Zhou, J. (eds), *Humanity Driven AI* (pp. 15–33). Cham: Springer.

PART II

MAKING DECISIONS WITH AI

8. Human judgment in the age of automated decision-making systems

*Dina Koutsikouri, Lena Hylving, Jonna Bornemark,
and Susanne Lindberg*

INTRODUCTION

In organizations across sectors, more and more people make decisions with the help of algorithms, or act upon automated decisions that are made without human involvement. We see, for example, that robots make decisions about financial assistance (Eubanks, 2018), and that parts of the employment service’s assessment support have been automated (Gal et al., 2020). The use of automated decision-making systems might give a sense of efficiency, quality, and impartiality (Ranerup and Henriksen, 2019; Neuroth et al., 2000; Zheng et al., 2021). However, studies (e.g., Hylving and Lindberg, 2022) show that challenges arise when we attach greater importance to the amount of data than to crucial details in making an appropriate decision. Overreliance on large amounts of data over human judgment reflects an expectation that automated decision-making systems (ADS) allow for more objective decisions than those made by humans, who may be influenced by prejudice, conflicts of interests, or fatigue (Lepri et al., 2021). This view is problematic, at least when artificial intelligence (AI)-based technology is applied to human activity. For example, it has been shown that knowledge-intensive work such as scientific practice also relies on (professional) judgment, which cannot be broken down into rules and automated (Ratti, 2019).

In automated decision-making, AI is based on calculating capacity and rational logic (that is, $1 + 1 = 2$). This, together with an immense amount of data that lays the ground for both pattern recognition and what stands out in the data is used when automated decisions are being calculated. Humans, on the other hand, make decisions guided by judgment (for example, aspects of not-knowing, sensibility, emotions, and experiential knowledge); asking questions such as: Is this reasonable? What is important in this specific situation? To what are we blind? Is this ethical? (Bornemark, 2018). This development accentuates the need for understanding what sort of intelligence humans have in relation to what AI aims to do (Smith, 2019). What has slipped under the radar of much scholarly attention is the role of human judgment to reinforce and augment AI-based decision-making and vice versa. As ADS are being rapidly deployed into the workplace, it is important to explore judgment since they lack the capabilities related to aspects of emotions, sensibility, and subjectivity that are at stake in “acting wisely” in a given situation, and which are the very core of human decision-making (Irwin, 1999). This prospect raises fundamental philosophical issues about the role of human judgment in relation to the capability

of ADS. The boundary becomes blurred between what human judgment is, and what the purpose of this competence is now that powerful technology such as AI exists to make decisions.

In this chapter we are interested in unpacking human judgment. We take as a starting point that judgment is opaque, and therefore it is difficult to explain how it works in decision-making. Interestingly, the same has been said about algorithms. Therefore, we explore the following questions: What are the different components of human judgment, and how are they manifested in human–AI decision-making? To achieve this exploration, we draw on Aristotle’s perspective of *phronesis*, which we understand as a kind of judgment, to compare human and automated decision-making.

We propose that human judgment is a continually evolving well of knowledge sources that guides action and decision-making in the world as it is. We find that judgment comprises at least eight elements: *what-ness*, not knowing, emotions, sensory perception, lived experience, intuition, *episteme*, and *techné*. By showing the resources of *phronesis* we contribute with a vocabulary to express what cannot (yet) be automated in decision-making. But also, that human involvement in an increasingly digitalized society needs to be cultivated in order to flourish and steer the direction of technology with the purpose of making “wise” decisions. Unpacking *phronesis* lays bare what sort of intelligence resides in humans, and what sort is displayed within the domain of emergent technologies such as ADS. *Phronesis* reflects of the way humans are, which is honed through culture, education, life experience, and contemplation. Awareness of what is different can inspire us to think more deeply about how “we might use the advent of AI to raise the standards on what it is to be human” (Smith, 2019, p. xvii). We believe that it can also help us to better understand how to cultivate human–AI decision-making in organizations.

ARTIFICIAL INTELLIGENCE

We witness today how AI is increasingly being deployed and entangled in our personal and professional lives, from before we are born (Dias and Torkamani, 2019), until we die (Lu, 2019; Wiederhold, 2019), and beyond. AI is intertwined in everyday activities ranging from dating and mating (Slater, 2013; Sumter et al., 2017) to epidemiologic research (Mayer-Schönberger and Cuckier, 2013) and is rapidly becoming a pervasive aspect of the present (Russell, 2019). While it is impossible to predict exactly how it will develop, or on what timeline, it is clear that the technology which enables AI is continuously advancing, as well as becoming increasingly affordable for the many (Fountain et al., 2019). How this shapes our world is as yet unknown (Baskerville et al., 2020); however, it is important to explore to drive the frontier of technology and knowledge forward, because it is the dominant technology of the future (Berente et al., 2021).

“AI” is an umbrella term that covers several different types of technologies and systems, including machine learning, robotics, computer vision, and natural language processing. Here we focus on one specific type of AI technology, namely automated

decision-making systems. ADS aim to aid or replace human decision-making based on rules and statistical and mathematical algorithms, often in context-sensitive settings where details make a difference, and where the application of judgment (phronesis) is critical (Kolbjørnsrud et al., 2016).

Automated Decision-Making Systems

ADS can take many forms and uses; at their most general level, ADS are algorithms that are used to collect, process, and model data to make decisions or recommendations for decisions, and then in turn use these decisions to improve the system itself (Araujo et al., 2020). The role of the human varies, from non-existent to having full autonomy. For example, the range goes from fully automated ADS that merely communicate and implement decisions, to recommender systems that the user can ignore (Araujo et al., 2020). Pre-existing and emerging biases have been an unrelenting issue for ADS (Dobbe et al., 2018). There are advocates of always ensuring that human beings have agency in automated decision-making, though it is not currently done (Wagner, 2019). Others instead focus on the process of designing and implementing ADS in a value-centric way (Dobbe et al., 2018). A third stream of research is focused on the governance perspective, arguing for the need of ethical auditing of ADS (Mökander and Axente, 2021).

The assumption is that ADS lead to more efficient, effective, and objective decisions; yet this is not always the case (Araujo et al., 2020). We have also seen systems that involve humans to “rubber stamp” automated decisions to avoid the necessary regulations placed on fully automated systems, which have led to misplaced liability and biased systems (Wagner, 2019; Mökander and Axente, 2021). Further, studies of AI advice and its loss and effects have shown that unique human knowledge can decrease, making humans act as “borgs,” without individuality (Fügenger et al., 2021).

ADS can be designed and built to emulate different human capacities and capabilities (Hussain, 2018). It is even considered to outperform humans in many domains (Boström, 2017). Yet, one of the problems with ADS is that even though they are supposed to emulate human capabilities, many building blocks are malfunctioning (Eubanks, 2018; O’Neil, 2016) or missing (Smith, 2019). Indeed, advocates of the human-augmented ADS highlight that not only does this need to be better understood, but also it may create new perspectives of the role of the human in relation to information technology (IT) in organizations (Teodorescu et al., 2021).

The intensification of implementation and usage of digital technology, including ADS, have in many ways put the human in the background. For one, it has been considered that by getting rid of “the human factor,” better and more objective decisions and accurate predictions can be reached. Yet, this has been proven wrong in many ways (Eubanks, 2018; O’Neil, 2016). In this light, we argue that we must get to grips with the human capacity for judgment, and how to sustain human agency in a time where ADS, on the surface, save us time by taking over more and more of our mundane practical judgments (for example, who to hire, whether to approve a loan, where to allocate resources). Here we approach phronesis and ADS from the vantage

point of discovering not only how judgment works, but also how ADS and humans can work together and augment each other.

PHRONESIS AS HUMAN JUDGMENT

To pick up some differences between human judgment and automated decision-making, it can be fruitful to turn to the Greek philosopher Aristotle. Aristotle presented a variety of knowledge forms, where it could be argued that some are close to the kind of knowledge that the automated decision-making system produces (epistemic knowledge, which is general and abstract), and others are closer to human judgment (phronesis, as discussed below) (Aristotle, 2009). While digital technology is superior in some respects to humans, such as visuospatial processing speed and pattern recognition, it lacks human reasoning, creativity, and empathy, what Aristotle called phronesis; often referred to as practical wisdom (Jeste et al., 2020; Kase et al., 2014). The Aristotelian notion of phronesis reflects the aspects that people in practice situations would refer to as being “the reasonable thing to do,” in relation to the particulars of the specific situation (Shotter and Tsoukas, 2014).

In Arendt’s philosophy of judgment there are two sides, that have sometimes been read as contradictory (Yar, 2000). She describes the person with judgment as an actor on one hand, and as a spectator on the other hand; while Marshall (2010) points out that both these two sides are needed in a sound judgment. We need to be able to act, even if we do not know everything in the situation; in that way the person with judgment is an actor. On the other hand, precisely because we do not know everything, we need to keep listening and be open to what we do not know, that is, maintain a spectating position.

Phronesis in relation to this can be considered a kind of judgment that focuses on the here and now, and the concrete situation at hand. Thus, it is therefore often referred to as situational knowledge (Shotter and Tsoukas, 2014). In practice, we often ask ourselves what we are supposed to do in a particular situation, and in relation to the person we are interacting with. It is about bringing timing into the equation, and knowing what, when, and precisely how to do something. It is closely associated to action even when not every aspect of the situation is known. The action-oriented nature of phronesis makes it critical to form sound ethical judgments (Begley, 2006). It is one’s capacity to direct action (Polansky, 2000) and is “a reasoned and true state of capacity to act with regard to human goods” (Aristotle, 2009). It begins from the apprehension of “what should be” and whether a particular action should be done in a particular circumstance (Baggini and Fosl, 2020).

Aristotle asserts that the highest form of human well-being is the life controlled by reason (Beauchamp, 1991). Thus, sound judgment plays an important role in being wise, and it refers to both theoretical reasoning (the apprehension of what the truth is) and practical reasoning (the apprehension of whether a particular action should be done), as well as ethics and moral action.

Here we draw on the Aristotelian view of phronesis and appropriate the concept in a way that is relevant to understand how it is embodied in how practitioners arrive at professional judgment in everyday decision-making situations. When carrying out this analysis, we rely on philosophical thinking based on prior work by Hannah Arendt and her portrayal of judgment, Edmund Husserl's (2014) description of the meaning of "lived experience," Nussbaum's (2001) description of the relevance of emotions in cognition, Merleau-Ponty's (1974) illustration of the role of sensory perception in sense-making, and Gadamer's (2004) view of phronesis as a model for his development of hermeneutics. It is on this understanding, coupled with empirical evidence drawn from explorations of practical knowledge where making judgments forms an important part of the job (for example, nurses, doctors, police officers, teachers), that we have sought to elucidate the components of human judgment. While we recognize that the list of components is not exhaustive, we suggest that these constitute key pillars for understanding the role of judgment in decision-making.

Although much of the analysis is couched in terms of the comparison between human capacities and the type of ability enabled by ADS, the aim is to illuminate that judgment is a type of phronesis that motivates human rather than machine action.

Components of Human Judgment

Calculating rationality describes how we turn to rules of abstractions and generalization to organize and make sense of our world (Hylving et al., 2022). We argue that too much of this type of rationality makes us lose contact with ourselves, others, and the specific situation in a way that disables us from developing judgment. Instead, we rely on external parameters to objectively guide our action. Practices built on phronesis, on the contrary, emphasize the subjective, emotional, and temporary, and our ability to "not know," but to learn to cope with insecurity, instability, and anxiety, and find ways to act in such terrains (Shotter and Tsoukas, 2014). Practices based on a calculating rationality, with rules and abstractions, and the opposite, where creativity and the not-known are in focus, are interdependent yet both are needed (Bornemark, 2018). But in modern societies capacities connected to phronesis have been suppressed, overlooked, and seen as a state of lack of better knowledge (Kristjánsson et al., 2021).

We delineate eight key components of phronesis that shed light on the importance of human agency in automated decision-making. These components are sorted out through a phenomenological investigation on how human decision-making is performed in inter-human professions. Here we build upon these, but also develop them further in the context of professional decision-making (Hylving and Koutsikouri, 2020). These components include not-knowing (openness to possibilities), whatness (what is important and valued highly), emotions, lived sensory perception, experiences (one's own and those of others), intuition (seeing a solution without deliberate thinking), episteme, and techne, and can be seen as fluid knowledge sources that enable attention to what the situation requires, rather than to what can be fully controlled through well-defined categories in a calculable system. Human judgment,

and to judge well, require a continuous movement between these components in an evaluative way, asking: what is important in this situation? How and when should the components be weighed in, and what potential conflicts or paradoxes do they entail?

On this basis, now that organizations are faced with the challenge of developing ethical, responsible automated decision-making systems with embedded artificial intelligence, phronesis, or human judgment, as a knowledge form takes on a new urgency. Next, we take a closer look at the components that we consider central in human judgment within the context of professional decision-making.

Not-knowing

When professionals make decisions the aspect of not-knowing is ever present (Souto, 2019). This form of knowledge relates to answering questions such as: Where does this problem (situation) begin and where does it end? What is important? What is at stake? For example, when an injured person goes to an emergency room for medical care, it is the job of the nurse to gather information and assess the patient's health information using evidence-informed tools including posing questions to make an appropriate assessment. This includes considering that the situation also contains dimensions of not-knowing. It means that while it is possible to consider information that is possible to capture from an initial analysis of the patient's health, there will always be aspects about a person's health status that are not possible to sort and gather. That is, a skilled nurse is acutely aware that what may seem a routine situation may well hide something more serious, and that it is important to remain sensitive to the possibility of not-knowing as part of deciding on patient health status. This entails the propensity to display humility and continuous listening.

Likewise, being able to discern when to take appropriate action even when facing an indeterminate situation (Shotter and Tsoukas, 2014): coming to a judgment then involves consciousness about one's own finite ability to know all aspects of a situation and hence perception of phronesis.

In this way it relates to the wicked problems described by Rittel and Webber (1973). They posit that wicked problems regard such questions as where the situation begins and where it ends, what the ultimate goal is, and what the central problem is. These problems do not have a final answer; rather, the way we temporarily answer these questions posits the way we understand the situation.

Automated decision-making systems do not have the capacity to relate to aspects of not-knowing as it cannot change its ultimate goals, and thus cannot emulate what is involved in this component of phronesis. For automated decision-making to function, the system requires data, known data, and it requires already-set goals (Russell, 2019). This technology can either use pre-organized data or learn from previously used data in decision-making.

Whatness

We take inspiration from Bornemark (2018, 2020) to explain whatness. She points to how practicing phronesis can make it possible to find alternative pathways in challenging and uncharted waters. It is the ability to attune to the particulars of the

situation and assess “what” is important in this moment, and which guides action (what we choose to do). All situations include many possible whatnesses, and the competence to pick up which are crucial in a certain situation can be trained, and humans can become better at this during the course of their professional development. Attuning to different whatnesses guides our values and our actions. A history teacher, for example, has to attune to what a student should learn in a certain situation, when the focus should be on the facts of a certain historical situation and when they should reflect upon what can be learned from this situation. The teacher also needs to pick up on when the important whatness is the learning content, and when it is more pressing to address a conflict between two students. Mobilizing phronesis requires one to be in relation to what is important in this specific situation. In new situations, a new whatness escapes a clear definition; it is sensed and has to be drawn out of the situation. Nonetheless, our lives are controlled by what is important.

In relation to decision-making the whatness of the situation is always central. When decision-making is easy, everyone involved agrees upon which whatnesses should guide the action; but when the situation is conflicted, we often see different opinions where the participants in the situation pick up on different whatnesses. The whatnesses are also closely connected to values and value formation. “What” appears as important shapes values, and values that we bring with us into the situation guide which whatnesses we pick up on.

Automated decision-making does not have this competence to relate to different whatnesses in different situations, and to pick up new whatnesses, as the whatness embedded in an automated decision-making system is pre-decided in the algorithm.

Emotions

Another central component of phronesis is emotions. Emotions routinely affect how and what we see. Emotions are often perceived as subjective and therefore something that is not relevant as information for action in professional contexts (Schwartz and Sharpe, 2010). Further, emotions are frequently associated with personal opinions and concerns. Foregrounding emotions as a knowledge source acknowledges the non-optical factors of being in a situation. The flow of sensory perceptions stemming from our environment is not neutral information. It carries vital information on what is important in a situation, and therefore provides a motivating influence on what to do next. Schwartz and Sharpe (2010, p. 71) suggest that it is emotion that compels us to act. Hence, emotion alerts us to something that demands our attention, and that action is required. In the same vein, Nussbaum (2001) posits that emotions are part of a cognitive process and are thus about information processing. In a professional capacity the key is to attune to one’s own as well as others’ emotions, and to acknowledge them as relevant information carriers for potential actions.

Strong emotions can shut down the ability to be sensitive to context and respond to the situation as it is now (Schwartz and Sharpe, 2010). Thus, the propensity to attune to but also relate to the layers of emotions in an impartial way is a prerequisite for developing judgment. Emotions are also necessary to make value judgments;

emotions point out what is important, and what we consider good and bad. Emotions thus point out the direction in a lived experience.

We argue in this chapter that current forms of automated decision-making systems are not alive and do not possess emotions, and hence do not have the capacity for emotional involvement that is required for acquiring and developing the type of expertise displayed by, for example, “caring” and “wise” professionals, judges, police officers, and teachers (e.g., Constantinescu and Crisp, 2022). But in phronesis, emotion is a central component.

Sensory perception

One of the knowledge sources that, for example, nurses in the emergency room (ER) heavily rely on is sensory perception. It is through the sensory organs, seeing, sensing, hearing, smelling, and tasting, that we perceive information about a certain situation. Without our bodies we are unable to relate to the particulars of the situation (its uniqueness). However, while sensations help to collect what is important in a situation, they are much richer than we can fathom and verbalize; one sensation can carry infinite opportunities to distil valuable knowledge (Bornemark, 2018). The challenge is to remain open to the incoming sensations despite the influence of prior knowledge and prejudice. The Covid-19 pandemic brought to light the importance of sensory perception in terms of foregrounding the limitations of digital platforms in, for example, education. Teaching online emphasized listening, while placing less importance on other sensory impressions compared to physical on-campus teaching, where it included most sensory organs and richer sensations to be experienced. Sensory perception is not only a collection of data, but is also intertwined with meaning and emotions, and part of our sense-making, where not everything is of equal importance (Merleau-Ponty, 1974). Professional judgment also includes the capacity to take both one’s own and others’ sense perception into account; the sense perception of, for example, a colleague can be of crucial value in working out situations and knowing what goals to pursue.

In sum, automated decision-making systems might record sound, vibrations, or temperature, and so on, but this is not the same thing as lived sensory perception as it is disconnected from lived sense-making and valuation. The introduction of automated decision systems in health care, for example, show that the capacity of AI is to build competence through data (from sensors); however, it cannot mobilize resonance in terms of forming meaningful connections with patients (and their social circumstances) (Lebovitz et al., 2021). Research in medical diagnosis shows that professional knowledge workers form their final judgment by synthesizing the AI knowledge claim with their own professional experience. This work points to the benefits of human–AI augmentation, but also that knowledge workers use their professional discretion to overrule the AI claim when they are uncertain about its output (Lebovitz et al., 2022).

Experiences

In hermeneutical phenomenology, phronesis and the capacity to act is closely intertwined with understanding, and Gadamer sets phronesis at the center of hermeneutical experience (Bobb, 2020; Gadamer, 2004). We also see it the other way around: a central element to phronetical knowledge is lived experience, involving sense perception, emotions, not-knowing, and picking up whatnesses. And here we can build upon a Husserlian phenomenology in which the stream of lived experience is foundational (Husserl, 2014). This knowledge source develops over time, and experiencing “layers and layers of concrete situations” increases the propensity for appropriate action. The attainment of phronesis then relies on experiences, and since each experience is unique, quantity matters. That is, opportunities for numerous experiences in a certain field support the development of phronesis. Experience paves the way for dealing with horizons of not-knowing by including them in the equation, rather than seeing uncertain elements as “noise” to be avoided. To leverage on individual and collective experience in organizations, however, relies on knowledge exchange. Professional experience that is relevant to a particular role or profession presents challenges in terms of upholding existing habits and behaviors which may not serve the greater good. An important part of cultivating this type of knowledge is to cultivate a culture of professional knowledge exchange to mitigate against biases and narrow perspectives taking root.

But learning from experience is not a process of generalization, it is not a mathematical exercise. Rather, earlier experiences, one’s own and those of others, function as background material in relation to which the contemporary situation can be reflected and understood in a richer way (Bornemark, 2020). Machine learning (ML), on the other hand, is capable of learning from past decisions and picking up patterns that human cognition does not see (e.g., Lebovitz et al., 2021). But as machine learning is not a lived stream of experience, it is not connected to emotions, not-knowing, valuations, and “wicked problems.” This account shows how AI and human judgment can complement each other, and how they differ.

Intuition

Within the realm of phronesis, the element of intuition denotes the unconscious knowledge process that also serves as a guiding light for professional conduct. In brief, this definition of intuition can be illustrated using the fictional example of an experienced fisherman looking for fish in new waters. The fisherman takes his boat out to sea to find a good fishing spot. While at sea he feels the wind direction and its velocity, tides, air, and water temperatures and from this information determines where to locate the boat to drop the line. The fisherman arrives at this conclusion without conscious processing; rather, the accumulated experience from prior pattern-recognition of the conditions of the sea (experience, sensory perception, emotions, and subjective knowledge) is at work here. In this way, the fisherman accumulates his experience over time, to be able to use intuition as a way to inform his decision. In other words, humans acquire and develop their ability for pattern recognition based on many different types of sensory data and previous experience

that together make up intuition. Such knowledge has also been understood as tacit knowledge: knowledge that is contextual and embodied (Polanyi, 1966).

In the same vein, machines (for example, ADS) can be described to possess the capacity for impeccable pattern recognition simply because of their capacity to process an immense amount of information. Indeed, ADS draw upon a gigantic amount of quantitative data to perform pattern-recognition (Russell, 2019). We argue that intuition is in part implicated in the intelligent machines; however, they do not display phronesis. From a phronetic perspective, while humans and machines display a propensity for pattern recognition, human intuition captures far more aspects (and properties) of social context, emotions, and meaning than the machine can. This points to the potential of ADS to increase awareness regarding “new” emerging patterns (for example, issues, risks, opportunities) on the horizon that may influence outcomes.

Episteme

Epistemic knowledge, according to Aristotle, is not part of phronetic knowledge. Rather than being focused on specific situations, epistemic knowledge abstracts and generalizes, and is focused upon that which is relevant and true to any situation. Humans continually try to formulate such knowledges in theories and in evidence-based materials, which we often term scientific knowledge. Nevertheless, today both theories and evidence-based materials need to be included in professional phronetic knowledge, for example of a doctor, a judge in court, or structural engineer in construction (Schwartz and Sharpe, 2010). Phronetic knowledge encompasses both episteme and *techné* (described next), and applying “judgment” helps to discern (and weigh up) whether to follow one episteme rather than another, or whether to follow new evidence or *techné* in a prescriptive manual. But when we sideline or forget phronetic knowledge and solely focus on epistemic knowledge, we run the risk of promoting work practices and routines that lead to poor performance and poor-quality outcomes for organizations and institutions. Episteme reflects general knowledge, and phronetic use of it entails discerning when and how to apply rules and principles that correspond with details of each situation (for example, emotions, subjective, experiential, and embodied knowledge). Overreliance on scientific knowledge and “what the computer says” (Collins, 2018) presents a danger that we lose sight of aspects of not-knowing which are inherent in complex problems and situations in professional practice. Following this, the rationality underpinning AI is based on a kind of epistemic knowledge, and this is where AI displays its strength and agency in supporting decision-making. The role of judgment, then, is to mediate between phronetic and theoretical knowledge, into a line of action that best serves the situation.

Techné

Techné represents the collection of prior experiences in relation to producing something (Bartlett and Collins, 2011). It means that to become an expert one must develop many lived experiences. A baker improves by baking many loaves; however,

the baker cannot transfer the specific details of making the “the perfect loaf.” Thus, the lessons learned through baking many loaves represent one kind of *techne* knowledge: embodied *techne* knowledge. Another kind of *techne* knowledge can be contained in recipes and manuals. Even if not every detail of making the perfect loaf can be contained in a recipe, some can. In this context, *phronetic* knowledge is at stake when, for example, an apprentice is unable to read a bread recipe correctly, and the baker subsequently responds to the situation in a manner that promotes confidence rather than shame. By understanding the role of *techne* in professional enterprises there is a possibility to broaden the view of what knowledge is. It is more than *episteme* (scientific knowledge and general across contexts) and *techne* (knowledge that can be transferred and learned), but they are both components of judgment in decision-making. While ADS have the capacity to gather recipes and manuals (for how things can be executed) to provide recommendations for action and decisions, they lack the capacity to also “feel” the situation through sensory perception to come to a judgment: they lack embodied *techne* knowledge. This scenario shows the limits of understanding judgment and reasoning as calculating competence, which is in the domain of automated decision-making.

In sum, human judgment is a kind of rationality that uses all of the above-mentioned components and moves within them (and probably also uses other components that have not become visible here). At the center of this *phronetic* judgment lies the capacity to relate to a specific situation, rather than to all situations. In different situations the different components become important. *Phronetic* judgment is in its essence connected to action, and to making decisions upon how to act in a certain situation. It becomes more important in complex, difficult, and new situations; whereas in “standard situations” it is more possible to lean toward, for example, a manual (following rules and principles) or habits. Automated decision-making can help out both in “standard situations” (even if it is a wicked problem to define these, and thus a definition that requires *phronesis*), and in complex situations where it can provide the *phronetic* judgment with data and patterns that might otherwise have been overlooked. This perspective of judgment provides us with a vocabulary to begin to better articulate and discuss the role of humans, and how human judgment can augment automated decisions and vice versa.

DISCUSSION AND CONCLUSIONS

In this chapter we explore the notion of *phronesis* in relation to automated decision-making, with the aim to show what is going on “under its hood.” We propose that *phronesis* is a synthesizing capacity which is particularly useful in complex situations where specific social and historical context matters. Given the width and depth of human judgment required for virtuous and responsible decision-making, we argue that human involvement in automated decision-making is not only essential, but should also consist of more than “rubber stamping” (Wagner, 2019). Humans should have agency in the decision-making process when the ques-

tions are of a wicked nature, or have potential impact on human beings or society, even though that would have repercussions for the efficiency and objectivity of ADS. Here we argue that AI does not have the potential to act wisely in complex situations, where human capabilities, such as *phronesis*, are required. In concrete decision-making situations, judgment enables attention to what the situation requires rather than what the algorithm dictates. Most of the time we rely on theoretical and analytical knowledge, since it appears to give us much more direction and a sense of being in control. Judgment is often not viewed as a reliable source of knowledge. The downside, then, is that we sideline what we are not able to “name” or what is often seen as subjective knowledge (for example, experiences, emotions, intuition), but that may be equally important to guide decisions. Given the emphasis on efficiency and control in organizations, there is a risk that judgment which demands more effort and energy “is under threat of AI-reckoning in decision-making” (Moser et al., 2022, p. 151). In other words, understanding judgment is more important than ever to yield greater trust in human decision-makers, but also to determine when ADS can augment decision-making, and when it may be detrimental.

While the advantages and capacities of ADS include following rules, collecting large amounts of data—that is, knowledge that lends itself to being measured and categorized—resources of *phronetic* judgment include an array of elements that are relevant to seeing the whole situation and searching for the appropriate course for action. Consequently, the relation between uniquely human knowledge (*phronesis*) and automated decision-making is still unexplored, yet has the potential to achieve social impact in both the short and the long term.

Importantly, focusing on judgment reflects the role of humans and how this aspect of decision-making is a black box. Thus, being involved in the loop of AI goes beyond auditing and altering algorithms: it also requires *phronesis*. It emphasizes the capacity when it is better to rely on evidence-based knowledge, prescriptions; when we need to rely on specifically human qualities such as not-knowing, sensory experience, emotions, intuition; and when we need to combine these two rationalities (Bornemark, 2018; Lebovitz et al., 2021). By better understanding the building blocks of human–ADS configurations, we can better understand how and why they can fail, causing suffering for the people who are affected by the decisions made by the systems, but also scepticism towards technology and automated decisions. More importantly, researching human–ADS configurations with a *phronetic* lens offers possibilities to develop applications that serve humans’ needs and values.

Scholars in the field of ADS reinforce the urgency to develop people-focused approaches in the design, use, and implementation of automated systems and technologies. On a deeper level it is a call for the need to rehumanise automation, and hence emphasize human agency in human–machine interaction; in current research termed human-in-the-loop (e.g., Grønsund and Aanestad, 2020). Against this background, this chapter sheds light on the role of human judgment in relation to AI represented in automated decision-making. The aim of the chapter is to spur discussions to inform and empower understanding of what judgment is, and its role in ADS and other human–AI interaction contexts. We should ask ourselves how we can make

space for “judgment” in contexts where following rules and manuals are prioritized over the need to base decisions on the specifics of the situation.

In this sense, and with regard to this *Handbook*’s topic of AI and decision-making, using the conceptual apparatus derived here can assist researchers and organizations to inquire (and intervene) into data-driven decision-making contexts, to realize the potential of humans to make wiser decisions. Without understanding how people incorporate information from algorithms with human judgment during decision-making processes, organizations run the risk of deploying AI in a way that reduces its true potential to support decisions in organizations.

Finally, there is a pressing need for more awareness of the “human black box,” namely the illusion that we understand humans better than algorithmic decision-making (Bonezzi et al., 2022), and its societal consequences. In other words, much attention is placed on explaining the algorithm, but the human remains unexplainable. Understanding human judgment, and how humans draw on the resources of phronesis, can contribute to greater trust in humans and human–AI decisions. Looking into the future, an improved understanding of judgment will be critical to cultivate human–AI-based decision-making that is informed by human judgment rather than algorithms.

REFERENCES

- Araujo, T., Helberger, N., and Kruijemeier, S. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI and Society*, 35, 611–623. <https://doi.org/10.1007/s00146-019-00931-w>.
- Aristotle (2009). *The Nicomachean ethics*. Oxford World’s Classics. Translated by David Ross, Revised with an introduction and notes by Lesley Brown. Oxford: Oxford University Press.
- Bartlett, R.C., and Collins, S.D. (2011). *Aristotle’s Nicomachean ethics*. A new translation. Chicago, IL: University of Chicago Press.
- Baggini, J., and Fosl, P. (2020). *The philosopher’s toolkit: A compendium of philosophical concepts and methods*. Hoboken, NY: Wiley-Blackwell.
- Baskerville, R.L., Myers, M.D., and Yoo, Y. (2020). Digital first: The ontological reversal and new challenges for information systems research. *MIS Quarterly*, 44(2), 509–523.
- Beauchamp, T.L. (1991). *Philosophical ethics: An introduction to moral philosophy*. New York: McGraw-Hill.
- Begley, P.T. (2006). Self-knowledge, capacity, and sensitivity: Prerequisites to authentic leadership by school principals. *Journal of Educational Administration*, 44(6), 570–589. <http://dx.doi.org/10.1108/09578230610704792>.
- Berente, N., Gu, B., Recker, J., and Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, 45(3), 1433–1450. <https://misq.org/skin/frontend/default/misq/pdf/CurrentCalls/ManagingAI.pdf>.
- Bobb, C.V. (2020). The place of phronesis in philosophical hermeneutics. A brief overview and a critical question. *Hermeneia*, 25, 29–36.
- Bonezzi, A., Ostinelli, M., and Melzner, J. (2022). The human black-box: The illusion of understanding human better than algorithmic decision-making. *Journal of Experimental Psychology General*, 1–9. <https://doi.org/10.1037/xge0001181>.

- Bornemark, J. (2018). The limits of Ratio: An analysis of NPM in Sweden using Nicholas of Cusa's understanding of reason. In Btuhaj, A. (ed), *Metric culture: Ontologies of self-tracking practices* (pp. 235–254). Bingley: Emerald Publishing.
- Bornemark, J. (2020). *Horisonten finns alltid kvar: Om det bortglömda omdömet*. Stockholm: Volante.
- Boström, N. (2017). *Superintelligence: Paths, dangers, strategies*, 2nd edn. Oxford: Oxford University Press.
- Collins, H. (2018.) *Artificial intelligence: Against humanity's surrender to computers*. Cambridge: Polity Press.
- Constantinescu, M., and Crisp, R. (2022). Can robotic AI systems be virtuous and why does this matter? *International Journal of Social Robotics*, 1–11. <https://doi-org.ezproxy.ub.gu.se/10.1007/s12369-022-00887-w>.
- Descartes, R. (1989). *The passions of the soul: An English translation of Les Passions De l'Ame*, Indianapolis, IN: Hackett Publishing Company.
- Dias, R., and Torkamani, A. (2019). Artificial intelligence in clinical and genomic diagnostics. *Genome Medicine*, 11(1), 1–12. <https://doi.org/10.1186/s13073-019-0689-8>.
- Dobbe, R., Dean, S., Gilbert, T., and Kohli, N. (2018). A broader view on bias in automated decision-making: Reflecting on epistemology and dynamics. *ArXiv.org*, ArXiv.org.
- Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. New York: St Martin's Press.
- Fountainaine, T., McCarthy, B., and Saleh, T. (2019). Building the AI-powered organization. *Harvard Business Review*, 97(4), 62–73.
- Fügener, A., Grahl, J., Gupta, A., and Ketter, W. (2021). Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *MIS Quarterly*, 45(3), 1527–1556. DOI:10.25300/MISQ/2021/16553
- Gadamer, H.G. (2004). *Truth and method*. London: Continuum.
- Gal, U., Jensen, T.B., and Stein, M.K. (2020). Breaking the vicious cycle of algorithmic management: A virtue ethics approach to people analytics. *Information and Organization*, 30(2), 100–301. <https://doi.org/10.1016/j.infoandorg.2020.100301>.
- Grønsund, T., and Aanestad, M. (2020). Augmenting the algorithm: Emerging human-in-the-loop work configurations. *Journal of Strategic Information Systems*, 29, 101–614. <https://doi.org/10.1016/j.jsis.2020.101614>.
- Hussain, K. (2018). Artificial intelligence and its applications goal. *International Research Journal of Engineering and Technology*, 5(1), 838–841.
- Husserl, E. (2014). *Ideas for a pure phenomenology and phenomenological philosophy. First book, General introduction to pure phenomenology*. Indianapolis, IN: Hackett Publishing Company.
- Hylving, L., and Koutsikouri, D. (2020). Exploring phronesis in digital innovation, In: *Proceedings of the 28th European Conference on Information Systems (ECIS), An Online AIS Conference, June 15–17, 2020*. https://aisel.aisnet.org/ecis2020_rp/78
- Hylving, L., and Lindberg, S. (2022). Ethical dilemmas and big data: The case of the Swedish Transport Administration. *International Journal of Knowledge Management (IJKM)*, 18(1), 1–16. DOI: 10.4018/IJKM.290021
- Hylving, L., Koutsikouri, D., Bornemark, J., and Lindberg, S. (2022). Ratio and intellectus: Towards a conceptual framework for understanding human and artificial intelligence. *ICIS2022*. https://aisel.aisnet.org/icis2022/adv_methods/adv_methods/3.
- Irwin, T. (1999). *Aristotle: Nichomacean ethics*, 2nd edn. Indianapolis, IN: Hackett Publishing.
- Jeste, D.V., Lee, E.E., Palmer, B.W., and Treichler, E.B.H. (2020). Moving from humanities to sciences: A new model of wisdom fortified by sciences of neurobiology, medicine, and evolution. *Psychological Inquiry*, 31(2), 134–143. <https://doi-org.ezproxy.ub.gu.se/10.1080/1047840X.2020.1757984>.

- Kase, K., González-Cantón, C., and Nonaka, I. (2014). *Phronesis and quiddity in management. A school of new knowledge approach*. London: Palgrave Macmillan.
- Kolbjørnsrud, V., Amico, R., and Thomas, R.J. (2016). How artificial intelligence will redefine management. *Harvard Business Review*, 2(1), 3–10.
- Kristjánsson, K., Fowers, B., Darnell, C., and Pollard, D. (2021). Phronesis (practical wisdom) as a type of contextual integrative thinking. *Review of General Psychology*, 25(3), 239–257.
- Lebovitz, S., Levina, N., and Lifshitz-Assaf, H. (2021). Is AI ground truth really “true”? The dangers of training and evaluating AI tools based on experts’ know-what. *MIS Quarterly*, 45(3), 1501–1525. DOI: 10.25300/MISQ/2021/16564.
- Lebovitz, S., Lifshitz-Assaf, H., and Levina, N. (2022). To engage or not with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization Science*, 33(1), 126–146. <https://doi-org.ezproxy.ub.gu.se/10.1287/orsc.2021.1549>.
- Lepri, B., Oliver, N., and Pentland, A. (2021). Ethical machines: The human-centric use of artificial intelligence. *IScience*, 24(3), 102249. DOI: <https://doi.org/10.1016/j.isci.2021.102249>.
- Lindberg, S. (2022). Ethical dilemmas and big data: The case of the Swedish Transport Administration. *International Journal of Knowledge Management (IJKM)*, 18(1), 1–16. DOI: 10.4018/IJKM.290021
- Lu, D. (2019). AI can predict if you’ll die soon – but we’ve no idea how it works. *New Scientist*. <https://www.newscientist.com/article/2222907-ai-can-predict-if-youll-die-soon-but-weve-no-idea-how-it-works/#>.
- Marshall, D.L. (2010). The origin and character of Hannah Arendt’s Theory of Judgment, *Political Theory*, 38(3), 367–393. <https://www.jstor.org/stable/25704821>.
- Mayer-Schönberger, V., and Cuckier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Boston, MA and New York, USA: Houghton Mifflin Harcourt.
- Merleau-Ponty, M. (1974). *Phenomenology of perception*. London, UK and New York, USA: Routledge & Kegan Paul / Humanities Press.
- Moser, C., den Hond, F., and Lindebaum, D. (2022). Morality in the age of artificially intelligent algorithms. *Academy of Management Learning and Education*, 21(1), 139–155.
- Mökander, J. and Axente M. (2021). Ethics-based auditing of automated decision-making systems: Intervention points and policy implications. *AI and Society*, 38, 153–171.
- Neuroth, M., MacConnell, P., Stronach, F., and Vamplew, P. (2000). Improved modelling and control of oil and gas transport operations using artificial intelligence, *Knowledge-Based Systems*, 13(2–3), 81–92.
- Nussbaum, M.C. (2001). *Upheavals of thought: The intelligence of emotions*. Cambridge: Cambridge University Press.
- O’Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Crown.
- Polansky, R.M. (2000). “Phronesis” on tour: Cultural adaptability of Aristotelian ethical notions. *Kennedy Institute of Ethics Journal*, 10(4), 323–336.
- Polanyi, M. (1966). *The tacit dimension*, 1st edn. London: Routledge & Kegan Paul.
- Ranerup, A., and Henriksen, H.Z. (2019). Value positions viewed through the lens of automated decision-making: The case of social services. *Government Information Quarterly*, 36(4), 101377. <https://doi.org/10.1016/j.giq.2019.05.004>
- Ratti, E. (2019). Phronesis and automated science: The case of machine learning and biology. In F. Sterpetti and M. Bertolaso (eds), *Will science remain human?* Springer. https://doi.org/10.1007/978-3-030-25001-0_8.
- Rittel, H.W.J., and Webber, M.M. (1973). Dilemmas in a general theory of planning. *Policy Sci*, 4, 155–169. <https://www.jstor.org/stable/4531523>.
- Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. New York: Penguin.

- Schwartz, B., and Sharpe, K. (2010). *Practical wisdom: The right way to do the right thing*. New York: Riverhead Books.
- Shotter, J., and Tsoukas, H. (2014). Performing phronesis: On the way to engaged judgment. *Management Learning*, 45(4), 377–396. <https://doi-org.ezproxy.ub.gu.se/10.1177/13505076145411>.
- Slater, D. (2013). *Love in the time of algorithms: What technology does to meeting and mating*. London: Penguin.
- Smith, B.C. (2019). *The promise of artificial intelligence: Reckoning and judgment*. Cambridge, MA: MIT Press.
- Souto, P.C.N. (2019). Ontological not-knowing to contribute attaining practical wisdom: Insights from a not-knowing experience in ‘samba-de-gafieira’ dance to the value of being and responding from within our practical experience and practical knowledge. *Learning, Culture and Social Interaction*, 21, 48–69. <https://doi.org/10.1016/j.lcsi.2019.01.008>.
- Sumter, S.R., Vandenbosch, L., and Ligtenberg, L. (2017). Love me Tinder: Untangling emerging adults’ motivations for using the dating application Tinder. *Telematics and Informatics*, 34(1), 67–78. <https://www.jstor.org/stable/26853723>.
- Teodorescu, M., Morse, L., Awwad, Y., and Kane, G.C. (2021). Failures of fairness in automation require a deeper understanding of human–ML augmentation. *MIS Quarterly* 45(3), 1483–1499. DOI:10.25300/MISQ/2021/16535.
- Wagner, B. (2019). Liable, but not in control? Ensuring meaningful human agency in automated decision-making systems. *Policy and Internet*, 11(1), 104–122. <https://doi-org.ezproxy.ub.gu.se/10.1002/poi3.198>.
- Wiederhold, B. (2019). Can artificial intelligence predict the end of life ... and do we really want to know? *Cyberpsychology, Behaviour, and Social Networking*, 22(5), 297–299. <https://doi-org.ezproxy.ub.gu.se/10.1089/cyber.2019.29149.bkw>.
- Yar, Majid (2000). From actor to spectator Hannah Arendt’s “two theories” of political judgment. *Philosophy and Social Criticism*, 26, 1–27.
- Zheng, L., Niu, J., Zhong, L., and Gyasi, J.F. (2021). The effectiveness of artificial intelligence on learning achievement and learning perception: A meta-analysis. *Interactive Learning Environments*, 1–15. <https://doiorg.ezproxy.ub.gu.se/10.1080/10494820.2021.2015693>.

9. Making decisions with AI in complex intelligent systems

Bijona Troqe, Gunnar Holmberg, and Nicolette Lakemond

Artificial intelligence (AI) is rapidly changing the landscape of society. Organizations and entire industries, including ones that deal with complex product and systems (CoPS) such as aviation, healthcare, and more, are using AI as a strategic and integral part of their products, and investing into reaping the benefits of AI in new types of business models as well as changes to daily operations (Dwivedi et al., 2021). Specifically, the integration of AI solutions in organizational decision-making has received astounding attention, from both the industry and scientific research (Shrestha et al., 2019; Trabizi et al., 2019; Puranam, 2021). Organizations across industries are implementing machine learning data-driven solutions for processing large amounts of data and supporting humans in daily decisions. Healthcare is a prominent example of an industry that has realized the potential of integrating AI in decision-making processes, with examples ranging from imaging and diagnostics (Dembrower et al., 2020), to diabetes management and drug discovery (Bohr and Memarzadeh, 2020).

These advancements potentially provide many societal benefits, such as better-informed decisions and better healthcare for society, but are also associated with several potential challenges. These involve a continuous transformation following evolving intelligent technologies, complex approval and certification processes and acceptance. The transformation will likely be inherently volatile at certain stages and lead to a more ambiguous decision-making environment with a more open collaborative landscape with multiple stakeholders. The integration of AI solutions in increasingly complex systems—that is, complex intelligent systems (CoIS) (Lakemond et al., 2021)—creates new challenges, not least in relation to decision-making. Questions such as who is the decision maker, what the decision processes look like, and what happens with the decision space when AI is integrated, cannot fully be answered with traditional organizational structures and strategies (Puranam, 2021). Therefore, as AI technologies become increasingly advanced, there is a growing need for researchers and practitioners to understand the new prerequisites of decision-making in CoIS and the possible organizational and societal implications of these technologies.

To increase the likelihood that the potential benefits of AI in complex systems can be achieved, the purpose of this chapter is to explore new implications for organizational decision-making with AI, where decision-making becomes reliant on a degree of human–AI collaboration. To address this, we draw on the illustrative example

of decision-making processes in the emerging field of personalized medicine. In particular, we focus on three main areas of decision-making, which we draw from the extensive research in decision-making theories and organizational studies: the decision-maker, the decision process (Langley et al., 1995), and inspired by a design perspective, the decision space (Hatchuel, 2001). By analyzing the dynamics that can be found in these three views on decision-making in relation to AI, we outline three main prerequisites of decision-making that offer new and potentially critical insights in how decision-making is shaped in CoIS. This new lens will broaden the established managerial decision-making perspective by accommodating characteristics of AI and complex systems.

The chapter contributes with a framework for identifying and explicating the new dynamic properties of decision-making in complex intelligent systems, based on a first exploration of decision-making processes in the case of personalized medicine within the healthcare sector. This contribution will be a first step towards a human–AI collaborative decision-making perspective that expands beyond the frames of organizational decision-making and resonates with new opportunities in a generativity paradigm and multi-faceted decision-making outcomes. By taking the perspective of CoIS into account, we aim at contributing towards the opportunities that AI and autonomous technologies present for better decision-making that is addressing societal benefits and enabling a responsible use of AI. Bringing together research in AI with management and organization is crucial for researching decision-making implications in the setting of CoIS.

DECISION-MAKING

Organizational decision-making has been a central concept in studies of management, organizations, and innovation for a long time. Literature in organizational theory has been preoccupied in understanding decision-making in organizations, whether it is the role of the individual, group, or organization (March, 1978; Simon, 2013; Kahneman, 2003), or centering around operational or strategic decision-making frameworks (Cyert and March, 1963; Cohen et al., 1972; Mintzberg, 1978). A fine example of structuring an understanding of decision-making is represented in the work of Langley et al. (1995) in the article “Opening up decision-making: the view from the black stool.” They identify main components of organizational decision-making, including perspectives from the decision-maker and the decision process, and offer a problematization of the relationships and dynamic linkages of decision-making as interwoven networks of issues. To frame the future of decision-making in CoIS, we use the decision-maker and the decision-making process as two perspectives (Langley et al., 1995), and add the decision space (Hatchuel, 2001) as a third perspective, providing complementary insights related to the decision in its context. The three views form the basis for our framework for understanding decision-making in CoIS. These are discussed in the next sections.

The Decision-Maker

A focus on the decision-maker has emphasized the role of humans in decision-making, including their individual characteristics and limitations that determine organizational outcomes (Langley et al., 1995). The theory of bounded rationality, grounded by Herbert Simon in the late 1940s, can be considered a landmark in understanding how humans make decisions, as the sequential process of listing alternative sets based on predefined problems and then choosing the alternative with the highest satisfactory value (Simon, 2013). He basically proposed that humans are incapable of making perfect rational decisions because of cognitive and time limitations. Hence, Simon (2013) proposed as an alternative view that humans make satisfactory choices, based on satisficing, and use “rules of thumb” as simple heuristics to guide these decisions.

This model of bounded rationality gave way to an increasing attention for research on cognitive processes and their effects on decision-making under uncertainty. Kahneman and Tversky devoted a large effort and experiments towards probability calculation and expansion of the utility theory by highlighting the role of framing and heuristics as central in decision-making (Tversky and Kahneman, 1974; Kahneman, 2003). This uncovered new challenges with the bounded rationality model, placing more weight on intuition, emotion, and other “non-rational” factors in cognitive decision-making, extending organizational decision-making under contexts of uncertainty.

As an example, in the medical context, the decision-maker role is filled by physicians, nurses, and other healthcare professionals. According to Li and Chapman (2020) medical decision-making is particularly difficult because it involves a lot of ambiguity, high risks, information overload, as well as a high number of stakeholders. Physicians and other medical decision-makers are often subject to cognitive overload, which comes as a result of processing large amounts of information, that is, risk probabilities, treatment efficacy, time and cost considerations, and more. A large part of such cognitive overload comes from numeracy factors (Li and Chapman, 2020), but also ambiguity, stress, and emotional overload are factors affecting the decision-makers’ ability to make decisions and pushing them towards satisficing. Another factor that contributes to decision-making in the medical context is the high risk–high uncertainty premises they are faced with, which makes them “systematically over- or underweight” decision outcomes (Li and Chapman, 2020), which again is a characteristic of the human decision-maker according to literature (Kahneman, 2003).

The Decision-Making Process

As a second view, decisions can be considered as “a system of decisional processes” (Langley et al., 1995). Decisions are not actions that happen at one point in time (*ibid.*), but are rather collective and iterative processes between actors in the organization, and interrelated with problems and goals, learning and organiza-

tional structures. In order to economize on bounded rationality, organizations build routines and standardized processes, which in turn help to define decision-making structures and processes (Gavetti et al., 2007). Simon (2013) views the organization as an entity where decision-making happens collectively as a result of hierarchy. The higher-level actors of the organization define the decision-making premises that then shape the decisions of the lower-level actors. This happens as a result of vertical authority and communication structures (ibid.). Langley et al. (1995) propose that decision-making processes can more resemble a network of decisions, driven both by rational processes and by emerging events. Such a perspective highlights the complexity of decision-making processes. Decisions do not flow from one point of the organization to another. For instance, when specialized information processing is needed from different organizational departments or functions, Simon (2013, p. 209) suggests that:

a communication process must be set up for transmitting these components from the separate centres to some point where they can be combined and transmitted in turn, to those members in the organisation who will have to carry them.

The bi-directional information processing flow shows the importance of organizational and decision-making structures and processes for knowledge integration. Additionally, this highlights the importance of a multitude of stakeholders in decision-making in organizations, as well as the possibility that these stakeholders enact and interpret information and decisions depending on the context of the decision-making.

The Decision Space

As a complement to the decision-maker and the decision-making process, we contend that the decision space is an additional perspective. The decision space can be understood as the space where alternative solutions for problems are identified and generated. The decision space is highly intertwined with the characteristics of the context. For instance, Simon's satisficing model of decision-making is a rule-based method, like chess, where complexity is often represented by a finite number of alternative positions (Hatchuel, 2001). Satisficing in this case works because it uses heuristic search on past experience patterns to find the solution (ibid.). However, this method does not work well where patterns of past experiences are not present, when the unknown goes beyond statistical uncertainty (ibid.). Making decisions in the unknown means discovering new alternatives and new states of the world. In this line, the decision space is connected to a theory of generativity of decision solutions in the (partly) unknown, based on design theory principles for innovation management (Le Masson et al., 2019). Such a perspective can be important as unknown states of the world may push decision-makers to fundamentally rethink their initial choices for that isolated event, and thus radically change decision-making. Thus, the goal of a generative decision-making framework, by focusing on the decision space,

is to identify unknowns that are challenging the process of decisions and to find ways to address this.

In the medical context, the decision space is tightly linked to the information and knowledge available to the physicians and other medical decision-makers. Connecting to the previously mentioned fact that medical professionals are constantly faced with information overload is an example of the fact that more information and alternative options do not equal better decision-making. In fact, Li and Chapman (2020) discuss how the opposite is true, and that more information can lead to cognitive overload, and therefore lower quality of decision-making. Sometimes, contracting the decision space by offering simplified choices can lead to better choice alternatives for decision-making (ibid.). Another important consideration for the decision space in the medical context is related to the generation and choice of future outcomes. Because of time and other uncertainty risk factors, medical decision-makers are restricted in the generativity of decision space, and might resort to making satisfactory decisions which may not be in line with future outcomes. Instead, immediate alternatives might be preferred not because of better quality, but because of context-imposed trade-offs.

The three perspectives form the central areas in our framework for understanding decision-making in CoIS. They represent different views that highlight different aspects of decision-making and together can create a more complete understanding. This is visualized in Figure 9.1. The view of the decision-maker is traditionally connected to the individual characteristics of human beings and their capability to make satisficing (bounded) rational decisions. The decision-making process is connected to sequential as well as anarchical aspects, not least when a network of decisions and stakeholders is involved. The decision space addresses the unknown and evolving generativity in relation to temporality and the context in which decisions are made. These three areas can be used to understand how the prerequisites of decision-making may change when AI becomes an integrated part in the decision-making process.

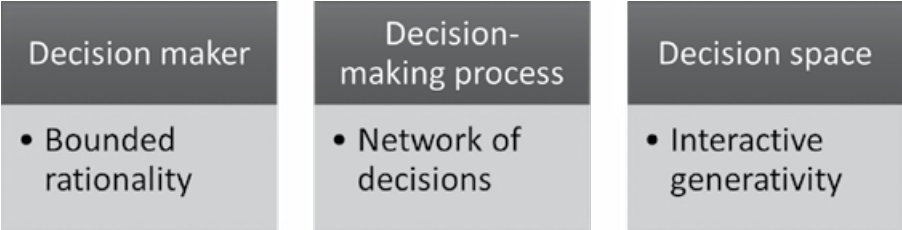


Figure 9.1 Central aspects of decision-making in organizations

AI

Artificial intelligence can be defined as non-human agents that display human-like cognitive functions, such as problem-solving and learning, and perform specific tasks with some degree of autonomy (Russel and Norvig, 2016; Dwivedi et al., 2021). Due

to the recent advances in AI development as well as increased availability of data and data-sharing platforms, more and more organizations are using AI tools and solutions as part of managerial tasks (Raisch and Krakowski, 2021). According to Dignum (2019), for an artificial agent to be considered intelligent, it needs to possess (among other properties) proactiveness, otherwise known as the ability to display autonomy. AI solutions can be autonomous in the sense that they can act and adapt without the interference of humans, and in terms of managerial processes, this can result in AI solutions taking over tasks previously performed by humans (Raisch and Krakowski, 2021). In the perspective of AI-based decision-making in organizations, Shrestha et al. (2019) also acknowledge the potential of AI taking over certain decision-making tasks, calling this structure of decision-making “full human to AI delegation.” According to the authors, this type of AI-based decision-making works when the problem and the decision search space are well defined. Here, the strengths of AI lie not only in the ability to process and analyze vast amount of data in a short time, overcoming what Simon previously noted as some of the most prominent factors of bounded rationality, but also in using machine learning algorithms to adapt and adjust decision accuracy and performance features (*ibid.*).

While autonomy is a property of AI that is being exploited in management, another property of AI can be argued to be augmentation, which according to Raisch and Krakowski (2021) can be defined as the ability of humans and AI to work closely together on a task. Shrestha et al. (2019) consider augmentation in decision-making as a hybrid decision-making structure, either as AI to human sequential decision-making, or as human to AI sequential decision-making. The former allows organizations to use the strengths of both AI and humans, usually in a sequential manner, where humans make the final, authoritative decision based on the input of AI. The latter follows the opposite sequence, where humans provide the initial input, usually on a small sample size, and AI is used to assist decision-making, usually with predictive modelling (*ibid.*).

The previously mentioned AI characteristics and possible decision-making structures are just two of the examples of how AI can be integrated in an organizational decision-making process; other types of configurations are also possible (Shrestha et al., 2019; Raisch and Krakowski, 2021). With the development of algorithms and extensive availability of data in numerous industries, organizations have not only increased the range of applications of AI in management, such as robotics vehicles, speech recognition, and more (Russel and Norvig, 2016), but also integrated dynamically different solutions in different decision-making structures.

AI IN THE CONTEXT OF COMPLEX SYSTEMS

As we are considering the context of emerging CoIS, it is necessary to position the understanding of AI in such a context in relation to the characteristics of complex systems. Hobday (1998) has characterized complex systems (denoted as complex products and systems, CoPS) as:

high-technology, business-to-business capital goods used to produce goods and services for consumers and producers. Unlike high volume consumption goods, each individual CoPS is high cost and made up of many interconnected, often customized parts, including control units, sub-systems, and components, designed in a hierarchical manner and tailor-made for specific customers. (Hobday, 1998, p. 689)

Differently from mass-produced products, CoPS are produced in small batches, or even as one-off products, which increases the focus on the systems engineering, project management, and system integration. Today, CoPS are largely present in several industries, including critical infrastructure and information and communication technology (ICT) to name a few (Lakemond et al., 2021). Complexity as a defining characteristic of CoPS is a result of the high degree of customization, product architecture, and new knowledge integration in the systems (Lakemond et al., 2021). CoPS may, to some extent, exhibit emergent behaviors throughout a long development as well as life cycle, with users, customers, system integrators, suppliers, and system engineering teams frequently interacting (Davies and Hobday, 2005). Besides the large degree of components, knowledge and skills, system integration in CoPS involves a multitude of actors which are brought together to produce the products or systems (Lakemond et al., 2021).

As a consequence of an increasing embeddedness of software and technology into CoPS, the complexity of managing such systems increases (Lakemond et al., 2021). With the integration of AI, amongst other things, a higher degree of intelligence and autonomy is expected to define the system, which in turn transforms these systems from CoPS to CoIS (complex intelligent systems). This autonomy can potentially disrupt already established system processes, configurations, and roles. Human authority in the system management and decision-making may be challenged by the increased autonomy stemming from AI, which creates further challenges, but also possibilities for the system management. Furthermore, new actors and new organizations—that is, data platforms, cloud systems, AI manufacturers, and vendors—can be expected to become part of the wider organizational context, as a results of AI embeddedness, contributing to further necessity to create an understanding of the capabilities that are imperative to master the management of CoIS. This extends to decision-making in and for CoIS, as AI not only potentially alters the authority of the decision-makers, but also has implications for organizational goals and decision spaces.

PERSONALIZED MEDICINE: AN OVERVIEW

Personalized medicine, also referred to as “precision medicine,” is an approach to medicine that focuses on the development of targeted therapies and medical treatment for individuals (Wang, 2022). While this field is not entirely new, having initiated with early experimentations in the 1960s (Duffy, 2016), it was not until the last decade that it really started to become the focus of policy makers and national strat-

egies for healthcare. In 2015, the United States launched the Personalized Medicine Initiative (PMI) (ibid.), and in Europe the International Consortium for Personalized Medicine (ICPerMed)¹ joins the effort of 40 European countries in research and advancement in the field of personalized medicine.

Personalized medicine has been considered a paradigmatic shift of medical philosophy, where instead of focusing on reactive and symptomatic treatments which include categorizing patients into “disease profiles,” the new approach is focusing on proactive prevention and treatment based on individual characteristics and personalized approaches (Duffy, 2016; Mesko, 2017). In addition to the entirely new philosophy of preventative care, personalized medicine has seen a dramatic rise due to the development of AI, genomics, robotics, and other similar types of disruptive technologies. With the refinement of AI algorithms, diagnostic tools, and cloud computation systems, physicians and healthcare practitioners can create a better understanding of the status of their patients and make better-informed decisions for their patients. In addition, due to the increasing implementation of electronic health record (EHR) and availability of individual patient data, such as medical history, genomics, lifestyle, and environmental factors, AI algorithms can assist healthcare practitioners to make decisions, and prescribe personalized and targeted treatments (Duffy, 2016; Kriegova et al., 2021).

COIS CHARACTERISTICS IN PERSONALIZED MEDICINE

The important endeavor of focusing on long-term health and preventative strategies (Kriegova et al., 2021) relies on continuous and complex integration of multiple actors and components, medical technologies, and high context uncertainty (see Table 9.1). Considering personalized medicine as a complex system (Rutter et al., 2017), its main components are characterized as the following: large electronic health record or cloud data base for each individual patient medical history and other relevant data; computational methods that can analyze and interpret this large amount of data; and predictive modelling and other approaches (Duffy, 2016) that support decision-making (Kriegova et al., 2021) for potential treatments and personalized therapy (as pictured in Figure 9.2). The system incorporates recurrent follow-ups and monitoring, and it involves both basic and applied research (Duffy, 2016; Pritchard et al., 2017).

Considering the previously mentioned components and their individual and systemic coupling, together with the several feedback loops and operational interactions, the system complexity in personalized medicine can be considerably high. Adding to this, the fundamental goal of providing tailor-made treatments creates further complexity for the decision-making of both medical professionals and patients (Wang, 2022). Finally, personalized medicine as a complex system is a high-uncertainty and high-risk context (Kriegova et al., 2021). The traditional evidence-based medicine is a high-uncertainty and high-risk context, and with diagnostics and treatments being personalized, more factors are considered and healthcare is more tailored to each

Table 9.1 System characteristics of personalized medicine

System characteristics	Personalized medicine
Multiple actors and disciplines	Integrated healthcare systems
High-complexity context	Personalized treatment
	Connecting multiple individual factors
Technological integration	AI, data-driven methods, robotics integrated to the existing healthcare practices

Source: Adapted from Duffy (2016).

patient. However, this also poses some challenges for decision-making, with high integration of technologies, actors involved, and patient information.

AI IN PERSONALIZED MEDICINE

The application of AI has made advances in the field of personalized medicine possible, and an increasing range of AI-supported tools is emerging in diagnosis and screening, patient health monitoring, and prediction. Examples of AI tools in diagnostics include cancer screening, diabetic retinopathy diagnosis, pathology, and more (Liu et al., 2020). Most of these tools are based on machine learning and deep learning models, which are able to analyze and learn from vast amounts of data. Availability of data, including electronic health records (EHR), patient medical history, lifestyle and environmental data (often available from health monitoring devices such as watches or phones) are proving substantial for qualitative clinical decision support systems, and for reducing patient risks and health challenges (Kriegova et al., 2021). AI tools are also being used for collecting, storing, and mining medical data, which can help physicians by saving considerable amounts of time in analyzing and structuring such data (Mesko, 2017).

However, the introduction of new AI-based tools in traditional medical practice also alters the role of physicians as the central decision-making actors (Li and Chapman, 2020). Traditionally, physicians have had very limited tools, resources, and computational skills, and have often relied on their experience, own judgment, and intuition to solve problems. Evidence-based guidelines and policies have been supporting decision-making. Physicians are often faced with a lack of time, a high amount of risk, and a need to make trade-offs. Furthermore, there may sometimes be a lack of information, but also sometimes an overwhelming amount of information available that creates difficulties for physicians and other decision-makers in the field.

It has been proposed that the increasing application of AI in several medical areas of personalized medicine, as well as the increased involvement of the patient, changes the focus on decision-making from the individual physician as the main decision-maker towards decision-making as a collaborative effort (Lu et al., 2023). The collaborative effort may involve a variety of actors. For instance, it involves col-

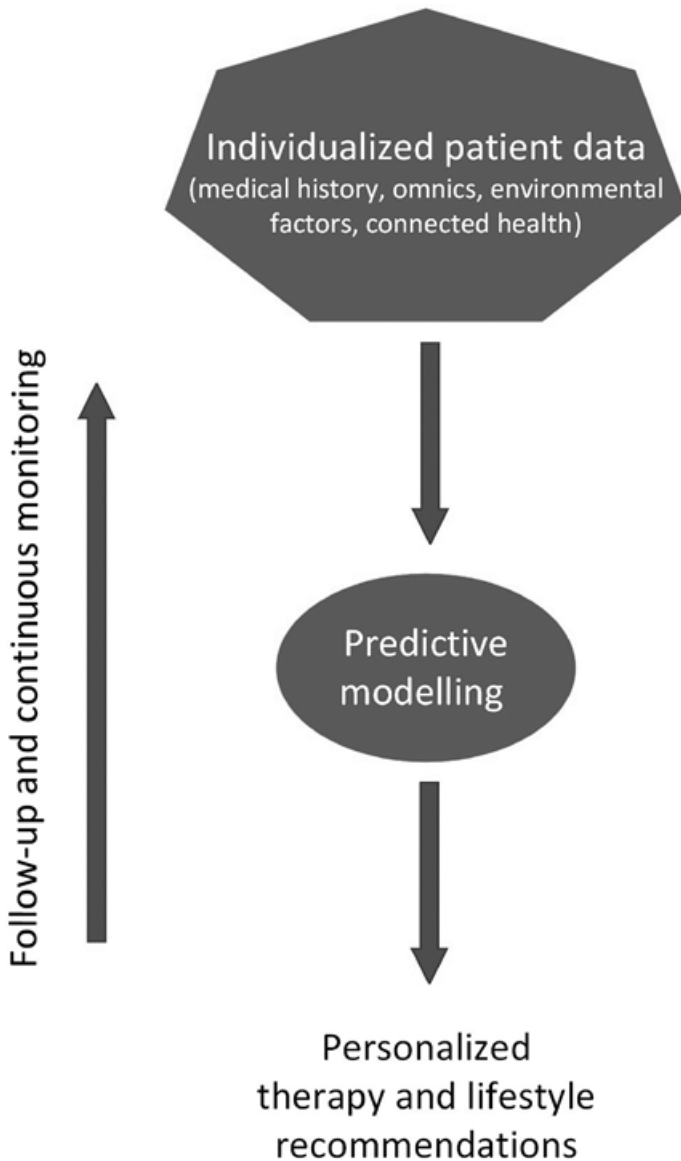


Figure 9.2 *Personalized medicine overview*

laborative engagement between industry and the health system (ibid.), but also physicians and patients that together evaluate medical factors and make better-informed health-related decisions (Kriegova et al., 2021). In personalized medicine, the role of the patient shifts from a passive receiver of care to an active contributor to their own health. To be able to provide targeted treatment, physicians and doctors interact with

patients, as well as patients increasingly engaging with several technological health tools, such as health trackers, to extend their knowledge around health issues.

Some AI systems could perhaps be considered increasingly as actors as well: that is, machine learning methods analyzing MRI images can collaborate with radiologists as independent readers, and thus become (semi)autonomous decision-makers in certain stages of breast cancer screening (Dembrower et al., 2020). In medical areas other than radiology, physicians and hospital staff are increasingly relying on AI as decision-supporting tools for providing individualized clinical solutions to patients, and even for operational decisions on an organizational level. Other AI solutions such as deep-neural networks are used for human genome interpretation and identification of complex diseases such as cancer on the DNA level (ibid.). Such an innovative approach allows physicians to have access to much better resources, such as deep learning and other learning and reasoning systems, to interpret the vast amount of individual patient data that can aid decision-making for diagnosis and treatments (Duffy, 2016; Mesko, 2017; Bohr and Memarzadeh, 2020).

In some parts of the world, these developments can make a great difference. For instance, in India, automatic image classification systems are now used to screen millions of people for diabetic retinopathy, a disease that causes blindness for more than 90 million people in the world (Yu et al., 2018).

NEW PREREQUISITES FOR DECISION-MAKING IN COIS

As described in the previous section, the context of personalized medicine clearly represents a changed landscape for healthcare, which could be described as a transformation from CoPS to increasingly CoIS. This implies a new logic, additional actors, increasing amounts of information available, and thus from a decision-making perspective a more complex decision situation. With these developments, the prerequisites for decision-making are changing, and a new understanding of decision-making needs to be reached to be able to reap the potential benefits that are associated with the shift towards personalized medicine in particular, and perhaps also the emergence of CoIS in other contexts more generally. Collaborative decision-making, change in organizational structures, and being context-driven are three important prerequisites which we highlight based on our empirical observations. Table 9.2 describes these new prerequisites connected to the decision-maker, the decision-making process, and the decision space. These are further discussed below.

Decision-Maker: Collaborative Decision-Making

As observed in the case of personalized medicine, in the context of CoIS, the presence of AI in decision-making is increasing. Different combinations of human and AI decision-making structures can be observed, including forms of delegation and sequential decision-making (Shrestha et al., 2019), as well as automation and aggregation of decision-making and organizational tasks (Raisch and Krakowski, 2021).

Table 9.2 New decision-making prerequisites for CoIS

Decision-making	New prerequisites for CoIS	Characteristics
Decision maker	Collaborative decision-making with AI	Human–AI collaboration Different combinations possible
Decision-making process	Change of organizational processes and structures	Changing roles Change of organizational processes
Decision space	Driven by an increasing number of elements in the context	Data-driven knowledge Generativity Requisite variety

The type of human–AI combination in decision-making varies with the type of decision and goal the organization is aiming for, as well as the type and function of the AI solutions. On the one hand, if the goal is to save time and critical resources, such as radiologist time, an automated solution (such as a machine learning solution used to scan MRI images) can be extremely helpful in saving an overloaded radiologist the job of sifting through a large number of images. On the other hand, such automation can make the AI solution partly a “co”-decision maker, with the AI solution receiving different degrees of authority. For instance, it can only serve as an input for the human decision-maker, or function as a second independent decision-maker. This reflects different degrees of automation and augmentation, providing different possibilities to create synergies and effective decision-making. Whatever the combination, using the strengths of the humans and AI is key to finding solutions to problems coming from the complexity and uncertainties of the context. In a context which is heavily characterized by data and data management, the amount of data available is likely to be beyond human cognition. There is clearly a computational approach needed for reaching better decision-making results (Kriegova et al., 2021). AI algorithms can better make sense of this data, directly influencing the decision-makers’ capacity to produce and evaluate alternative choices. The process of interpreting and predicting based on large amounts of data, in combination with the humans’ experience and sense making, helps to overcome the traditional impediment of numeracy, which is often the culprit of numerous failed diagnoses and risk estimations (Li and Chapman, 2020). Thanks to this combination of data processing algorithms and humans’ tacit knowledge, decision-makers can better tackle uncertainty related to complex and unknown cases (Mesko, 2017; Shrestha et al., 2019), and expand their rationality beyond bounded rationality. This is a balancing act on the traditional rationality versus intuition paradox: AI helps humans to overcome their bounded rationality by evolving their problem-solving, and humans expand AI’s contextual and historical limitations by exercising their tacit knowledge (Jarrahi, 2018; Shrestha et al., 2019; Langley et al., 1995).

Decision-Making Process: Change of Organizational Processes and Structures

With the emerging integration of human–AI collaborative decision-making, organizations need to prepare for ways to adapt processes and structures around such a new collaborative framework. Such structural changes in decision-making need to take into account the safety-critical nature of decision-making in healthcare, and build on knowledge integration and coordination between new and existing actors.

In regard to safety criticality, Lakemond et al. (2021) note that when organizations integrate evolving technologies, they need to pay attention to maintaining safety and reliability, particularly in high-safety contexts, such as personalized medicine. With most AI solutions integrated in CoIS, the reliability of the data and data sources is one of the most crucial aspects of safety. Aspects such as credibility of data source, accuracy, relevance, and biases are critical for the functionality of machine learning and other types of AI solutions which depend on data. If the data are not up to date or checked for biases, the input of AI results in decision-making can have serious consequences for safety and reliability, particularly in high-risk decisions, including risks to human life. In addition, being unable to adhere to safety principles can seriously undermine trust in decision-making, particularly related to decision inputs from AI. That is why data management is critical for safety in decision-making in CoIS. Another aspect of safety criticality in decision-making can be considered regulatory frameworks around the implementation of AI solutions. Before deciding to implement AI in decision-making processes, organizations need to consider performing risk analysis as well as analyzing safety compliance of such solutions. This can be achieved with continuous cooperation with AI providers and developers, regional and national frameworks for safety and ethics. The risks for implementing AI solutions need to be mitigated, and aspects of accountability need to be considered.

The integration of AI decision-making processes requires organizations to consider how such integration can fit into existing operational workflows and the existing knowledge base of the organization. This can create potential changes in organizational roles as well as knowledge integration strategies. Regarding roles, the increased integration of AI solutions in decision-making requires humans to have an increased interaction with these technologies, and a need for understanding of ways to operate and interpret their results. In the case of personalized medicine, physicians and other healthcare staff who interact with AI need to have a baseline knowledge of computational language and how to interpret the decisions before they can make sense and progress through the decision-making. That can contribute to additional complexity within the systems. In this data-intensive computing era, knowledge acquisition requires new scientific methods, and knowledge is represented in immensely complex and rapidly accumulative datasets (Miller, 2019). Additionally, in this data-intense context, knowledge acquisition can be difficult. Organizations are expected to build novel knowledge acquisition, retention, and integration capabilities from the collaborative decision-making processes, to increase overall organizational knowledge and improve decision-making in the future. Another challenge for health professionals connected to knowledge derived from data is how to translate technol-

ogies to the human condition (Miller, 2019). Additionally, different types of data structures require different computational approaches.

Humans can easily transfer past experiences and expertise to new tasks; AI generalizes poorly to new datasets and this causes considerable failures (ibid.). Finally, organizations in CoIS are expected to interact with a growing number of actors, including technology developers and vendors, data platforms, cloud systems, and much more. This increase in stakeholders creates additional challenges for CoIS, including added complexity and possible difficulties in coordination.

Decision Space: Driven by an Increasing Number of Elements in the Context

The combination of strengths of humans and AI in different areas of decision-making demand that organizations understand not only the strengths and weaknesses of both, but also how to capture learning and knowledge that is generated from the context. Learning from previous errors and several sources is a characteristic of high-reliability organizations (Roberts and Bea, 2001). Reliability in CoIS can be observed through two main characteristics: enactment and requisite variety. People need to create meaning in their context to produce reliability (Weick, 1987). This means that within a given system, people create both problems and solutions, and it is within their own perception of the context and their decisions that reliability is achieved. Narration, sense-making, and storytelling make possible decision generativity based on scenario creation and evaluation in cases when the system is too complex for linear and rational decision-making (ibid.). In the latter sense, it has to do with both the analytical and intuitive limitations that humans deal with in the complex challenges. The introduction of AI creates additional opportunities when navigating decision-making. As previously noted, decision-making in CoIS is continuously evolving, meaning that the process of decision-making is made of several decision points over time, using multiple sources of evidence, as well as input from AI. According to Jarrahi (2018), combining the analytical strengths of AI and the intuitive strengths of humans can leverage the strengths of both. When AI is combined with humans in decision-making, it can create the opportunity for humans to obtain and consider a larger data and input set than they would have been able to themselves, creating in this way a different framing of the problem and solution space in decision-making. Not only that, but also with AI's ability to process and analyze this data better and faster than humans, the human decision-maker can focus on finding better and creative solutions for complex problems. AI systems can generate more alternative solutions for well-defined and isolated problems, and compare these solutions for humans to evaluate the best solutions, augmenting in a way Simon's model of intelligence, design, and choice (Langley et al., 1995). This generative decision-making perspective pushes decision-makers to fundamentally rethink their initial choices for that particular isolated event. For instance, it can provide additional opportunities for personalized medicine, where physicians and healthcare decision-makers need to change their medical decision-making models, from segregating patients by disease types and subtypes, towards treating each patient

individually based on their personal genomic, historical, and environmental data (Duffy, 2016). Additionally, by involving AI in such a generative model, humans can focus on expanding knowledge about the problem, iterating between concepts shared between them and AI, and focus on finding creative solutions that come as a result of the intuitive decision-making approach. This decision-making model would also bring personalized medicine one step closer to reaching purposeful healthcare, by allowing physicians to spend more of their efforts and time with the patient, and to exercise the human touch (Mesko, 2017).

Furthermore, the reliability principle allows organizations to capture and gather a variety of information and knowledge (requisite variety) which will enable them to better diagnose and cope with problems in the systems (Roberts and Bea, 2001). On the one hand, too much richness introduces the inefficiencies of overcomplication, and too little media richness introduces the inaccuracy of oversimplification. On the other hand, ignoring the media and information may result in inaccuracy or oversimplification. The authors suggest that a team of divergent individuals has more requisite variety than a team of homogenous individuals. Collective knowledge, which is greater than individual knowledge, increases requisite variety, which in turn improves reliability. In the context of CoIS, as presented previously, this heterogeneity and diversity of knowledge is represented in terms of both individuals and actors, as well as technological diversity, including AI. However, issues can arise when trying to manage larger heterogenous groups, such as delegation of responsibility and trust (Weick, 1987).

CONCLUSION

The fast-paced integration of AI in complex systems is proving potentially valuable for decision-making, but is not without new challenges and implications for organizations. The main contribution of the chapter is an initial framework for understanding organizational decision-making when AI becomes an integrated part of complex and increasingly intelligent systems. The framework contains three areas that need to be addressed, that is, the decision-maker, the decision-making process, and the decision space. The chapter outlines several new prerequisites for decision-making in CoIS, including the increasing importance of understanding human–AI collaborative decision-making, necessary changes in organizational processes, structures, and the role of the context in which human–AI decision-making is taking place (Table 9.2). The emerging field of personalized medicine is used to illustrate how human–AI decision-making is anchored in a complex context. There are clear indications that the three areas become increasingly intertwined in a context with high complexity, where traditional human-based expert approaches, devices, and imaging support, and an increased integration of AI solutions, including the use of data from various sources, form a new decision landscape. By outlining this and taking it into account in further research and practice, a more profound understanding of the main challenges and opportunities related to AI–human decision-making can be achieved.

The context of personalized medicine is of course very specific and peculiar, but many of the challenges and perspectives needed in other contexts, such as intelligent transportation, smart cities, and so on, are rather similar. The three general areas are suggested to be applicable in many different contexts, for instance related to autonomous driving contexts, the areas of decision-making, decision-making process, and decision space are relevant, not least when a temporal perspective is considered (for example, perspectives of initial system design, incremental updates, and the actual operation of the system). An awareness of the three areas could potentially assist practitioners in understanding the new decision-making landscape and its implications, when complex systems and organizations are turning into increasingly complex intelligent systems and organizations. This could guide new practices in human–AI decision-making. A combination of several of the areas point at important aspects to be addressed, such as the agency distribution between human and AI, the increased embeddedness of an understanding in its context, that is, situatedness, and the importance of having a temporal perspective.

Given the new emerging decision-making landscape, it is not yet fully clear how decision-making processes will materialize in the future. For instance, the involvement of AI in decision-making may not only raise questions of creating appropriate and well-functioning systems, but also raise issues regarding responsibility and accountability. This is clear not least in a medical context, but is also relevant for all systems that fulfil a critical role in society. Furthermore, new types of actors, such as algorithm providers, system engineers, and even data platforms and legal frameworks, are entering the landscape. This may pose additional challenges and a change in the surrounding landscape of organizational decision-making, potentially increasing complexity even further.

While AI becomes increasingly important, it is clear that AI alone will not solve the century's most pressing societal issues. In the field of personalized medicine and other private and public contexts, researchers, professionals, policy-makers, and so on cannot be expected to be coasting through the process and watch as technological advancements are made. Rather, strategic mindfulness about the potential of AI in making better decisions in the field is necessary, including a deep exploration of the challenges related to its potential and its application. This most certainly requires a transition that may be evolutionary, but still requires radically new perspectives on organizations and organizing (Pritchard et al., 2017). This includes an awareness of the ethical and social issues that arise as a result of AI application in complex intelligent systems in general, and medical decision-making in particular. Most probably, the management and organizational implications will go hand in hand with policies and regulations for ethical standards in applied AI (Duffy, 2016). In addition, in parallel with ethical regulations, policy-makers must consider issues of accountability and responsibility standards; for instance, medical decision-making under uncertainty (Mesko, 2017), which will help with building medical practices for hospitals and care centers.

While the proposed framework is an initial step for outlining some of the main new prerequisites for decision-making in CoIS, we are aware of several potential

limitations. Firstly, the term “AI” is an overarching term that includes different kinds of computational technologies, with different degrees of autonomy and complexity. Taking into account the diversity of AI technologies, it is important to consider that the implications of different technologies need to be considered differently in organizations. Secondly, as the framework itself suggests, AI is a very context-dependent technology. This characteristic implies that different types of decisions, organizations, and industries need to consider contextual and situational characteristics of decision-making and decision outcomes. This can have especially important implications when considering training data sources and quality, but also can be related to organizational structures and strategy. Finally, what this proposed framework does not address is the discussion around transparency, interpretability, and trust, which in our opinion are an essential part of AI implementation in CoIS, especially in the context of healthcare.

This chapter represents a first exploration of AI–human decision-making in the context of complex systems. Currently, to our knowledge, there exists no organizational decision-making perspective that incorporates AI in the context of complex systems. Also, the technical challenges and opportunities of AI have just started to become expanded and disseminated into the managerial theories of decision-making (see, e.g., Puranam, 2021). The three areas outlined in the chapter—the decision-maker, the decision-making process, and decision space—all need to receive further attention and in-depth research efforts to explore the findings of this research further. In addition, in order to capture the decision-making dynamics, the interaction between the three proposed perspectives needs further investigation. Such insights can contribute to a fuller understanding that is essential to overcome challenges related to the implementation of AI–human decision-making in general, and AI-based personalized medicine in the healthcare sector particularly.

ACKNOWLEDGMENT

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program—Humanities and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation.

NOTE

- 1 <https://www.icpermed.eu/en/icpermed-medicine.php>.

REFERENCES

- Bohr, A., and Memarzadeh, K. (2020). The rise of artificial intelligence in healthcare applications. In Bohr, A., and Memarzadeh, K. (eds), *Artificial intelligence in healthcare* (pp. 25–60). Elsevier.
- Cohen, Michael D., March, J.G., and Olsen, J.P. (1972). A Garbage Can Model of Organizational Choice. *Administrative Science Quarterly*, 17, 1–25.
- Cyert, R.M., and March, J.G. (1963). A behavioral theory of the firm. *Englewood Cliffs, NJ*, 2(4), 169–187.
- Davies, A., and Hobday, M. (2005). *The Business of Projects: Managing Innovation in Complex Products and Systems*. Cambridge University Press.
- Dembrower, K., Wählin, E., Liu, Y., Salim, M., Smith, K., Lindholm, P., Eklund, M., and Strand, F. (2020). Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: A retrospective simulation study. *The Lancet Digital Health*, 2(9), e468–e474.
- Dignum, V. (2019). *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer.
- Duffy, D.J. (2016). Problems, challenges and promises: Perspectives on precision medicine. *Briefings in Bioinformatics*, 17(3), 494–504. <https://doi.org/10.1093/bib/bbv060>.
- Dwivedi, Y.K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., and Eirug, A. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 57, 101994.
- Gavetti, G., Levinthal, D., and Ocasio, W. (2007). Perspective—Neo-Carnegie: The Carnegie school's past, present, and reconstructing for the future. *Organization Science*, 18(3), 523–536.
- Hatchuel, A. (2001). Towards design theory and expandable rationality: The unfinished program of Herbert Simon. *Journal of Management and Governance*, 5(3/4), 260–273.
- Hobday, M. (1998). Product complexity, innovation and industrial organisation. *Research Policy*, 26(6), 689–710.
- Jarrahi, M.H. (2018). Artificial intelligence and the future of work: Human–AI symbiosis in organizational decision-making. *Business Horizons*, 61(4), 577–586.
- Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, 93(5), 1449–1475.
- Kriegova, E., Kudelka, M., Radvansky, M., and Gallo, J. (2021). A theoretical model of health management using data-driven decision-making: The future of precision medicine and health. *Journal of Translational Medicine*, 19(1), 1–12.
- Lakemond, N., Holmberg, G., and Pettersson, A. (2021). Digital transformation in complex systems. *IEEE Transactions on Engineering Management*, 71, 192–204. doi: 10.1109/TEM.2021.3118203.
- Langley, A., Mintzberg, H., Pitcher, P., Posada, E., and Saint-Macary, J. (1995). Opening up decision making: The view from the black stool. *Organization Science*, 6(3), 260–279.
- Le Masson, P., Hatchuel, A., Le Glatin, M., and Weil, B. (2019) Designing decisions in the unknown: A generative model. *European Management Review*, 16, 471–490. <https://doi.org/10.1111/emre.12289>.
- Li, M., and Chapman, G.B. (2020). Medical decision-making. In Sweeny, K., Robbins, M.L., and Cohen, L.M. (eds), *The Wiley Encyclopedia of Health Psychology*, Vol. 2 (pp. 347–353). Wiley.
- Liu, S., Ko, Q.S., Heng, K.Q.A., Ngiam, K.Y., and Feng, M. (2020). Healthcare transformation in Singapore with artificial intelligence. *Frontiers in Digital Health*, 2, 592121.
- Lu, C.Y., Terry, V., and Thomas, D.M. (2023). Precision medicine: Affording the successes of science. *NPJ Precision Oncology*, 7(1), 3.

- March, J.G. (1978). Bounded rationality, ambiguity, and the engineering of choice. *Bell Journal of Economics*, 9(2), 587–608.
- Mesko, B. (2017). The role of artificial intelligence in precision medicine. *Expert Rev Precis Med Drug Dev*, 2(5), 239–241.
- Miller, D.D. (2019). The medical AI insurgency: What physicians must know about data to practice with intelligent machines. *NPJ Digital Medicine*, 2(1), 62.
- Mintzberg, H. (1978). Patterns in strategy formation. *Management Science*, 24(9), 934–948.
- Pritchard, D.E., Moeckel, F., Villa, M.S., Housman, L.T., McCarty, C.A., and McLeod, H.L. (2017). Strategies for integrating personalized medicine into healthcare practice. *Personalized Medicine*, 14(2), 141–152.
- Puranam, P. (2021). Human–AI collaborative decision-making as an organization design problem. *Journal of Organization Design*, 10(2), 75–80.
- Raisch, S., and Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review*, 46(1), 192–210.
- Roberts, K.H., and Bea, R. (2001). Must accidents happen? Lessons from high-reliability organizations. *Academy of Management Perspectives*, 15(3), 70–78.
- Russell, S.J., and Norvig, P. (2016). *Artificial Intelligence: A Modern Approach*. Pearson Education.
- Rutter, H., Savona, N., and Glonti, J.B., Bibby, S., Cummins, D.T., et al. (2017). The need for a complex systems model of evidence for public health. *The Lancet*, 390(10112), 2602–2604. doi:10.1016/S0140-6736(17)31267-9.
- Shrestha, Y.R., Ben-Menahem, S.M., and Von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review*, 61(4), 66–83.
- Simon, H.A. (2013). *Administrative Behavior*. Simon & Schuster.
- Trabizi, B., Lam, E., Girard, K., and Irvin, V. (2019). Digital transformation is not about technology. *Harvard Business Review*. <https://hbr.org/2019/03/digital-transformation-is-not-about-technology>.
- Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, 185(4157), 1124–1131.
- Wang, X. (2022). New strategies of clinical precision medicine. *Clinical and Translational Medicine*, 12(2), 1–3.
- Weick, K.E. (1987). Organizational culture as a source of high reliability. *California Management Review*, 29(2), 112–127.
- Yu, K.H., Beam, A.L., and Kohane, I.S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, 2(10), 719–731.

10. Addressing the knowledge gap between business managers and data scientists: the case of data analytics implementation in a sales organization

Stella Pachidi and Marleen Huysman

INTRODUCTION

The explosion of data generated through the digitalization of every facet of organizational life, along with the tremendous growth in computing power, have given rise to several learning technologies that are capable of supporting decision making processes across all levels in the organization's hierarchy (Faraj et al., 2018). For example, neural networks are commonly used to support decisions about loan request approvals, by analyzing parameters of bank clients such as age, solvency, or credit history based on historical data representing past loan approval decisions (Benbya et al., 2021). Aiming to fulfil the ideal of rationality, organizations increasingly introduce these tools, hoping to make better-informed and faster decisions (Davenport et al., 2010).

Despite the promises of learning algorithms, only a few organizations have managed to successfully implement such technologies and leverage their full potential (Deloitte, 2020). While much emphasis is placed on the economic and strategic benefits gained from automating and augmenting decision making procedures with the introduction of learning algorithms, the associated changes in the nature of work, the ways in which knowledge is produced, the decision making culture, and the integration of digital talent are often neglected or taken for granted. Our objective in this chapter is to explore those organizational challenges associated with an organization's shift to data-driven decision making and discuss potential ways through which organizations may overcome these challenges.

We suggest taking a practice perspective on decision making, which pays attention to the everyday situated actions that people perform in the workplace (Nicolini, 2012). Looking at the work practices allows us to understand how actors not only individually but also collectively make judgments and decisions in their everyday work. Thus, a practice perspective enables us to look at the practical changes in how people act and interact when their organization adopts tools to shift to data-driven decision making, and thus helps us to analyze the associated challenges that managers need to address in the workplace. Because knowledge and views on how decisions are made are deeply engrained in people's work practices (Carlile, 2002), a practice perspective helps us to explain the difficulties that people face in changing

their practices when learning algorithms are introduced in their work, and allows us to understand how they cope with those difficulties.

We report on a qualitative study performed in the business-to-business sales department of TelCo, a large telecommunications organization located in a Western European country, focusing on the challenges that arose when data analytics was introduced to support account managers in their everyday planning decisions. We found that account managers struggled to change their decision making practices because their view of what kind of data mattered and how that informed judgments and actions was substantially different from the data and methods inscribed in the analytics tool. Their views also clashed with those of the data scientists, who were highly focused on the (assumed) superiority of their tools and failed to collaborate with the account managers to enable the transformation of the sales work practices.

In the following sections, we outline the theoretical background of this study and subsequently present our research methodology and case description. We then continue with the analysis of our findings. Finally, we discuss the theoretical insights for information systems and organizations literature and unpack the practical implications for managers and organizations.

LITERATURE REVIEW

The Consequences of Learning Algorithms for Organizational Decision Making

Organizations are always determined to process as much information as possible, in order to satisfy their strive for rationality, that is, choosing the optimal based on a rational calculation of alternatives and their consequences (March and Simon, 1958; March and Olsen, 1975). Given that humans are bounded by cognitive limits and time available to search for and process information, such rational decision making is hardly ever perfect (Simon, 1976) which historically led to the development and adoption of tools that are believed to enable rationality (Jarzabkowski and Kaplan, 2015; March, 2006; Pachidi and Huysman, 2017). The underlying assumption is that the more information a tool can process, the more complete examination of alternatives it will be able to offer, and thus the more rational the supported actions will be.

Nowadays organizations, in their strive for rationality, increasingly rely on actionable insights produced by learning algorithms to support or at times even automate various decisions. Learning algorithms include a category of algorithmic technologies including data analytics, recommender systems, and artificial intelligence that provide responses, classifications, dynamic predictions, and recommendations by adjusting output to the large datasets they are fed with (Faraj et al., 2018). As we explain in the following paragraphs, learning algorithms encapsulate distinct views and methods for what information matters for a specific decision, how that is produced, and what decision criteria are considered.

Learning algorithms prioritize information toward types of objects that can be captured in digital form and even quantified (Kitchin, 2014). For example, analytics algorithms used in marketing prioritize past transactions, social media entries, and demographic data as digital representations of the social sphere (Boyd and Crawford, 2012; Kitchin, 2014). Furthermore, the selection of data fed into those algorithms, along with the practices of generating, collecting, packaging, and preprocessing this data, are all dependent on the material configurations used as well as the humans involved in these processes – including data scientists, data curators, but also humans who are involved in generating the data (Gitelman, 2013; Leonelli, 2016).

Learning algorithms employ a quantitative mode of inquiry that includes the use of statistical methods, machine learning, and other quantitative processes chosen by data analysts, data scientists, or other professionals with quantitative and digital expertise who are involved in the programming of the algorithms. The processes followed are usually black-boxed from the users, who are presented with the insights produced, or at the very best with a high-level description of the selected quantitative methods (Annany and Crawford, 2018). Other details of the analytical process, even moments of doubt such as when a risk analyst may suspect that a model is overfitting, are usually opaque or fully hidden (Amoore, 2019). Yet, people in organizations tend to treat insights produced via quantitative modes of inquiry as accurate, fair, and objective (Annany, 2016; Porter, 1996; Waardenburg et al., 2022). Thus, managers tend to grant high authority to learning algorithms (and to the experts involved in their making) by allowing them to even run automatically to interact with customers, provide financial advice, adjust product prices, or even control employee performance (Faraj et al., 2018; Introna, 2016; Kellogg et al., 2020).

Unavoidably, the introduction of learning algorithms to transform how judgments, decisions and evaluations are made in an organization will be likely associated with fundamental changes in the regime of knowing. Regimes of knowing include what kind of knowledge matters, what valuation schemes are used for evaluating objects, people, and performances, and who has the authority to define how the work should be organized to ensure a skilled performance (Pachidi et al., 2021). Thus, to understand the organizational challenges that managers and leaders face in the shift to data-driven decision making one needs to pay attention to deeper issues and conflicts associated with the production of knowledge, and to the interactions between the actors involved. In the next sub-section, we discuss how a knowing perspective can be helpful in this endeavor.

Taking a Knowing Perspective to Understand Organizational Shifts to Data-Driven Decision Making

When it comes to studying organizational shifts to a data-driven decision making culture, the practice perspective helps to trace the actual changes that people perform in their work activities and the negotiations made around those changes. Taking a practice perspective focuses our attention on the activities through which actors make decisions, and on their interactions with others around those activities

(Feldman and Orlikowski, 2011; Nicolini, 2012). This allows us to look at the collective arrangements around which specific types of information are deemed valuable, how accountability and authority become allocated, and tacit assumptions around how valuable knowledge should be created and choices should be made (Pachidi et al., 2021).

From a practice-based approach, choices are not treated as the outcome of deliberate prior planning or intention, but instead, their non-deliberate emergence through everyday practical coping is the focus of attention (Nicolini, 2012). This is because choices and decisions are viewed as enactments of practical knowledge that is acquired through the performance of a practice and becomes materially inscribed. From a practice-based approach, knowing is inextricably related to how people perform a practice. This means that people who are engaged in the same practice are likely to share knowledge and develop shared understandings, interpretations, and interests (Carlile, 2002). However, when faced with a different practice, actors may easily face misunderstandings, communication problems, and conflicts (Bechky, 2003; Carlile, 2004). In the case of digital transformation efforts, when digital experts are introduced to transform the organization's processes, products, and services, incumbent practitioners may struggle to collaborate with the digital experts due to their (often) fundamentally different practices. Given that decision making is deeply seated in people's practices, knowledge boundaries will most definitely arise when a new group of experts (data scientists, data analysts, and so on) enter the workplace to introduce tools for data-driven decision making.

At times of transformation, such as when an organization shifts to a data-driven decision making culture, managers and practitioners are likely to become reflective about their work and ways of acting, and vocally express their differences (Boland and Tenkasi, 1995; Levina and Vaast, 2005). This happens because when practices break down temporarily, practitioners distance themselves from their customary ways of acting and reflect on their actions (Yanow and Tsoukas, 2009). In such cases, practitioners articulate their theoretical reflections and compare the different perspectives they have with other practitioners from their community. These comparisons may even take the form of evaluating what is a good way of acting, and what is right or wrong (MacIntyre, 1981; Nicolini, 2012). Thus, at breakdown moments, knowing in practice becomes temporarily visible as practitioners' reflections involve analyzing what knowledge matters, and through what practices that is obtained (Van den Broek et al., 2021; Pachidi et al., 2021). Contextualized beliefs, meanings, values, standards of excellence, routines, and situated knowing that are deep-seated in work practices may come to the surface through their reflections, and be contested by other practitioners.

Looking at actors' theoretical reflections on knowing may help us to gain a better understanding of how they cope with breakdowns in their practices that will unavoidably arise when the organization introduces learning algorithms to shift to data-driven decision making, as well as how they contest their situated views with data scientists or other experts developing and advocating the use of these tools. Qualitative studies of transformation processes are needed to address these issues. In the next section,

we introduce the research setting in which we observed and analyzed the contestation that unfolded when data analytics was introduced to drive planning choices in sales work. While we have analyzed the process of transforming the regime of knowing in other publications (Pachidi et al., 2021), here we focus on the conflicting strong views of actors about what knowledge matters, and how decisions should be made, which deeply affected the change process.

RESEARCH METHODOLOGY

We performed an inductive longitudinal study in TelCo, a telecommunications organization. We spent 24 months studying what happened after analytics was introduced in TelCo's Sales, specifically the department of business-to-business sales that targeted medium-sized enterprises. The introduction of data analytics was meant to make everyday planning and decisions involved in sales work more data-driven.

Research Setting

In Sales, account managers were assigned to a fixed set of 250–300 customers. They were responsible for maintaining a relationship with those customers to identify sales opportunities and generate leads. The account managers worked in pairs consisting of one internal and one external account manager. Internal account managers worked every day from the office and communicated with the customers over the phone. The external account managers worked mostly on the road, visiting customers in their offices.

The Customer Intelligence team consisted of data scientists and analysts who used analytic techniques to analyze customer data and offer insights to Marketing. The data scientists' educational background was in engineering and econometrics, and they had much experience in quantitative data analysis. Their job was to offer data insights to their "internal customers," that is, marketers, campaign managers, product managers, and others. After having established their data analytics services in Marketing, the Customer Intelligence team aspired to start offering analytics services also to Sales.

Customer Intelligence introduced data analytics in Sales in January 2012, along with the introduction of the customer lifecycle management (CLM) way of working. The data scientists developed the CLM model, which provided a list of all medium-sized customers with several statistics, metrics, and forecasts about each customer's potential to buy a service at a specific time. The data scientists argued that the CLM model would help the account managers "to contact the right customer at the right time and with the right offer." The analysts ran the algorithms once a quarter to generate new data-based insights such as predictions regarding the customer's potential for a specific campaign.

In our qualitative study, we analyzed the tensions that emerged between the account managers and the analysts since the introduction of the CLM model. We

focused on understanding the practices of the account managers and the analysts, their views on the CLM model, as well as how they collaborated.

Data Collection

Our main source of evidence consisted of semi-structured interviews with data scientists, account managers, sales team managers, and other roles (Weiss, 1995). The interviews focused on people's work practices and their views of the CLM model. It was during those moments when people explained why the CLM model was or was not useful that they would explicate their views on what kind of knowledge mattered.

Interviews were complemented by ethnographic observations, conducted by the first author as a passive participant (Spradley, 1980). The first author observed account managers by shadowing them in order to understand how they worked and how they used the CLM model. She would sit next to an account manager for the whole day, to see how they conducted their everyday work, such as how they planned which customers to call or how they contacted their customers. Similarly, the first author also shadowed data scientists, and specifically those involved in the construction of the CLM model. She observed how they worked while preparing queries and algorithms for the CLM model. While shadowing, the account managers articulated what they had been doing on their computer, or what they had been discussing with a contact person on the phone. The analysts would explain the code that they were developing, along with their views on the CLM model, and how they expected it to be used by the account managers.

Another type of observation included observing meetings, including quarterly kick-off presentations where the data scientists presented the new version of the CLM model to the account managers. Finally, documents were used to triangulate information from the interviews, and in particular to verify retrospective information.

In total, we performed 78 interviews with a total recorded time of 73 hours, and 85 hours of observations, and we analyzed 75 documents.

Data Analysis

We analyzed the data following a process research approach to understand how things unfolded (Langley, 1999). We started open coding in parallel with our data collection. Initially, we focused on aspects related to how the analysts and account managers worked, in order to understand their practices. Furthermore, we coded for events related to the introduction of the CLM model in Sales and how the two groups interacted. Based on the codes that we had developed, we created an event list which captured the time sequence of events at TelCo. The event list helped us to construct the case narrative, which consisted of a detailed story that helped to identify the chronology of events, as well as linkages and patterns between different types of events, and establish analytical themes (Pettigrew, 1990). Following several iterations between our data and the literature, we developed second-order codes around

the respondents' reflections, including what kind of information matters and how decisions should be made.

In the next section, we provide the descriptive first-order narrative of how the decision making culture in TelCo Sales became data-driven, and we analyze how the clash between account managers and data scientists unfolded. In the discussion section, we theoretically reflect on the findings and provide recommendations for practice.

INSIGHTS FROM THE FIELD STUDY AT TELCO

In this section, we analyze how the clash between the analysts and account managers unfolded with the introduction of analytics in TelCo Sales. We have organized our narrative into four main phases that are fundamental in understanding why and how the deep-seated knowing epistemic differences were significantly impeding the collaboration between account managers and data scientists, and the integration of the CLM model in account managers' work practices. Epistemological differences were significantly impeding the integration of analytics with the account managers' knowing practices. In the first phase, the introduction of the CLM model brought the account manager's practice face-to-face with the analyst's practice. This triggered both groups to reflect on what kind of knowledge matters to them. Their epistemic differences came to the surface, and the clash between the two groups arose. In the second phase, both groups performed attempts to resolve their tensions. However, it turned out to be difficult for them to integrate their different practices. Thus, in the third phase, the analytics model was forced onto Sales and the organization appeared to converge towards a data-driven way of working. However, in the fourth phase, it became evident that the relational way of working was still enacted by account managers who had diffused in other sales channels. They fostered the re-emergence of the old knowing practices in new contexts.

The Clash between Account Managers and Data Scientists Due to Epistemic Differences

The data scientists from Customer Intelligence introduced their analytical model to Sales with the goal to establish customer lifecycle management (CLM) as a way of working that would bring efficiency to the sales process. Their motto was that the CLM model would help account managers to contact "the right customer, at the right time, and with the right offer." The output of the model was presented in a spreadsheet format, which contained a list of all customers and the different predictions. The data scientists assigned the customers to different customer segments (A, B, C, D) depending on each customer's revenues and potential. Also, for each portfolio, the data scientists calculated the phase of the customer lifecycle.

At the start of the introduction of the CLM model, the data scientists gave a kick-off presentation to the nine sales teams. Although they explained how it could

help them to approach their customers more efficiently and effectively, they received a rather negative reaction from the sales teams. The account managers worked in a very different way compared to the data scientists. The knowing practices of the two groups—that is, the practices through which they generated and used knowledge in their work—were very different.

Account managers worked with the ultimate goal to generate sales opportunities that would eventually turn into leads and later into orders. They approached each customer as unique, and aimed to understand how their business worked to best serve them. They mainly gathered information and identified selling opportunities in conversation with their customers. Thus, they found it important to sustain a personal relationship with the contact persons (for example, by remembering personal details) to build trust with them. Account managers also had access to several information systems in which they could find relevant information such as the dates on which the customers' subscriptions expired. They used the customer relationship management (CRM) system to store information about the customers and potential sales opportunities. They also often kept their own shadow administration, such as Excel files.

On the other hand, data scientists had very different knowing practices with which they addressed the sales task. These practices involved collecting data that they later analyzed by running different models to create data insights relevant to Marketing and Sales. More specifically, they spent time thinking about how to create predictive models to support Marketing and Sales, and what variables each model should include. They collected data by running queries in TelCo's databases, or by acquiring data from external sources such as Nielsen. They used different types of data analysis techniques such as decision trees and regression models to create customer profiles and run various predictions; for example, about the impact of campaigns. They provided analytics insights to their "internal customers" (that is, marketers and sales employees) in the form of a presentation, in which they visualized the insights and provided advice on what people should do based on the outcome of their analysis.

As the knowing practices of the two groups of practitioners were so different, collaboration was almost impossible. The data scientists experienced resistance from the account managers and their sales team managers, who did not want to use the CLM model. As the tensions intensified, it became clear that each group did not trust the knowledge of the other, and they were thus unwilling to adjust their practices. For example, the data scientists argued that it was impossible for an account manager who deals with 250–300 customers to know each customer perfectly. They insisted that the predictions of the CLM model were important to use for serving the customers effectively. On the other hand, the account managers found that the information in the model was not up-to-date, since it was based on historical data, and questioned the effectiveness of the model, especially since it only gave probabilities.

More fundamentally, the two groups had deep-seated differences about what kind of knowledge mattered and how decisions should be taken. Account managers prioritized relationship-based, context-specific knowing of their customers; maintaining a good relationship with the customer was important in order to know their challenges and plans for the future, so that they could identify their needs and

translate them into sales opportunities. For the account managers, deciding when to contact which customer, how to talk to them, about which portfolio, and so on, was done intuitively. They often referred to the practice of generating sales opportunities as “feeling” the customer’s needs, or “feeling” when there was potential for sales.

On the other hand, the data scientists believed that the insights of the quantitative analysis constituted knowledge about TelCo’s customers and how they should be approached. In other words, if a model showed X about customer Y, then Y needed to be approached with the Z campaign, and so forth. For the data scientists, customer behavior could be modeled with data by applying different segmentation techniques, which yielded customer profiles. By analyzing the data with prediction models, they could also assign probabilities to the future actions of the customers. They believed that “hard” data that came from the databases was more trustworthy information than the interpretation of the account managers, which could be falsified.

Efforts to Resolve Tensions

To eliminate resistance, the data scientists tried to find a way to show that their model was effective, hoping that they would gain more support from management in this way. Thus, they asked the account managers to register the use of the model, by assigning a specific code every time they stored a lead (that is, when a customer wanted to receive an offer) in the CRM system if they had created that lead with the help of the CLM model. In this way, they could track how many leads were created with the model.

Furthermore, the data scientists relied on a campaign manager who took on the mission to promote the CLM model to the account managers. Together with the data scientists, the campaign manager visited each sales team at the start of every quarter to present the new analytics insights.

The campaign manager identified one sales team manager who was more positive about the data analytics model than the rest. She was a former marketer and had recently started working as a sales team manager. She wanted to appear as a change leader in Sales and was willing to support the data campaign manager and the data scientists in getting the model established in Sales, even though she did not believe that the predictions of the CLM model could be more accurate than the account manager’s interpretation. Therefore, she encouraged her account managers to register the code regarding the use of the CLM model even when they did not use it, to help establish this more structured way of working in Sales. The data scientists saw that team manager as an ambassador who could promote the CLM model to the other sales team managers. Since her team registered the CLM code in the CRM system, they started including a benchmark in their presentations to the sales teams, to show that certain account managers were successful with using the model.

In addition, the data scientists started updating the model regularly, based on the feedback that they got from the account managers. They often added information that the account managers had requested (such as expiry dates of contracts, which they normally had to search for in different databases), as long as they found it relevant. In

this way, they could meet the information needs of the account manager and turn the CLM model into a “complete customer view.”

The account managers appreciated the additional information that the data scientists were adding to the CLM model. It was much easier to find the expiry dates of customer contracts in one file, instead of having to search in different systems. From their side, they also attempted to improve their collaboration with the data scientists, by selectively using parts of the CLM model. More specifically, they used the CLM model as a backup check, or when they did not feel “in control” (that is, if they had not had contact with a specific customer recently), or on focus days (when they had to make “cold calls” to increase sales in a specific portfolio).

While each group was trying out parts of the other practice, they were also reflecting on the other epistemic practices. For example, the data scientists eventually acknowledged that the knowledge generated by the account managers, who had direct contact with the customers, could also be valuable. Similarly, the account managers, while using part of the CLM model on certain occasions, were also reflecting on when data analytics could be useful. They further tried to understand how the data scientists calculated the insights. They also acknowledged that the information from the CLM model was useful, albeit under certain boundary conditions. For example, they would have to first confirm whether the recommendation regarding the potential sales opportunity was correct by contacting the customer, or they would interpret it combined with “their own intelligence.”

This interplay between trying out parts of each other’s knowing practices led to further refinement of their differences: each time they acknowledged that the other way of knowing could also be useful sometimes, they would conclude again that their way of knowing was better. The data scientists kept believing in the primacy of the algorithmic, data-based way of knowing. They were convinced that the registration of leads with the CLM code represented the actual use of the CLM model and consequently did not see any value in adding the account managers’ interpretations (stored in CRM) to the model. The account managers, on the other hand, argued that most times when they used the analytics model, they had to conclude that they already knew the information in it. While they acknowledged that the model could sometimes be useful, they insisted that the most important knowledge came from talking to their customers.

Thus, each group would stick to their own views around what knowledge mattered and how choices should be made, and it was very difficult for the two teams to combine their different epistemic practices into a hybrid one. Thus, the account managers had to turn to symbolic ways to deal with the pressure from the data scientists: they attended the meetings with the data scientists just to show conformity, while most of them started registering leads in the CRM system with the CLM code symbolically, without actually having used the model. The data scientists did not question the data with the registered leads in the CRM system. Instead, they expected that the data represented the actual use of the CLM model.

Appearing to Shift Towards Data-Driven Decision Making

The symbolic actions of the account managers, combined with the fact that the data scientists assumed that the registration data represented the actual usage of the model, eventually led to validating the CLM model at TelCo. One year after introducing the model, the data scientists presented all data regarding leads generated with the use of the CLM model to higher management, who started formally supporting the CLM model. In a meeting with all sales teams, the Marketing and Sales director announced that this should be a standard way of working for Sales. After this, the CLM model became more and more established in Sales. For example, it got incorporated into the training for new account managers.

Having support from higher management, the data scientists put more pressure on the sales teams to use the analytics. This, however, triggered even more symbolic registration; many account managers would register leads with the CLM code even if they had not used the model, just because it was considered to be an established way of working and higher management expected them to use it.

The account managers had not expected that the validation of the CLM model would have even more severe consequences for them. As the telecom industry had been suffering from increased competition and lower revenues, TelCo had to simplify its processes and reduce its costs. This led to the decision to reorganize Sales by firing most of the account managers and outsourcing their work to business partner organizations. As the CLM model appeared to be an efficient and effective way of working, higher management presented analytics as the way of working towards efficiency.

Thus, the sales function in TelCo appeared to have shifted toward data-driven decision making. The data scientists no longer needed to prove the effectiveness of their model. The sales employees realized that their symbolic actions had ended up establishing analytics as the way of working at Sales, but also that the model was eventually replacing them, since it appeared to fit with the changes towards more efficiency. Most account managers were fired, while the ones who stayed had to work in a very different way, which entailed less nurturing of the relationship with the customers and more creation of business opportunities. On the other side, the analytics practice was expanding in TelCo Sales: the CLM model was used to communicate sales opportunities to the business partner organizations. The data scientists also started expanding the model to more sales channels.

Old Knowing Practices Re-emerge in New Contexts

While data-driven decision making had appeared to become dominant in TelCo Sales, it soon became evident that this was only on the surface. Many of the account managers from Sales got hired by the business partner organizations, or remained in TelCo and moved to other sales channels. Those people had not changed their views or epistemic practices. They did not increase their use of the CLM model, but only used it occasionally, in similar ways as before, for example as a backup check.

Other employees at the business partner organizations also resisted change in their way of working, and hesitated to use the predictions from the CLM model. Instead, they would only use contract-related information that was included in the model, and would not act upon the predictions.

Furthermore, the relational knowing practices remained prevalent in other sales channels where the data scientists had tried to expand the CLM model. Thus, TelCo was realizing that there were boundaries where analytics may or may not be useful. For example, the data scientists reflected that the CLM model should have a different purpose in Sales Large, where an account manager dealt with only 15–20 large enterprises and already knew those customers well enough.

DISCUSSION

In this chapter, we have examined the challenges that organizations face when they introduce learning algorithms to shift to data-driven decision making. In our study at TelCo, where data analytics had been introduced to transform work in Sales, we found that data scientists and account managers had orthogonal epistemic practices, and their conflicting views around what knowledge mattered and how decisions should be made did not allow them to integrate their ways of working. Eventually, data analytics had to be forced upon the sales work.

Learning algorithms encapsulate distinct epistemic practices that are often orthogonal to the existing practices in the workplace (Pachidi et al., 2021). Our study highlights that when organizational members attempt to combine fundamentally different epistemic practices, their different views on what kind of knowledge matters and how decisions are made can come to the surface and intensify the clash between them. This can breach managers' expectations that introducing learning algorithms will make processes more effective and efficient, as it exposes the deeper issues faced in the implementation process, which often tend to be neglected. More specifically, our study illustrates that sometimes it is important for practitioners (and for researchers who study their practices) to articulate and comprehend those different views. Those fundamental differences may significantly impede the collaboration process between different groups of practitioners who are required to combine different knowing practices.

Our study suggests that tensions arise when practitioners encounter new knowing practices with which they must combine or replace existing practices. Indeed, in the case of TelCo, tensions emerged between the data scientists and the account managers, as the account managers resisted using the CLM model and expressed their resistance in their meetings with the data scientists. We already know from the literature on knowledge boundaries (Bechky, 2003; Carlile, 2002, 2004) that when groups who perform different knowing practices have to collaborate, they may face tensions due to their different situated understandings, different interests, and so forth. Our study shows that when actors have to combine their knowing practices and integrate their approaches to decision making, deep-seated differences in their views

about what kind of knowledge matters, or how decisions should be made, come to the surface. At TelCo, as the data scientists and account managers were experiencing difficulties trying to collaborate, they reflected on what kind of knowledge mattered to them. Those reflections were articulated in the kick-off meetings where they met each other, but also spontaneously during our interviews with them when they described what they did in their work and when they mentioned the CLM model. The account managers would argue that they did not want to use the model since what mattered in finding sales opportunities was contacting their customers and having a good relationship with them. The data scientists would suggest that the insights of their model were based on facts from databases, and these were necessary in order to know which customers to contact, when, and about which portfolios. Such fundamental views cannot be easily changed, and thus impede the collaboration further, to the extent that one epistemic practice may need to be imposed on the organization. In fact, it was so difficult for the account managers to act upon the predictive insights from the CLM model that they ended up resorting to symbolic actions and mainly pretended to use the analytics.

Our study contributes to the theories of knowing in practice by showing how deep-seated views around what kind of knowledge matters, and how decisions are reflected upon, shared, contested, and negotiated by practitioners. The epistemic differences mentioned above come to the surface when the practitioners encounter fundamentally different knowing practices because they have to collaborate to address the same task (for example, data scientists and account managers serving the customers in different ways). Such instances constitute severe breakdowns in the flow of practice that cause the practitioners to stop being absorbed in their practice and to engage in a more analytic and theoretical reflection. This reflection often includes an evaluative orientation towards the things that matter within their practice, such as what kind of knowledge matters, how to prioritize different types of information, and ultimately how to make judgments and decisions. In addition, our study contributes to the literature on collaboration across knowledge boundaries (Bechky, 2003; Carlile, 2002, 2004; Majchrzak et al., 2012) by emphasizing the need to understand epistemic differences and to engage in deep knowledge sharing to collaborate and eventually merge their knowing practices.

In closing, our study offers some managerial implications. First, our study helps to explain the difficulties that many organizations face when employing learning algorithms to shift to a data-driven decision making culture. It shows that employees' resistance is much more than mere resistance to change, or even fear of becoming obsolete. Instead, resistance may be related to deep-seated views about what kind of knowledge matters, and how decisions are made, which are inscribed in people's practices and come to the surface during times of breakdown. Second, our study shows that integrating digital talent into the organization is more complex than bringing in a new team of digital experts. It also requires developing the digital skills of non-digital natives, to help them understand how the introduced tools work and the potential value they may offer. But also, vice versa, it requires exposing the digital natives to the work practices of the practitioners who they are expected to support, to

better understand and appreciate the existing epistemic practices and how they may be enhanced, rather than replaced, through the algorithmic tools. This could happen, for example, by having the digital natives shadow the business practitioners for some time, or even having them temporarily work in the practitioners' role to experience first-hand what their work entails, as part of their induction program. Third, our study suggests that data scientists and managers need to engage in deep knowledge sharing to understand their epistemic differences and investigate ways to integrate their different approaches. This might be achieved by having the different groups collocated frequently to be exposed to the different practices. Another way to accomplish deep knowledge sharing can be through participatory design, which requires the users to participate in the design development process from very early on, and thus inscribe their epistemic practices in the algorithm (Waardenburg and Huysman, 2022). Fourth, our study shows that focusing on key performance indicators (KPIs) can be counter-productive when focusing on transforming the organizational culture in digital transformation. In our case, the number of leads that were registered with the code that indicated the use of analytics was not representative of the actual use. When setting such numerical targets, employees may simply try to game the numbers. At the very least, leaders need to question how the numbers are generated. Crucially, focusing on supporting the employees as they struggle with the changing nature of their work is more critical than focusing on whether they hit the assigned KPIs. Last, but certainly not least, our study highlights the essential role of proactive leadership in driving the organization's transformation. In the case of TelCo, leadership did not monitor the change process closely, but instead relied on the data presented by the data scientists, which had been produced through symbolic actions. Leaders need to develop their digital skills not only to inspire the workforce but also to become actively engaged in the transformation process.

REFERENCES

- Amoore, L. (2019) Doubt and the algorithm: On the partial accounts of machine learning. *Theory, Culture Society* 36(6): 147–169.
- Ananny, M. (2016) Toward an ethics of algorithms: Convening, observation, probability, and timeliness. *Science, Technology, and Human Values* 41(1): 93–117.
- Ananny, M., and Crawford, K. (2018) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media and Society* 20(3): 973–989.
- Bechky, B. (2003) Sharing meaning across occupational communities: The transformation of understanding on a production floor. *Organization Science* 14(3): 312–330.
- Benbya, H., Pachidi, S., and Jarvenpaa, S. (2021). Special issue editorial: Artificial intelligence in organizations: Implications for information systems research. *Journal of the Association for Information Systems* 22(2): 10.
- Boland Jr, R.J., and Tenkasi, R.V. (1995) Perspective making and perspective taking in communities of knowing. *Organization Science*, 6(4): 350–372.

- Boyd, D., and Crawford, K. (2012) Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication and Society* 15(5): 662–679.
- Carlile, P.R. (2002) A pragmatic view of knowledge and boundaries: Boundary objects in new product development. *Organization Science* 13(4): 442–455.
- Carlile, P.R. (2004) Transferring, translating, and transforming: An integrative framework for managing knowledge across boundaries. *Organization Science* 15(5): 555–568.
- Davenport, T., Harris, J., and Morison, R. (2010) *Analytics at Work: Smarter Decisions, Better Results*. Harvard Business Press.
- Deloitte (2020) *Thriving in the Era of Pervasive AI: Deloitte's State of AI in the Enterprise*, 3rd Edition. Deloitte Insights. <https://www2.deloitte.com/content/dam/Deloitte/cn/Documents/about-deloitte/deloitte-cn-dtt-thriving-in-the-era-of-persuasive-ai-en-200819.pdf>. Accessed on March 28, 2022.
- Faraj, S., Pachidi, S., and Sayegh, K. (2018) Working and organizing in the age of the learning algorithm. *Information and Organization*, 28(1): 62–70.
- Feldman, M.S., and Orlikowski, W.J. (2011) Theorizing practice and practicing theory. *Organization Science* 22(5): 1240–1253.
- Gitelman, L. (2013) *Raw Data is an Oxymoron*. MIT Press.
- Introna, L.D. (2016) Algorithms, governance, and governmentality: On governing academic writing. *Science, Technology, and Human Values* 41(1): 17–49.
- Jarzabkowski, P., and Kaplan, S. (2015) Strategy tools-in-use: A framework for understanding “technologies of rationality” in practice. *Strategic Management Journal* 36(4): 537–558.
- Kellogg, K.C., Valentine, M.A., and Christin, A. (2020) Algorithms at work: The new contested terrain of control. *Academy of Management Annals* 14(1): 366–410.
- Kitchin, R. (2014) Big Data, new epistemologies and paradigm shifts. *Big Data and Society* 1(1): 1–12.
- Langley, A. (1999) Strategies for theorizing from process data. *Academy of Management Review* 24(4): 691–710.
- Leonelli, S. (2016) *Data-Centric Biology*. University of Chicago Press.
- Levina, N., Vaast, E. (2005) The emergence of boundary spanning competence in practice: Implications for implementation and use of information systems. *MIS Quarterly* 29(2): 335–363.
- MacIntyre, A. (1981) *After Virtue: A Study in Moral Theory*. Duckworth.
- Majchrzak, A., More, P.H., and Faraj, S. (2012) Transcending knowledge differences in cross-functional teams. *Organization Science* 23(4): 951–970.
- March, J.G. (2006) Rationality, foolishness, and adaptive intelligence. *Strategic Management Journal* 27(3): 201–214.
- March, J.G., and Olsen, J.P. (1975) The uncertainty of the past: Organizational learning under ambiguity. *European Journal of Political Research* 3: 147–171.
- March, J.G., and Simon, H.A. (1958) *Organizations*. Wiley.
- Nicolini D (2012) *Practice Theory, Work, and Organization: An Introduction*. Oxford University Press.
- Pachidi, S., Berends, H., Faraj, S., and Huysman, M. (2021) Make way for the algorithms: Symbolic actions and change in a regime of knowing. *Organization Science* 32(1): 18–41.
- Pachidi, S., and Huysman, M. (2017) Organizational intelligence in the digital age: Analytics and the cycle of choice. In Galliers, R.D. and Stein, M.K. (eds), *The Routledge Companion to Management Information Systems* (pp. 391–402). Routledge.
- Pettigrew, A.M. (1990) Longitudinal field research on change: Theory and practice. *Organization Science* 1(3): 267–292.
- Porter, T.M. (1996) *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press.

- Simon, H.A. (1976) *Administrative Behavior. A Study of Decision-Making Processes in Administrative Organization*, 3rd edn. Free Press, Collier Macmillan Publishers.
- Spradley, J. (1980) *Participant Observation*. Holt, Rinehart & Winston.
- Van den Broek, E., Sergeeva, A., and Huysman, M. (2021) When the machine meets the expert: An ethnography of developing AI for hiring. *MIS Quarterly* 45(3): 1557–1580.
- Waardenburg, L., and Huysman, M. (2022) From coexistence to co-creation: Blurring boundaries in the age of AI. *Information and Organization* 32(4): 100432.
- Waardenburg, L., Huysman, M., and Sergeeva, A.V. (2022) In the land of the blind, the one-eyed man is king: Knowledge brokerage in the age of learning algorithms. *Organization Science* 33(1): 59–82.
- Weiss, R.S. (1995) *Learning from Strangers*. Free Press.
- Yanow, D., Tsoukas, H. (2009) What is reflection-in-action? A phenomenological account. *Journal of Management Studies* 46(8): 1339–1364.

11. Constructing actionable insights: the missing link between data, artificial intelligence, and organizational decision-making

Arisa Shollo and Robert D. Galliers

INTRODUCTION

It goes without saying that the growth in available data across the globe is exploding (e.g., Shirer and Rydning, 2020), resulting in the global data-intensive technologies market valued at US\$271.83 billion in 2022 and projected to grow to US\$745.15 billion by 2030 (Fortune Business Insights, 2023). Organizations are striving to “democratize” data and make it available to employees for better decision-making and improved business performance (Davenport, 2010; Chen et al., 2012; Torres et al., 2018). While such technologies are rapidly transforming, many organizations are trying to make sense of and create value from their data, but only a few manage to reap the promised benefits. This applies to the whole range of data-intensive technologies from the simpler (for example, business intelligence and descriptive analytics) to the more complex ones (for example, machine learning, neural networks and prescriptive analytics). In a 2021 survey of 85 Fortune 1000 companies, only 24 percent indicate that their organization is data-driven (Bean, 2021), and researchers in general find that leveraging business intelligence and analytics (BI&A) and artificial intelligence (AI) is no trivial task (Brynjolfsson et al., 2018; Tarafdar et al., 2019; Shollo et al., 2022). Over the last decade, academic and practitioner-oriented discourses have focused primarily on the possibilities that AI and BI&A offer (Coombs et al., 2020; Davenport and Ronanki, 2018; Shrestha et al., 2019), but fall short in uncovering the orchestration practices that take place for the promised outcomes to materialize. Consequently, organizations and researchers alike are increasingly focusing on the human component of realizing value from data (Shollo and Galliers, 2016; Günther et al., 2017; Torres and Sidorova, 2019; Shollo et al., 2022), with sociomaterial theorizing becoming more prominent (e.g., Stein et al., 2014). Thus, the discussion is slowly changing from how to become data-driven, to a greater focus on becoming insight-driven (Davenport et al., 2019).

Researchers have long debated how best to use data-intensive technologies for improving results throughout the data transformation process (Sharma et al., 2014), resulting in two primary perspectives on AI and BI&A. What might be termed the traditional perspective emphasizes the technological capabilities of AI and BI&A systems as drivers of better knowledge output (Rowley, 2007; Olszak, 2016;

Torres and Sidorova, 2019), implying that leveraging insights is a more or less straightforward task. Conversely, the practice perspective emphasizes the importance of people—such as data analysts (Shollo and Galliers, 2016), data scientists and domain experts (Grønsund and Aanestad, 2020; Joshi et al., 2021; Ghasemaghaei and Turel, 2021; Shollo et al., 2022; van den Broek et al., 2021)—as well as the processes associated with interpreting AI and BI&A output for actionable insight (Grønsund and Aanestad, 2020; Shollo and Galliers, 2016). However, while there is a plethora of studies focusing on how to use data-driven technologies to generate insights (Abbasi et al., 2016; Chen et al., 2012; Kitchens et al., 2018), there is less coverage of what enables this AI and BI&A output to be turned into actionable decisions (Sharma et al., 2014).

Recently, the topic of actionable insight has surfaced as a concept which aids understanding of outcomes that facilitate action and turn out to be insightful (Dykes, 2016; Torres and Sidorova, 2019). This concept calls on data analysts and data scientists not only to use data-intensive technologies merely to generate insights, but also to mobilize data, technologies, and people for actionable decision-making: for knowing (cf. Orlikowski, 2002) in other words, leading to doing. Thus, there are a growing number of calls for research on how and when insight becomes actionable, as well as the challenges and efforts required throughout the data transformation process that might inhibit or encourage their creation (Sharma et al., 2014; Günther et al., 2017). Understanding the concept of actionable insight, and elaborating on how organizations might implement such insight, could hold the key for organizations to leverage their data to become not only data-driven, but also insight-driven (cf. Davenport et al., 2019). In light of the above, this chapter attempts to answer the following research question: How might organizations facilitate actionable insight through data-intensive technologies?

To guide the empirical investigation of this research question, we begin by providing some background on data-driven decision-making processes, particularly focusing on how actionable insights have been treated in the scholarly literature thus far.

BACKGROUND

Making better, more informed decisions within organizations is at the core of the debate around such data-intensive technologies as AI and BI&A (Wamba et al., 2015). Researchers appear to accept that AI and BI&A can create value in organizations by enabling superior decision-making processes, which in turn would lead to improved performance (Davenport, 2013; Sharma et al., 2014). The practice by which managerial decisions are based on insights from data analysis rather than intuition or opinion is known as data-driven decision-making (DDD) (Provost and Fawcett, 2013). Proponents of DDD often refer to the “objectivity” of the data, the “strength of large numbers” and the “computational power” of data-intensive technologies as a way to reduce human biases in decision-making processes (Brynjolfsson et al., 2011; Chen et al., 2012; McAfee et al., 2012). Consequently, or so the argument

goes, better decisions enabled by the use of analytics have the potential to optimize or transform tasks, business processes, and business models (Brynjolfsson and McAfee, 2017). For example, LaValle et al. (2011) attempt to show how top-performing organizations use analytic insight to inform both strategic and operational decisions. Likewise, Brynjolfsson et al. (2011) demonstrate the benefits of DDD practices. Their findings indicate a 4–6 percent increase in productivity for companies that have adopted and practice DDD. Similarly, Chen et al. (2012, pp. 1166–1168) suggest that big data analytics can help organizations to “better understand ... business and markets and make timely business decisions.” More recent studies argue that AI technologies form a new, economically important general-purpose technology with vast implications for output and welfare gains (Brynjolfsson et al., 2018; Brynjolfsson et al., 2023).

The enthusiasm arising from the perceived potential of the massive amounts of data that have become available for analysis has contributed to the diffusion of a DDD approach in managerial decision-making (Barton and Court, 2012; McAfee et al., 2012; Brynjolfsson and McElheran, 2016). Yet, these calls for DDD often imply that data use is a relatively straightforward process. As such, they fail to acknowledge the different ways in which practitioners make sense of data to inform their decisions and actions (Galliers et al., 2017; Martin and Golsby-Smith, 2017). The suggestion is that multiple forms of data are first simply turned into information via analysis (using data-intensive technologies), and then combined with managerial judgment and expertise to create actionable insights.

As managerial judgment is a human process of cognition, decision-makers are, however, often prone to cognitive biases in using simplistic heuristics (Tversky and Kahneman, 1974) based on the information available in a particular context. This can lead to low decision quality. The “heuristics and biases” school of thought has continuously made the point that human decision-makers can be faulty decision-makers because of their limited predictive abilities. Hence, DDD is portrayed as superior to human judgment alone (Grove and Meehl, 1996; Grove et al., 2000). Nevertheless, others argue that new biases can appear when subjective decisions are taken during the development of DDD artifacts (Suchman, 2002; Lycett, 2013; Boyd and Crawford, 2012; Kitchin, 2014).

From Data to Value

Sharma et al. (2014) identify three stages of the DDD process: Data to Insight, Insight to Decision, and Decision to Value. This process was earlier called the “information value chain” (Koutsoukis and Mitra, 2003; Abbasi et al., 2016). In the following, we use the DDD value creation process of Sharma et al. (2014) to illustrate the phases through which organizations can create value through data-intensive technologies.

The first, Data to Insight, stage includes the core insight generation process. Much of the research on AI and big data analytics focuses on this stage, specifically on technologies and techniques for storing, integrating, computing, and analyzing data from various sources (Wamba et al., 2015). The outputs of these technologies are

often (automated) prediction scores based on past data or (automated) prescriptions based on the predictions. Big data analytics is defined as “statistical modeling of large, diverse, and dynamic data sets of user-generated content and digital traces” (Müller et al., 2016, p. 289), combining descriptive, diagnostic, predictive, and prescriptive analytics methods to obtain answers. With AI technologies, there are two broad types: symbolic AI rule-based technologies and connectionist AI technologies, based on machine learning (ML) (Chollet, 2019; Legg and Hutter, 2007). Since the resurgence of ML in the 1990s and the rise of “deep learning” in the 2010s, ML has increasingly become the dominant approach to building AI systems (Berente et al., 2021; Haenlein and Kaplan, 2019). Nonetheless, the rule-based approach still plays an important role and is used pervasively in industry. Understanding the insight generation process is important for understanding how the use of AI and BI&A leads to improved performance.

In the second stage, Insights to Decisions, insights about customer trends, supplier performance, or competitor behavior are just pieces of information that need to be embedded into strategic and operational decisions to generate value (Lycett, 2013; Sharma et al., 2014). Previous studies have questioned the taken-for-granted assumption that big data analytics leads to better decisions, while noting the psychological, contextual, and organizational factors that might influence this relationship (LaValle et al., 2011; Davenport, 2013; Sharma et al., 2014; Shollo and Galliers, 2016; Thiess and Müller, 2018).

Sharma et al. (2014), point out that there are many options on how to act upon an insight; some less obvious than others. The process of developing and evaluating these options impacts the quality of the decisions. In particular, human biases and satisficing behavior due to bounded rationality (cf. Simon, 1960) might negatively influence the search for options as well as their assessment. Notwithstanding, Frisk et al. (2014) show how big data analytics can help organizations to develop new creative options.

To enable the conversion from insights to actionable decisions, LaValle et al. (2011) and Thiess and Müller (2018) suggest embedding analytics in operational business processes and daily workflows, rather than producing standardized reports that are often not read and are therefore not acted upon. Organizations are distinctive, however, and embedding requires the understanding and definition of their distinctive business processes. Thus, decision-making processes and organizational attention have a strong influence on successfully converting insights into decisions. For example, Shollo and Galliers (2016) found that, even in instances where analytics are part of daily non-automated workflows, decision-makers act on insights only if there is an associated management focus. In their case, when management focus shifted, decision-makers ceased to act on these insights. Given the alternative views expressed in this debate, it can be argued that further research is needed to identify the process and conditions under which insights lead to decision quality improvements.

As noted above, the third stage concerns Decisions to Value. In order for decisions to generate value, successful implementation is required. This is rarely guaranteed, however. One contributing factor is decision acceptance, which influences

decision-makers' commitment in following through on their decisions (Sharma et al., 2014). Researchers have suggested that AI and BI&A have the capability of transforming decision execution by allowing enhanced visibility of firm operations and improved performance measurement mechanisms (McAfee et al., 2012; Habjan et al., 2014; Huang et al., 2014; Shollo and Galliers, 2016; Kitchens et al., 2018). In particular, Habjan et al. (2014) and Shollo and Galliers (2016) illustrate how big data analytics increased decision acceptance due to the enhanced visibility of operations and the decision model, while Huang et al. (2014) and Kitchens et al. (2018) propose that the use of analytics combined with structural changes enables operational agility.

At the same time, however, advanced AI technologies such as ML and deep learning might lessen visibility, as it is less easy to understand or explain how the algorithm produces the predictions, and which data points matter in the production of those insights (Gerlings et al., 2021). Additionally, even if all the steps are taken well (that is, the right questions are asked, new insights are generated, leading to good decisions being made, which are successfully implemented), it is still uncertain whether actual value will result, particularly for strategic decisions (Sharma et al., 2014).

Decisions and Computational Analysis

Certainly, not all decisions can be made under the auspices of computational analysis. Simon (1960) was the first to distinguish between programmed and non-programmed decisions. Programmed decisions are well structured: the procedure concerning how to reach the solution is knowable, with predefined instructions to be followed. These decisions are characteristically repetitive and follow similar patterns. Conversely, non-programmed decisions are not planned; there is no set procedure and no fixed pattern is followed. These decisions are generally one-time decisions that are often complex and can have long-term impact. Programmed decisions are more adequately addressed by computational analysis and automation. Thus, arguably, non-programmed decisions can only partly be addressed by computational analysis, with the output of this analysis being used as input into decision-making.

In their seminal article, "A framework for management information systems," Gorry and Scott-Morton (1971) argue that "information systems should exist only to support decisions." Their perspective is thus on organizational decision-making, characterizing managerial activity in terms of these decisions. In particular, they distinguish between the type of the managerial activity that a decision is associated with (operational, tactical, and strategic; based on Anthony, 1965), and the kind of decision itself (programmed and unprogrammed; based on Simon, 1960). In this way they create nine categories of the purposes and problems of information systems activity. They suggest that decisions which are structured and belong to operational activities are more easily (and likely to be) automated, in comparison with more strategic and non-programmed decisions. Thus, it is argued that when it comes to strategic decision-making, a computational approach is adequate only for a small subset of strategic decisions: that is, when the ends and the means are

known (Thompson, 1965/2003). This means that, for a specific decision task, an objective can be specified and a procedure for how to achieve the objective exists. In an AI context, Shrestha et al. (2019) make a similar argument by investigating those conditions under which a decision should be made by AI, or by a human, or in a hybrid human–AI configuration. They also argue that decision tasks can be delegated to AI systems in decision-making scenarios where: the decision objective is well structured; the accuracy of the prediction is more important than interpretability of the decision-making process; the set of options is large; decision-making speed is critical; and replicability of decision outcomes is desirable. Under these conditions, they argue that AI can generate actionable insights, as the AI system will generate insights and act upon these based on the decision objective.

Actors in the Data-Driven Decision-Making Process

As organizations are increasingly adopting DDD processes, new actors are emerging to facilitate value creation. In the last decade or so, the role of the data scientist has emerged in organizations and has increased considerably (Davenport and Patil, 2022). The 650 percent growth rate in data scientists' jobs since 2012 is an indication of this trend (Columbus, 2017). As a new profession, data scientists have attracted the attention of both academics and practitioners. Numerous popular science and academic articles (Davenport and Patil, 2012) seek to understand and define the data scientist role in organizations: "the people who understand how to fish out answers to important business questions from today's tsunami of unstructured information ... think of him or her as a hybrid of data hacker, analyst, communicator, and trusted adviser." The required combination of technical and communicative skills is also noted by Debortoli et al. (2014) and Van der Aalst (2014).

Thousands of data scientists are already working in organizations, and recent empirical studies demonstrate their key role in creating value from AI and BI&A (Shollo et al., 2022; Grønsund and Aanestad, 2020). Most of the studies undertaken thus far have been focused on the skills of data scientists and their education and background (Davenport and Patil, 2012; Van der Aalst, 2014; Debortoli et al., 2014). Further, these studies (except for Debortoli et al., 2014) tend to be conceptual, highlighting the need for data scientists; and prescriptive, focusing on what their role should be in organizations and how they should go about addressing business problems.

Another emerging role that is commonly referred to is that of domain experts. As the title implies, domain experts tend to be business people who have experience in a particular domain (for example, inventory management, credit risk, anti-money laundering, clinical trials). Empirical studies show that these domain experts work closely with data scientists, and provide their business knowledge in defining business problems and creating computational solutions to them. For example, Grønsund and Aanestad (2020) show how domain experts are involved in auditing practices, and work closely in evaluating and monitoring these systems in organizations. Van den Broek et al. (2021) go a step further in an attempt to cement this collaboration by

demonstrating how the situated knowledge of domain experts is critical in training data scientists, in order to augment the development of learning algorithms during their deployment in organizations.

Notwithstanding this recent research effort, empirical studies on how organizational actors navigate the complex organizational decision-making processes, or even transform them to turn the insights they produce into good decisions, remain scarce. The question remains: how do they make such insights actionable? What is their involvement in creating options, evaluating these options, and securing commitment around the insights and the final, agreed option?

Against this background, this chapter aims to further extend our knowledge on how actionable insights are produced by reflecting on a case study in a pharmaceutical organization. To illustrate the human endeavor of creating actionable insights, we chose a well-structured operational decision where, according to extant theory, computational analysis would be the appropriate approach to address the decision and its implementation (that is, the creation of actionable insights), given that it should be reasonably straightforward with a clear procedure to follow.

METHOD

The rest of this chapter builds on a case study of creating actionable insights from data, using data-intensive technologies. The aim of the study was to understand how the organization went about creating actionable insights, who was involved, and what processes rendered insights “actionable.” The case study was part of a larger project on how the case organization, “Pharma,” became a more insight-driven organization. We conducted 16 interviews with central actors (four developers/business analysts, three of whom were interviewed twice; six users; and three managers). We also collected information from the company’s website, including organizational charts, and the vision and mission as communicated to stakeholders. The aim of the data collection was to document the processes through which the organization created actionable insights. Rather than deducing that actionable insights are a default outcome when using data-intensive technologies, we allowed ourselves to be surprised by the empirical world. As a result, this case study provides a specific and well-documented example of the processes through which actionable insights are generated in Pharma; an example of what may also characterize actionable insights generation processes in other organizations, without predicting that they will.

THE PHARMA CASE

Pharma is a European pharmaceutical company, employing more than 6000 people across the globe. Pharma is no stranger to large amounts of data, which are amassed across the various departments (for example, Finance, Supply Chain, Production, Sales, and Marketing). Much of the data are stored in a SAP Business Warehouse

system, and are subsequently accessed for input into a number of analytical tools. Multiple departments have created dedicated BI&A functions with a view to leveraging their data for managerial insight and decision-making. For example, the Finance department has established a Center for Analytics (CFA), looking to support the commercial part of Pharma with better access to and use of its data through reports and tools built in BI&A software. The Integrated Business Planning department likewise has dedicated BI&A developers working on reports and tools to improve supply chain efforts and decisions, but without it being a fully fledged center of excellence like the CFA. While each of these initiatives is born out of individual departments, the solutions they build are made available to and are expected to be used across many departments and functions in Pharma. This has so far resulted in multiple BI&A reports spanning both classic finance metrics (for example, gross sales, net sales, and commercial profit and loss), supply and inventory reports for optimizing stock situations (for example, the Inventory Coverage Tool), and cross-departmental reports which combine data from multiple parts of the organization (for example, the Sales and Operations dashboard).

The ambition is to have BI&A reports penetrate departmental silos and achieve widespread adoption for improved data-driven insight and decision-making throughout Pharma. However, developers, managers, and users alike indicate that this is no easy task. Developers find that it is difficult to move users away from their old legacy systems and customized Excel files, even if they are time-consuming and ineffective. The users themselves find that some of the BI&A reports do not suit their needs, due to technical difficulties and “analytical insight” that is not exactly as they would like. Managers find that the underlying source data in Pharma’s systems are still inconsistent, of poor quality, and that reports do not always provide sufficient insight for decision-making. However, the same employees also find that the BI&A reports, given two years of hard work, are now better than they were at the beginning. Source data is improving, reports are getting better, and employees are starting to see the value of using BI&A in their work.

Pharma is thus still in the middle of its data-driven journey, and although many challenges have emerged while working with these technologies, some have been resolved. The organization is slowly but surely leveraging the potential of the BI&A technology. Below, we analyze how Pharma works with data-intensive technologies in pursuit of actionable insights.

ANALYSIS OF THE PHARMA CASE

The Commitment to Actionable Insights: A Platform Perspective

We observed that BI&A is used for facilitating the translation of raw data into an insight output, with employees highlighting how people and processes are interwoven and impacting the BI&A processes. What was especially interesting to us was that the generation of actionable insights from BI&A was viewed as a platform,

rather than as a system where the development and the consumption of BI&A reports are integrated. Looking at BI&A as a platform allowed the organizational members to better understand the process of generating actionable insight, by integrating BI&A as a tool as well as a process, where different actors played a critical role in the development of actionable insights.

On the one hand, Pharma's BI&A systems are being used in the traditional sense: to transform raw, unstructured data input into a structured knowledge output. Thus, BI&A might save time on analytical tasks by automating data refreshes and model updates, or provide access to data that were previously hard to come by. Indeed, there seemed to be both a desire among users to use BI&A for providing "knowledge and information on the spot" (Senior Manager), that can be used to "make decisions or take action faster" (Head of Supply), and an equal desire from managers and developers to provide it: "BI in a nutshell is this whole ETL process to be. You find data, transform and manipulate it, load it into an output. It could be Power BI or even Excel. That's the definition to me. To make data actionable" (Senior Business Analyst). This emphasizes how the key purpose for BI&A, according to managers and users in Pharma, was to provide actionable insight.

On the other hand, the organization is also aware of the more human components in the equation, acknowledging that the output of BI&A needs to go through a human actor in order to become an action: "You want to take a lot of data and tell a story with it. Based on your story and the data, the users of that report are enabled to interpret and make decisions" (Developer). The human actor is central to the process of extracting insight from BI&A, in the sense that the BI&A system might deliver insights, but not insights that any developer or user can find and take advantage of. Rather, the developer or user must leverage their own business expertise and analytical skills to explore what the BI&A reports have to offer, interpret what they see, and extract insights:

I think over time, I kind of adapt it to the style here to give them an action plan instead of giving them possibilities and a lot of abstract metrics. If I provide them with a concrete action, then they are happy to just get it. But if I use something a bit more abstract like probability they will have questions about what we should do. (Senior Developer)

This awareness among the developers that the human actors are central to the generation of actionable insights has resulted in greater focus being given to the involvement of end-users and stakeholders in the BI&A development process. They are being involved in an iterative development process where they participate in scoping and feedback meetings (Senior Business Analyst), but also in joint community sessions where the BI&A reports and processes are discussed:

Once a month we have a community session ... There are participants from the Financial Performance team who govern the market guidelines. Some from Financial Management Information who own the data used by the market. We gather those who are making the reports, those who use them, those who make the rules and those who provide the data. (Business Analyst)

This implies a departure from what might be termed the traditional view given to the BI&A systems: the insight output and the users of the systems and the developers do not exist independently from each other; rather, they coexist in a platform- and community-like entity, where developers, managers, and users all participate in iterative processes of BI&A report development and consumption. Likewise, the insight output of BI&A is thus a combination of technical transformation and structuring of data, together with the consumption of BI&A reports by the users utilizing their expertise.

The Different Actors that Participate in the Generation of Actionable Insights

Developers: producers and trainers

Developers play a large role in shaping the data and the insight that is provided to the information consumers. They describe their role in the BI&A platform as “providing insight across the business” (Senior Business Analyst). This implies that the developers are more than just information providers, since they must have sufficient business understanding to shape the insight that the users need: “I always have the business problem in mind. Always. Else we won’t make solutions that are made for the users” (Developer and Business Analyst). In addition, the developers tap into the collaborator task of translating analytical results, as they help to train users not only on how to use the technical features of BI&A, but also on how to use specific reports in the users’ specific areas of work: “What really works is when I go train smaller groups. Then we can discuss how to apply the reports in their specific work” (Business Analyst).

Users: consumers and co-producers

Users are those who leverage the BI&A reports for better decision-making themselves. In terms of business functions in Pharma, the users span positions such as Finance Business Partners and more operational employees with only peripheral contact with the platform. They are employees who have direct influence on decisions regarding sales estimates, forecasts, production, and supply chain planning. Historically, they have performed most of their analyses by manually extracting data from SAP into tools such as Excel, or by using some of the few standardized reports directly in the SAP system. Today, they are expected to perform the same analyses using the BI&A reports made by the developers, for increased quality and efficiency. As these users possess excellent knowledge about business processes, they fit the Information Consumer category well. However, like the developers, the argument can be made that users are not just consumers. Users who reach a certain technical and analytical level of expertise are seen coming up with new ideas and asking for new features and reports to be developed, thereby helping to innovate the BI&A platform. They add context to BI&A insights, are typically users of multiple BI&A reports at one time, and continuously ask for new data sources to be included, as explained by the Finance and Analytics Business Partner for instance. Furthermore, as the Head of

Supply reported, they are even included in the development process by the developers, and assist in shaping the BI&A reports to meet their particular needs.

Managers: production and consumption facilitators and consumers

The Director of Corporate Planning and Reporting explained that managers are employees in supervising roles who help to align decisions and operational efforts with the overall organizational strategy. In terms of their role and involvement with the BI&A platform, they also help to ensure a strategic link, often requesting new reports and assisting the developers in scoping and aligning them with business objectives (as illustrated by the Director of Product Costing). Managers also often act as facilitators by pointing to the right experts (both technical and business), allowing developers and users to go through the necessary iterations, and providing context on strategic directives from top management (as depicted by the Director of Integrated Business Planning during our study). These characteristics again span a range of information worker archetypes. Their contributions include innovating, scoping, and steering the BI&A platform through new report ideas, assisting other users in decision-making by helping to raise good questions, and providing business context. Additionally, though, they are sometimes just business users who simply extract numbers or insight for monthly reporting to higher management, or to make other decisions within their own department themselves.

Actionable Insights for Well-Structured, Operational Problems

One report in Pharma's BI&A portfolio that provides actionable insights for management concerns a well-structured, well-recognized issue in commercial organizations: the Inventory Coverage Tool. This report is an inventory tracking tool "allowing employees to view past, current and planned stock on all products across markets at stock locations" (Developer, Integrated Business Planning). One page of the report that provides prescriptive insight is the Stock Out List, which highlights stock locations that are already completely empty of certain products. "Each 'stock out' is supplemented by comments made by the users responsible for the products and locations, providing further contextual information" (Senior Manager). This could be a comment indicating that a listed stock out is just a data error (that is, not a true stock out), or it could be a confirmation that a particular location has indeed run completely dry of a product. This gives users short-term actionable insight to solve stock out issues across the supply chain. This combination of insight and qualitative contextualization allows for immediate action. Another page yielding more long-term actionable insight is the Trench Analysis, which allows users "to see the stock deliveries planned for the coming months" (Developer, Integrated Business Planning). The page includes user-friendly color coding to highlight where stock will be high or low were the current plan to be followed; it also includes the same comment functionality as the Stock Out List for qualitative contextualization, enabling the user to take immediate mitigating action, "to avoid over/under-stocking products on location" (Senior Manager).

One of the major challenges of generating actionable insight that we identified was the anchoring of reports and insight in business processes. A challenge which Pharma manages employs agile development ideas, in which reports are quickly pushed to the consumers for concrete feedback and suggestions: “We realized that an iterative process is crucial ... It needs a certain quality, all must be correct. But the presentation and visuals, it’s okay to launch a minimum viable product we can learn from and iterate on” (Business Analyst). In addition, the iterative development involves multiple scoping and feedback meetings involving consumers and other stakeholders in the process: “We gather people in meetings. Encourage them to have different opinions ... Here it came to light that ratios were very important for the report. That was something I had not thought about at all” (Business Analyst). In general, this requires great effort from the developers in terms of not only acting as information producers but also taking on a greater collaborative project management responsibility: bringing stakeholders together and connecting the data to business problems in collaboration with both managers and end-users. When done well, this results in reports which are better aligned with business needs, at both a strategic and a managerial level, but which also give operational employees the insight they need (Business Analyst).

This development effort by the developers is concurrently found to be supported by an equivalent enablement effort, to overcome the identified challenges in terms of both adoption and leveraging report functionality. The developers at Pharma find that users need training in using the BI&A software to understand the general technical features of BI&A reports: “Before I started these trainings, many didn’t really get it ... That you can bookmark, reset. It’s just the whole way of working with a new tool” (Developer). In addition to the general training, the developers identified a need for specific training showing users how to use reports for specific business processes, and hence facilitating the business anchoring of BI&A reports:

What really works is when I train smaller groups. Then we can discuss how to use the reports in their specific use case. “On workday 16 we need to sign off on this forecast vs. last month”. Well then you need to put your filters like this and this to get the data you need. (Developer, Integrated Business Planning)

An important lesson from these training sessions has been that the users need to “get their hands on” the tools to improve their general analytical capabilities. Simply seeing somebody else using the reports is not enough: “The users must get their hands dirty. Not just see it. They must do it. It’s very important ... It makes a huge difference if you show that there is a drop down here, or you make them use the drop-down” (Senior Business Analyst). This further solidifies how the developers may act as information collaborators, as they, in this case, act as trainers and facilitators who help to upskill information consumers.

Actionable Insight for Ill-Structured Problems

We observed that when it comes to more ill-structured problems and non-programmed decisions, the organization used Investigatory Actionable Insights. This kind of insight might not lead to immediate action or a decision, but enables the generation of deeper and better questions for further analysis. One example is managers using reports to raise questions and ask for explanations from operational employees: “I have daily and weekly contact with everybody. If they see something, they are good at warning me. And I look a lot at data and ask questions too. Sometimes I see things they haven’t.” (Director, Corporate Planning and Reporting). There are even examples of upper management at Vice President level using the reports for inquiring about data that appear odd in some way: “I’ve seen our VP taking screenshots of our tables of the stock situation, maybe he filtered on a specific site, showing that we had too much stock the last 8 months. Then he sends it out and asks, ‘what are we gonna do about this?’” (Developer, Integrated Business Planning). In a similar vein, many regular users utilize BI&A reports to inform recurring business tasks such as Monthly Landing Estimates, or Demand Reviews, which are subsequently used to retrospectively analyze and optimize operations: “For me it’s the starting and end point of the demand review process. We look at it to say where are we now with our latest plan version versus target? That is the start of market demand discussions” (Business Partner). This type of inquiry-inspiring insight, which provides a starting point for discussion and further analysis, is not about enabling immediate organizational action, but employees still perceive this type of insight as actionable. However, it is also pointed out that this type of insight only provides value if the questions eventually lead to more concrete action in the form of decisions or behavioral changes: “The questions you ask must cause a behavioral change or something else to happen. Otherwise, it just creates noise” (Business Analyst).

DISCUSSION AND REFLECTIONS

The Pharma case shows that constructing actionable insights through data-intensive technologies, even for well-structured operational problems, requires continuous interaction between developers, users, and managers. According to extant theory, however, well-structured operational problems would be characterized as programmed or programmable decisions (Simon, 1960; Gorry and Scott-Morton, 1971) that are ripe for being dealt with by a computational approach (Thompson, 1965/2003), and easily automated and executed by an algorithm. AI technologies are supposed to take over the decision-making and associated tasks of well-structured operational decisions, such as inventory management. Nonetheless, even in this instance, we demonstrate how creating actionable insights and overcoming the challenges associated with such work is a continuous task, not a one-time endeavor. So, the question arises as to why there should be a need for continuous interaction

between different stakeholders and BI&A data and insights in the process of constructing actionable insights.

As succinctly pointed out by Dobbe et al. (2021, p. 2), “the question is no longer what computers can or cannot do, but how to structure computation in ways that support human values and concerns.” In this case, the human value that the effort is directed to is “Action.” As such, actionable insights are an emergent property that arises from these interactions among both technical elements (such as data) or intelligent agents (predictive models) and human agents, processes, and the supporting infrastructure (Dobbe et al., 2021).

This continuous interaction focuses on understanding the insights (produced through data-intensive technologies) by interpreting them (cf. Walsham, 1994) from different perspectives (for example, those of developers, users, or managers), contextualizing them within the business domain, and then reconceptualizing at a group level with a view to understanding and agreeing how the insights are to be actioned. These ongoing processes of conceptualizing, subjectifying, and contextualizing insights underline the emergent situated nature of actionable insights. This is because, for humans to act on insights, they need to be convinced of their legitimacy; they must believe and accept that they are indeed “insights” and be able to justify those insights and the actions to be taken. Justification plays a crucial role in qualifying the veracity of data and the conclusions arising. It is partly in light of this that the recent new field of explainable AI has emerged as a response to the AI “black box” issue (Gerlings et al., 2021). Justifications are achieved through iterative situated interactions between the actors that allow them an understanding and acceptance of insight provenance. Developers acting as information collaborators attempt to facilitate an iterative process of understanding stakeholder needs and input, and connecting analytics to business problems for improved insight. Users are continuously attempting to improve their analytical capability for deriving more out of the insight through the training that is facilitated by the developers. The resulting analytics community efforts on the part of the participating information workers further help in supporting this ongoing process, facilitating the creation of insight that is fit for action. These interactions between information workers and the BI&A platform during report development and consumption also indicate the blurring boundaries between development and consumption (Waardenburg and Huysman, 2022).

Organizing for Making Decisions with AI

AI applications are insight-producing machines that sometimes may themselves act on the output they produce. The Pharma case indicates that if data-intensive technologies are to be leveraged, work needs to be organized differently.

First, when making decisions with AI, organizations need to ensure ongoing feedback channels for stakeholders to assign appropriate meaning to AI-enabled, data-driven insights and possible actions. Authentic participatory approaches are more relevant than ever, given the need to make sense of complex insight generation systems (Bødker and Kyng, 2018). Otherwise, organizations might struggle

when facing unexpected change (Pachidi et al., 2021) or unintended consequences (Stice-Lusvardi et al., 2023). Managers, users, developers, and other relevant stakeholders should form units taking care of the decisions made with AI; understanding its limitations and being in control of their actions in order to do so, as appropriate. Second, until now, we have talked about AI/ML or analytics projects as though they have a clear beginning and end. Yet, as the Pharma case illustrates, working with data-intensive technologies and ensuring that actionable insights are produced is a continuous, collective endeavor; one that is close to the notion of organizational knowing (Orlikowski, 2002), where the focus is as much on the ongoing processes and actions as the outcome. Thus, we contend that we need a different vocabulary from that which we currently use in describing work with AI machines. AI projects, AI implementation, AI use, are outmoded concepts based on a focus on the tool (the technologies) and a static consideration of information systems (IS). Working with AI in the current era implies a blurring of the boundaries between development, implementation, and use (Waardenburg and Huysman, 2022). Third, continuous interaction is a characteristic of making decisions with AI, even when working and making decisions with off-the-shelf AI systems; where the user has little insight into the “backstage” of the AI system, the data being input, and lacks training in its use and application. This is evident, for example, when working with large language models (LLMs) where, in order to gain meaningful and actionable insight, one has to enter into a discussion with the chatbot and learn the art of prompting (for example, specifying conditions).

It is apparent, therefore, even from this limited case study, dealing as it does with a relatively simple, programmed decision (that is, inventory management), that future studies are needed, focusing on new theorizations of the work behind the construction of actionable insights as well as on the new ways of working. While our findings are but a first step towards understanding the intricacies of the generation processes of actionable insights in organizations, future studies can build on these foundations to provide additional understanding and insight. There is clearly a necessary qualitative element in this future agenda (cf. Simeonova and Galliers, 2023). At this stage, we simply ask for caution in overly extolling the virtues of artificial intelligence, wishing rather that AI might more appropriately come to stand for “actionable insights.”

REFERENCES

- Abbasi, A., S. Sarker, and R.H.L. Chiang (2016). “Big data research in information systems: toward an inclusive research agenda.” *Journal of the Association for Information Systems*, 17(2), 1–32.
- Anthony, R. (1965). *Planning and Control Systems: A Framework for Analysis*. Graduate School of Business Administration, Harvard University, Boston.
- Barton, D., and D. Court (2012). “Making advanced analytics work for you.” *Harvard Business Review*, 90(10), 78–83.
- Bean, R. (2021). “Why is it so hard to become a data-driven company?” *Harvard Business Review Digital Articles* (2021), 1–5.

- Berente, N., B. Gu, J. Recker, and R. Santhanam (2021). "Managing artificial intelligence." *MIS Quarterly*, 45(3), 1433–1450.
- Boyd, D., and K. Crawford (2012). "Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon." *Information, Communication and Society*, 15(5), 662–679.
- Bødker, S., and Kyng, M. (2018). "Participatory design that matters—facing the big issues." *ACM Transactions on Computer–Human Interaction (TOCHI)*, 25(1), 1–31.
- Brynjolfsson, E., L.M. Hitt, and H.H. Kim (2011). "Strength in numbers: how does data-driven decisionmaking affect firm performance?" *SSRN Electronic Journal*. Available at SSRN: <http://ssrn.com/abstract=1819486>.
- Brynjolfsson, E., D. Li, and Raymond, L.R. (2023). *Generative AI at Work* (No. w31161). National Bureau of Economic Research.
- Brynjolfsson, E., and A. McAfee (2017). "The business of artificial intelligence: what it can—and cannot—do for your organization." *Harvard Business Review Digital Articles*, 7, 3–11.
- Brynjolfsson, E., and K. McElheran (2016). "The rapid adoption of data-driven decision-making." *American Economic Review*, 106(5), 133–139.
- Brynjolfsson, E., Rock, D., and Syverson, C. (2018). "Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics." In Ajay Agrawal, Joshua Gans and Avi Goldfarb (eds), *The Economics of Artificial Intelligence: An Agenda* (pp. 23–57). University of Chicago Press.
- Chen, H., R.H.L. Chiang, and V.C. Storey (2012). "Business intelligence and analytics: from big data to big impact." *MIS Quarterly*, 36(4), 1165–1188. <http://www.jstor.org/stable/41703503>.
- Chollet, F. (2019). *On the Measure of Intelligence*. ArXiv:1911.01547 [Cs]. <http://arxiv.org/abs/1911.01547>.
- Columbus, L. (2017). "LinkedIn's fastest-growing jobs today are in data science and machine learning." *Forbes*, December 11. <https://www.forbes.com/sites/louiscolumbus/2017/12/11/linkedins-fastest-growing-jobs-today-are-in-data-science-machine-learning/?sh=6f2383a451bd>.
- Coombs, C., D. Hislop, S.K. Taneva, and S. Barnard (2020). "The strategic impacts of intelligent automation for knowledge and service work: an interdisciplinary review." *J. Strateg. Inf. Syst.*, 29(4), 101600.
- Davenport, T.H. (2010). "Business intelligence and organizational decisions." *International Journal of Business Intelligence Research*, 1(1), 1–12. <https://doi.org/10.4018/jbir.2010071701>.
- Davenport, T.H. (2013). "Analytics 3.0." *Harvard Business Review*, 91(12), 64.
- Davenport, T.H., E. Brynjolfsson, A. McAfee, and H.J. Wilson (2019). *Artificial Intelligence: The Insights You Need from Harvard Business Review*. Harvard Business Press.
- Davenport, T.H., and D.J. Patil (2012). "Data scientist: the sexiest job of the 21st century." *Harvard Business Review*, 90(5), 70–76.
- Davenport, T.H., and D. Patil (2022). "Is data scientist still the sexiest job of the 21st century?" *Harvard Business Review*, 90, July 15. <https://hbr.org/2022/07/is-data-scientist-still-the-sexiest-job-of-the-21st-century>.
- Davenport, T.H., and R. Ronanki (2018, January 1). "Artificial intelligence for the real world." *Harvard Business Review*, January–February. <https://hbr.org/2018/01/artificial-intelligence-for-the-real-world>.
- Debortoli, S., O. Müller, and J. vom Brocke (2014). "Comparing business intelligence and big data skills." *Business and Information Systems Engineering*, 6(5), 289–300.
- Dobbe, R., T.K. Gilbert, and Y. Mintz (2021). "Hard choices in artificial intelligence." *Artificial Intelligence*, 300, 103555.

- Dykes, B. (2016). "Actionable insights: the missing link between data and business value." *Forbes*, 1–9. <https://www.forbes.com/sites/brentdykes/2016/04/26/actionable-insights-the-missinglink-between-data-and-business-value/#47b8359f51e5>.
- Fortune Business Insights (2023). *Big Data Analytics Market Report*. <https://www.fortunebusinessinsights.com/big-data-analytics-market-106179> (accessed December 9, 2023).
- Frisk, J.E., R. Lindgren, and L. Mathiassen (2014). "Design matters for decision makers: discovering IT investment alternatives." *European Journal of Information Systems*, 23(4), 442–461.
- Galliers, R.D., S. Newell, G. Shanks, and H. Topi (2017). "Datification and its human, organizational and societal effects." *Journal of Strategic Information Systems*, 26(3), 185–190.
- Gerlings, J., A. Shollo, and I.D. Constantiou (2021). "Reviewing the need for explainable artificial intelligence (xAI)." In *54th Annual Hawaii International Conference on System Sciences, HICSS 2021* (pp. 1284–1293).
- Ghasemaghaei, M., and O. Turel (2021). "Possible negative effects of big data on decision quality in firms: the role of knowledge hiding behaviours." *Information Systems Journal*, 31(2), 268–293.
- Gorry, G., and M. Scott-Morton (1971). "A framework for management information systems." *Sloan Management Review*, 13(1), 56–79.
- Grove, W.M., and P.E. Meehl (1996). "Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: the clinical–statistical controversy." *Psychology, Public Policy, and Law*, 2(2), 293.
- Grove, W.M., D.H. Zald, B.S. Lebow, B.E. Snitz, and C. Nelson. (2000). "Clinical versus mechanical prediction: a meta-analysis." *Psychological Assessment*, 12(1), 19.
- Grønsund, T., and M. Aanestad (2020). "Augmenting the algorithm: emerging human-in-the-loop work configurations." *Journal of Strategic Information Systems*, 29(2), 101614.
- Günther, W.A., M.H. Rezazade Mehrizi, M. Huysman, and F. Feldberg (2017). "Debating big data: a literature review on realizing value from big data." *Journal of Strategic Information Systems*, 26(3), 191–209. <https://doi.org/10.1016/j.jsis.2017.07.003>.
- Habjan, A., C. Andriopoulos, and M. Gotsi (2014). "The role of GPS-enabled information in transforming operational decision making: an exploratory study." *European Journal of Information Systems*, 23(4), 481–502.
- Haenlein, M., and A. Kaplan (2019). "A brief history of artificial intelligence: on the past, present, and future of artificial intelligence." *California Management Review*, 61(4), 5–14.
- Huang, P.-Y., S.L. Pan, and T.H. Ouyang (2014). "Developing information processing capability for operational agility: implications from a Chinese manufacturer." *European Journal of Information Systems*, 23(4), 462–480.
- Joshi, M.P., N. Su, R.D. Austin, and A.K. Sundaram (2021, March 2). "Why So Many Data Science Projects Fail to Deliver." *MIT Sloan Management Review*. <https://sloanreview.mit.edu/article/why-so-many-data-science-projects-fail-to-deliver/>.
- Kitchens, B., D. Dobolyi, J. Li, and A. Abbasi (2018). "Advanced customer analytics: strategic value through integration of relationship-oriented big data." *Journal of Management Information Systems*, 35(2), 540–574.
- Kitchin, R. (2014). "Big Data, new epistemologies and paradigm shifts." *Big Data and Society*, 1(1), 2053951714528481.
- Koutsoukis, N.-S., and G. Mitra (2003). *Decision Modelling and Information Systems: The Information Value Chain* (Vol. 26). Springer Science & Business Media.
- LaValle, S., E. Lesser, R. Shockley, M.S. Hopkins, and N. Kruschwitz (2011). "Big data, analytics and the path from insights to value." *MIT Sloan Management Review*, 52(2), 21.
- Legg, S., and M. Hutter (2007). "Universal intelligence: a definition of machine intelligence." *Mind. Mach.* 17(4), 391–444. <https://doi.org/10.1007/s11023-007-9079-x>.

- Lycett, M. (2013). “‘Datafication’: making sense of (big) data in a complex world.” *European Journal of Information Systems*, 22(4), 381–386.
- Martin, R.L., and T. Golsby-Smith (2017). “Management is much more than a science.” *Harvard Business Review*, 95(4), 128–135.
- McAfee, A., E. Brynjolfsson, T.H. Davenport, D.J. Patil, and D. Barton (2012). “Big data: the management revolution.” *Harvard Business Review*, 90(10), 60–68.
- Müller, O., I. Junglas, J. vom Brocke, and S. Debortoli (2016). “Utilizing big data analytics for information systems research: challenges, promises and guidelines.” *European Journal of Information Systems*, 25(4), 289–302.
- Olszak, C.M. (2016). “Toward better understanding and use of business intelligence in organizations.” *Information Systems Management*, 33(2), 105–123.
- Orlikowski, W.J. (2002). “Knowing in practice: enacting a collective capability in distributed organizing.” *Organization Science*, 13(3), 249–273.
- Pachidi, S., H. Berends, S. Faraj, and M. Huysman (2021). “Make way for the algorithms: symbolic actions and change in a regime of knowing.” *Organization Science*, 32(1), 18–41.
- Provost, F., and T. Fawcett. (2013). “Data science and its relationship to big data and data-driven decision making.” *Big Data*, 1(1), 51–59.
- Rowley, J. (2007). “The wisdom hierarchy: representations of the DIKW hierarchy.” *Journal of Information Science*, 33(2), 163–180. <https://doi.org/10.1177/0165551506070706>.
- Sharma, R., S. Mithas, and A. Kankanhalli (2014). “Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organisations.” *European Journal of Information Systems*, 23(4), 433–441. <https://doi.org/10.1057/ejis.2014.17>.
- Shirer, M., and J. Rydning (2020). “IDC’s Global DataSphere Forecast shows continued steady growth in the creation and consumption of data.” *International Data Corporation (IDC)*. <https://www.idc.com/getdoc.jsp?containerId=prUS46286020>.
- Shollo, A., and R.D. Galliers (2016). “Towards an understanding of the role of business intelligence systems in organisational knowing.” *Information Systems Journal*, 26(4), 339–367. <https://doi.org/10.1111/isj.12071>.
- Shollo, A., K. Hopf, T. Thiess, and O. Müller (2022). “Shifting ML value creation mechanisms: a process model of ML value creation.” *Journal of Strategic Information Systems*, 31(3), 101734.
- Shrestha, Y.R., S.M. Ben-Menahem, and G. Von Krogh (2019). “Organizational decision-making structures in the age of artificial intelligence.” *California Management Review*, 61(4), 66–83.
- Simeonova, B., and R.D. Galliers (eds) (2023). *Cambridge Handbook of Qualitative Digital Research*. Cambridge University Press.
- Simon, H.A. (1960). *The New Science of Management Decision*. Harper & Row.
- Stein, M.K., S. Newell, E.L. Wagner, and R.D. Galliers (2014). “Felt quality of sociomaterial relations: introducing emotions into sociomaterial theorizing.” *Information and Organization*, 24(3), 156–175.
- Stice-Lusvardi, R., P.J. Hinds, and M. Valentine (2023). *Legitimizing Illegitimate Practices: How Data Analysts Compromised Their Standards to Promote Quantification*. Available at SSRN 4321298.
- Suchman, L. (2002). “Located accountabilities in technology production.” *Scandinavian Journal of Information Systems*, 14(2), 7.
- Tarafdar, M., C.M. Beath, and J.W. Ross (2019). “Using AI to enhance business operations.” *MIT Sloan Management Review*, 60(4).
- Thiess, T., and O. Müller (2018). “Towards design principles for data-driven decision making—an action design research project in the maritime industry.” In: *ECIS Proceedings. Portsmouth: European Conference on Information Systems (ECIS)*.

- Thompson, J.D. (1965/2003). *Organizations in Action: Social Science Bases of Administrative Theory*. Transaction Publishers.
- Torres, R., and A. Sidorova (2019). "Reconceptualizing information quality as effective use in the context of business intelligence and analytics." *International Journal of Information Management*, 49(May), 316–329. <https://doi.org/10.1016/j.ijinfomgt.2019.05.028>.
- Torres, R., A. Sidorova, and M.C. Jones (2018). "Enabling firm performance through business intelligence and analytics: a dynamic capabilities perspective." *Information and Management*, 55(7), 822–839. <https://doi.org/10.1016/j.im.2018.03.010>.
- Tversky, A., and D. Kahneman. (1974). "Judgment under uncertainty: heuristics and biases." *Science*, 185(4157), 1124–1131.
- van den Broek, E., A. Sergeeva, and V.M. Huysman (2021). "When the machine meets the expert: an ethnography of developing AI for hiring." *MIS Quarterly*, 45(3), 1557–1580.
- van der Aalst, W.M.P. (2014). "Data scientist: the engineer of the future." In Mertins, K., Bénaben, F., Poler, R., and Bourrières, J.-P. (eds), *Enterprise Interoperability VI* (pp. 13–26). Springer.
- Waardenburg, L., and M. Huysman (2022). "From coexistence to co-creation: blurring boundaries in the age of AI." *Information and Organization*, 32(4), 100432.
- Walsham, G. (1994). "Geoff Walsham: interpreting information systems in organizations." *Organization Studies*, 15(6), 937–937.
- Wamba, S.F., S. Akter, A. Edwards, G. Chopin, and D. Gnanzou (2015). "How 'big data' can make big impact: findings from a systematic review and a longitudinal case study." *International Journal of Production Economics*, 165, 234–246.

12. It takes a village: the ecology of explaining AI

Lauren Waardenburg and Attila Márton

RIGHT PLACE, RIGHT TIME

November 26, 2018, 7:20 p.m. A gray car flees through the city and the police follow. CityPol officers Robert and Max¹ were on car patrol duty, following orders to keep an eye on a specific parking garage in the city center. The predictive policing AI system had forecasted a car burglary in that area that evening. Noticing a gray car exiting the parking garage in an unusual rush, Robert flashed the stop sign. But the driver ignored it, instead hitting the accelerator across a red traffic light during peak-hour traffic, turning a routine patrol into a full-on chase.

At first glance, this sequence of events seems relatively straightforward. However, it presents only a fraction of the decision-making process involved in the police arriving at the right place at the right time. In fact, this episode is part of a larger series of events that unfolded over nearly 11 months, observed by Lauren in her intensive, 31-month ethnographic fieldwork at the police headquarters of a large Dutch city (which we call “CityPol”). This series, and what it means for making decisions with artificial intelligence (AI), is what we analyze in this chapter. To do this, we first go back to where the story started.

ELEVEN MONTHS EARLIER

January 8, 2018. CityPol’s department managers² have their weekly meeting to decide how to allocate police resources. For this, they draw on a predictive policing AI system that has been in use for almost two years. Its main aim is to help police managers to determine the types of crimes that are likely to become an issue and are therefore in need of attention. Tailor-made by data scientists hired by the Dutch police, the system is based on a machine learning model to predict “hotspots”—including the time (in four-hour intervals) and area (in 125m² squares)—where a crime will likely take place one week in advance. The decision to develop this tool in-house was driven by widespread societal critique against the United States version PredPol, which was reported to intentionally profile individuals.³ In an effort to stay away from such outcomes, the Dutch police decided to maintain control over the development of their algorithm and not to include individual-level data and variables. The result is a relatively simple machine learning model based on logistic regression that uses aggregate demographic and geographic data (for example, total number of

addresses, average house prices, number of male and female inhabitants, average age of inhabitants) and crime-specific data (for example, time since last similar crime, number of similar crimes in the last two weeks).

The AI system is intended as a tool to stimulate preventive action for the police to come up with strategies and actions to deter predicted crimes (Waardenburg et al., 2019). Often, such preventive measures mean increasing police patrols, and thus police visibility, to discourage criminal activities. Its application is generally considered a success by CityPol's managers, as the rates of nearly all crimes it was applied to (for example, disorderly conduct, house burglary, robbery) decreased by up to 50 percent compared to the years before. At CityPol, the only major exception is car burglaries, of which the incidence rate has not declined, despite increasing police patrols. As a consequence, the AI system continues to produce car burglary predictions, prompting the police managers to focus their key resources on this single crime category.

For the police to look towards an AI system to make better decisions is not surprising, given the increased societal belief in the power of AI to solve organizational problems (e.g., Daugherty and Wilson, 2018; Davenport, 2018). Of course, this tendency to seek help in making decisions is not new, since throughout history humanity has turned towards some form of higher, transcendental authority, including religion, logic, and ideology, to provide guidance. With the widespread emergence of AI, the hope is that we have found the 21st century solution to improve our decision-making processes, by which insights do not rely on religious revelation, logical reasoning, or ideological adherence, but on technological computation. The world, thus, becomes a computational problem to be solved by machine learning algorithms that autonomously find patterns between large numbers of data points. This may lead to insights beyond what humans can achieve on their own and, by extension, allow for better decisions (Agarwal and Dhar, 2014; Agrawal et al., 2018; Davenport and Kirby, 2016; Domingos, 2015; Ford, 2018; Leavitt et al., 2020).

The move towards AI systems to improve decision-making by processing more and more data represents a long-standing, yet flawed, assumption that decisions are nothing but choices amongst a limited number of options. A decision, however, is a much more complex accomplishment because one could always have decided otherwise (Luhmann, 2000). This, in turn, invites others to question why one decided one way and not another. Hence, at the core of every decision lies a paradox: what makes a decision into a decision is that it cannot be decided (Sweeting, 2021). Or, to put it the other way around: "only those questions that are in principle undecidable, we can decide" (Foerster, 2003, p. 293).

Formal organizations, as social systems, have evolved to deal with this decision paradox by providing the means to cover it up; be it through authority, remuneration, tradition, duty, or other institutions (Luhmann, 2000; Kallinikos et al., 2013). Think, for instance, of the multiple responses available when one questions why a certain organizational decision was made: "because the boss (or the consultant) says so," "because I pay you to do what I say," or "because this is how we have always been doing things." To be sure, none of these responses actually answers the question as

to why a certain decision was made and not another; they, rather, bring an end to an otherwise infinite regress of “why” questions (such as, “why was the decision to make you the boss and not somebody else?”, or “why was the decision to have a boss and not something else?”).

Indeed, turning towards an ultimate, transcendental authority serves this very same purpose of covering up the decision paradox by, for instance, referring to religion (“because it is God’s will”), reason (“because it is the logical thing to do”), or ideology (“because it is what capitalism dictates”). Drawing on digital technology, in particular AI, then just adds another option of “because the machine says so.” It is in those terms that AI supports organizing; not by providing some new kind of objective decisions, as is the commonly held belief (Ackoff, 1967), but by covering up the paradox that whenever we make a decision, we also communicate that we could have decided otherwise (Luhmann, 2000).

If we agree that decisions are paradoxical and that covering this up is an organizational necessity, why is it then important to return to the paradox when it comes to AI? As we demonstrate in this chapter, it is important because AI, with its advanced computational capabilities to process large amounts of data, comes with the illusion that decisions can finally be turned into problems that can be mathematically solved (Alaimo and Kallinikos, 2021). However, a simple mathematical problem (that is, with only one correct solution) is not a decision. Or, to put it the other way around, if a question could be solved mathematically, it would not require a decision in the first place (Smith, 2019). What, then, happens when a tool that is intended as a cover-up is believed to make decisions?

WHAT’S IN THE BOX?

April 24, 2018. The pressure is increasing on CityPol, as they just figured out that for this week’s predicted hotspot, there are 32 more registrations of car burglaries than this time last year. For over three months, back-office intelligence officers have been processing the output of the predictive policing AI system, sharing with police officers where and when to go. In turn, the police officers, patrolling the indicated hotspots, have been recording the time and place of every car burglary in the police database, thus feeding new data back into the AI system to improve its predictive model. Yet, all these efforts have not resulted in a decrease in the number of car burglaries, nor in any arrests. As a consequence, the police officers realize that just following the algorithm wherever it leads them is not enough. Rather, to be able to act, the police need to know why the AI system is producing these predictions. Since such explanations are not available, they struggle to find the best strategy to catch the burglars or prevent the car burglaries from happening in the first place.

For the police, to ask why the algorithm sends them to certain areas at certain times does not mean questioning the efficacy of the algorithm—that is, they are not asking why the algorithm decided that one area is a future hotspot and not another—but is an expression of their doubts about their own understanding of why the car burglaries

continue to happen. Addressing this practical need involves trying to understand how the AI system views the world (Lebovitz et al., 2021). This is not a philosophical endeavor to determine whether these kinds of AI systems are intelligent or not (Bridle, 2022; Bostrom, 2014; Smith, 2019), but a matter of cognition and, in particular, explainability (Hafermalz and Huysman, 2022; Vassilakopoulou et al., 2022).

Employing technological tools to make better decisions comes with the promise that humanity, as the designer of those tools, will always be able to understand, at least in principle, how they work and therefore explain their outcomes. However, this assumption is contested in the case of current AI systems that use machine learning algorithms (Smith, 2019). These systems “create their own models of reality” (Hafermalz and Huysman, 2022, p. 12), and thus can become “opaque” (Burrell, 2016; Christin, 2020) or “black-boxed” (Anthony, 2021; Pasquale, 2015). That is to say, although AI systems are a tool of our own design, it is getting more and more difficult to understand how their machine learning algorithms find patterns in data and, consequently, to explain the reasoning behind their outputs (Faraj et al., 2018; Glaser et al., 2021; Alaimo and Kallinikos, 2021). Ironically, our efforts to explain the world by means of computation have resulted in new problems in explaining the computations themselves.

What makes the increasing inability to explain AI particularly relevant for organizational decision-making is that, while it may have been (and in certain domains still is) appropriate to cover up the decision paradox by simply accepting the opacity of a higher transcendental authority (be it God, reason, or ideology), “because the algorithm says so” does not seem to be enough to make AI systems work in practice. In the case of CityPol, their need for explainability is not because of philosophical, ethical, legal, and/or political concerns—as is typically the focus of attention—but of practical relevance in their efforts to clamp down on car burglaries. To put it more generally, explaining AI can become a necessary practice for people to do their job and, thus, for an organization to perform (Anthony, 2021; Hafermalz and Huysman, 2022; Lebovitz et al., 2022; Vassilakopoulou et al., 2022; Waardenburg et al., 2022).

The observation that explaining AI may be a practical necessity for organizational performance further complicates the so-called “explainable AI” debate (Burrell, 2016; Hafermalz and Huysman, 2022). To be sure, not all scholars agree that AI systems can, in principle, be explained. Some argue that because AI systems will inevitably become black boxes (e.g., Ananny and Crawford, 2018; Christin, 2020; Grimmelmann and Westreich, 2017; Introna, 2016), there is an inherent incompatibility between human and machine reasoning that cannot be overcome (e.g., Bostrom, 2014; Burrell, 2016). This incompatibility gives reasons for concern with regard to the accountability (e.g., Burke, 2019; Burton et al., 2020; Schulzke, 2013; Von Krogh, 2018) and trustworthiness (e.g., Bader and Kaiser, 2019; Christin, 2017; Glikson and Woolley, 2020; Lebovitz et al., 2021; Robbins, 2019) of AI-based decision-making that may even escalate into a threat for democracy (Christin, 2020).

Others argue that AI can be explained, but disagree on how explainability is to be accomplished (Hafermalz and Huysman, 2022). For one, the discourse about the ethics and regulation of AI typically emphasizes the societal implications of remain-

ing in the dark about machine reasoning (e.g., Coeckelbergh, 2020; Dignum, 2019). The proposed remedy is to provide transparency about the data and algorithm used, which can generate explanations suitable to those involved with the system. The responsibility for those explanations, in turn, lies with the computer scientists creating these systems. Making AI systems thus understandable for non-AI experts then also allows policymakers to determine who is responsible for the outputs generated by AI.⁴

By contrast, computer scientists place the responsibility of explaining AI mainly on the side of its users. For this purpose, they distinguish between two types of explainable systems (Doran et al., 2017): interpretable systems rely on relatively simple statistical models which allow users, with the requisite expertise, to examine how an algorithm processes input into output; while comprehensible systems, by comparison, typically rely on deep learning and neural networks, which makes their computation inaccessible to direct inquiry. Hence, a comprehensible system needs to, as it were, explain itself, by producing symbols (for example, words or visualizations) that are comprehensible to humans, allowing them to infer an explanation (e.g., Schmid et al., 2016; Van der Maaten and Hinton, 2008; Doran et al., 2017).

Viewed in this purview, CityPol's predictive policing AI system can be categorized as an interpretable system. As it is running a relatively simple regression model, the data scientists who created the model can unpack the mathematical calculations in order to explain how a certain input is processed into a prediction by the algorithm. However, getting the mathematics behind the predictions explained is not practically relevant, let alone helpful, for supporting CityPol in their decision-making to fight car burglaries. The reason, as we explained above, is that decisions are complex accomplishments (unsolvable due to their paradoxical nature) rather than mathematical problems (to be solved by technological computation).⁵ In other words, the explanation that is needed at CityPol is not a lesson in mathematics, but indications about, for example, who might be a suspect, or the modus operandi of the car burglars, so that the police officers do not feel left in the dark about what to pay attention to when patrolling. Deriving an explanation, therefore, means that the CityPol officers have to find indicators that go beyond technological computation.

THINKING OUTSIDE THE BOX

May 29, 2018. For over a month, the police officers have continued to follow the predictions during their patrols, but this time they also reported observations other than the time and place of the burglaries, such as location-specific details. This kind of data work is an important complement to the predictive policing AI system, which merely predicts areas that can include all sorts of locations (for example, open streets, public parking lots), each with a different possible explanation as to why car burglaries happen there. However, as more and more data is being collected, the predictions become more targeted to one specific hotspot, and with the additionally recorded details and observations, the intelligence officers start to see a pattern: the

car burglaries happen mainly in parking garages. This insight helps the police to understand why following the predictions to show increased police presence has not deterred criminals from burglarizing cars: in parking garages, car burglars can hide and operate in dark corners, out of sight of police patrols.

This is a crucial insight: the effectiveness of the AI system depends on crime details that are not included in the algorithm. Knowing that the car burglaries happen in parking garages helps the police to understand that increasing their presence through patrols is of little help to prevent them. In this case, the way to reduce the crime rate is to actually catch the criminals. For that, they need to know who or what they are looking for. Hence, as the police pay more attention to the locations, they realize that most of the predictions revolve around one hotspot containing only one parking garage with only a single entrance. Using this parking garage as the key focus, the intelligence officers then use camera footage and reports of previous car burglaries to find further insights. These lead to an important shift away from the aggregate patterns provided by the AI system to individual indicators. More specifically, they start to look for characteristics of potential suspects. In this case, they find a small, gray car with two passengers that shows up in the footage right around the time cars are burglarized in the parking garage. Finally, the police have found an actionable insight, pointing towards something concrete to look out for when they are patrolling a hotspot predicted by the AI system.

The extent to which CityPol has to go to get from following generic predictions, providing the most likely times and areas of car burglaries, to creating detailed explanations of their antecedents and patterns, shows the amount of work that it can (and often does) require to make AI systems work in practice; work that typically goes unnoticed or is willfully hidden behind the myth of AI as a panacea for organizational woes (Ekbja and Nardi, 2017; Ens and Márton, 2021; Gray and Suri, 2019; Márton and Ekbja, 2021; Waardenburg and Huysman, 2022).⁶ Importantly, this work of arriving at a practically relevant explanation (in the sense that the police can act, for example, by looking out for a particular gray car during their patrol) involves a wide variety of interactions beyond the mere explanation of what is happening “inside” the machine. It involves, for instance, the intelligence officers who keep track of the car burglary predictions, and find patterns in the additional details shared by the police officers, who, based on these predictions, adjust their patrol and collect further information.

Tracing these interactions reveals a wider cognitive system of feedback loops involved in explaining AI. As described above, the most immediate feedback starts with the AI output and its interpretation by the intelligence officers, which then triggers police officers into action. In turn, the police officers report on the car burglaries, based upon which the data scientists retrain the model to come up with better predictions. However, by the same token, the police officers also interact with criminals (by, for instance, recording their crimes and trying to catch them), who in turn interact with car-owners parking their cars in parking garages to be easily burglarized. As a consequence, the incidence rate of car burglaries remains high enough for the department managers to respond by funneling resources into fighting

that particular type of crime, which by extension impacts the data collected and fed into the AI system, closing yet another loop. In other words, the car burglars and their victims (after all, the police depend on both to continue their behaviors to affirm the pattern as predicted by the AI system) are just as much part of the explanation as are the police officers, managers, intelligence officers, and the AI system itself.

This complex pattern of feedback loops, connecting victims, criminals, and police, is of course part of wider patterns connecting city councils (supporting the use of AI for predictive policing), news media (reporting on the rise and fall of crime rates), socio-economic factors (increasing the likelihood for some citizens to turn to crime, and for others to be able to afford a car), city infrastructures (providing roads and parking garages), car manufacturers and the oil industry (marketing individualized transportation), and so on (see Figure 12.1).

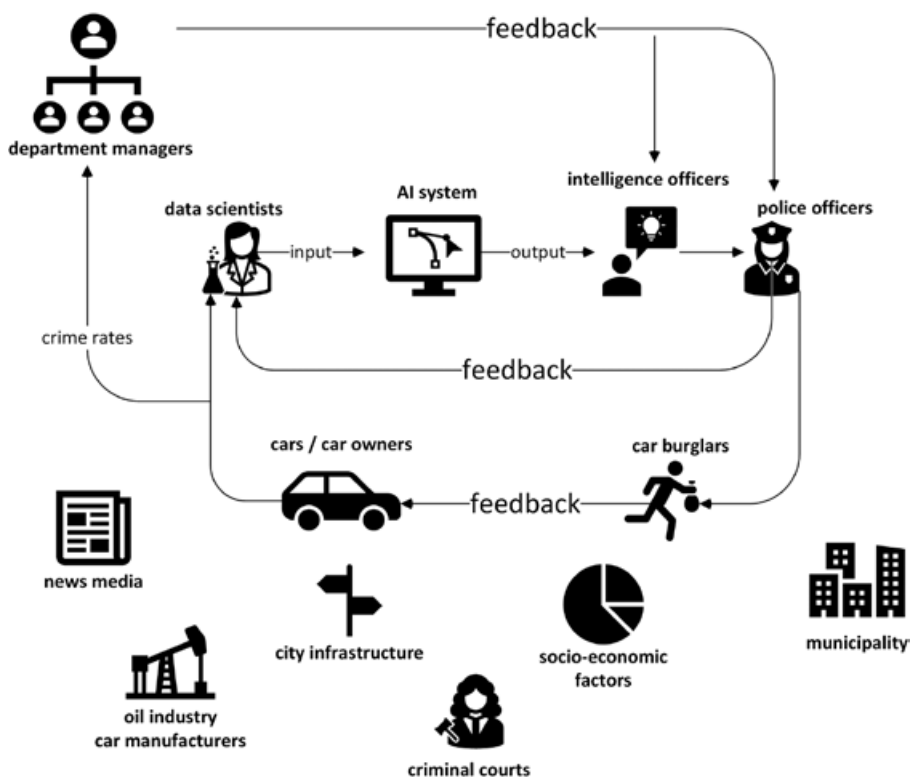


Figure 12.1 *An ecology of explaining AI*

As this pattern demonstrates, while the question of why the algorithm came up with its predictions is simple, the explanation is radically complex, because it defers to an entire ecology of unbounded, open-ended interactions and interdependencies (Márton, 2022). Put simply, it takes a village to explain AI, which makes it impos-

sible to tell where the cognition of the machine ends and that of the police begins. Rather, cognition is established in mutual relationships; it is between the machine, and the police officers, and intelligence officers, and citizens, and city councils, and lawyers, and ministries, and, not to forget, crime victims, suspects, criminals, and on and on with no end (Bateson, 2000; Zundel et al., 2013; Mikołajewska-Zajac et al., 2022). Explaining AI predictions, therefore, requires stepping away from merely trying to open the black box of the machine learning algorithm, into the village or ecology in which the AI system is participating.

WRONG PLACE, WRONG TIME?

Back to November 26, 2018, 11 months after the police initially turned their attention to car burglaries. This evening, there is a prediction for that one particular parking garage. Police officers Robert and Max are on car patrol and drive straight toward the entrance. At that moment, a small gray car, occupied by two people, exits. Robert says that he believes the car is “interesting,” picking up on the clue that the potential car burglars are also driving in a small, gray car. At that moment, the driver hits the accelerator and the police officers speed after it, flashing their stop sign. Ignoring the stop sign, the gray car flees across a red traffic light on a street filled with peak-hour traffic, and the police follow. While maneuvering through the traffic, Max shouts that the suspects are throwing something out of the car. Robert immediately shares this with the control room, so that another police couple can follow up and figure out what it is that was thrown out. Later, this turns out to be a laptop bag. The chase continues and even when the suspects get stuck between the police car and a roadblock, they do not yet give up. They jump out of the car and try to run away. However, they cannot outrun Robert and Max, and are caught and handcuffed after a short chase by foot. In the meantime, the police couple who found the laptop bag went back to the parking garage to find a car with a broken window. Finally, after 11 months of trying to explain the car burglary predictions, it turned out to be about chasing a gray car. It is “over and out” for the car burglars, or so it seems.

If there is one thing that the 11-month trajectory of CityPol shows, it is that making decisions with AI requires making decisions about AI. More specifically, decisions need to be made about what reality the predictions not just reflect, but also create. Explaining those decisions, then, is not just about unpacking the computational methods used for the machine learning model, but about recognizing the ecology of unbounded, open-ended interactions and interdependencies. As a case in point, at CityPol, “computationally speaking,” the AI system could only indicate the estimated time in four-hour blocks and the estimated area in 125m² squares. All other indicators—that is, the focus on parking garages, leading to one specific parking garage and the small gray car with two occupants—were details added to the AI system’s predictions, not produced by it.

However, in the current debate about explainable AI, this part of the explanation, or as we rather argue, the actual explanation, is ignored. As a consequence, the call

for explainability is not a way to uncover which decisions are made in relation to AI, but serves to cover up those decisions that make predictions work in practice. In other words, the current call for computational explanations serves to hide the decision paradox that comes with the practical explanations required to make decisions with AI. What uncovering the computational methods can do is show the scope in which AI systems can contribute to decision-making (for example, pointing at specific areas to allocate police resources). It does not, however, reveal the decision-making process. The problem, then, is not that decisions need to be made about AI to make them work in practice, but that it becomes invisible and, therefore, goes unquestioned that one could have just as easily decided otherwise.

Let us look once more at CityPol to understand the widespread and potentially harmful consequences of overlooking (or ignoring) this cover-up. As described above, the AI system used at CityPol was designed not to include individual-level data or variables, in an attempt to go against profiling critiques. While these technical conditions did not change over time, since the features used for the machine learning model were not adjusted in these 11 months, the explanation that made the predictions useful in practice was individualized to the level of one particular gray car. Interestingly, no one questioned this narrowing down, nor whether it could have been decided otherwise. Instead, everyone seemed to agree that “the machine said so.” But the machine never did.

In the case of CityPol’s use of an AI system to fight car burglaries, narrowing down to one indicator seemed to be the final key to success. And while a success story like the one we just recounted is easy to accept as a blueprint for making the right decisions with AI, there is a flip side to it. A careful reader might have noticed that the suspected car burglars were not caught red-handed. Instead, the chase was instigated by their car resembling what the police had been nudged to look out for and respond to, for months. The question that remains is whether the people who were arrested actually carried out the car burglary, or were just in the wrong place at the wrong time.

WHAT IS THE BOX?

In conclusion, if it becomes normalized to find and target subjective, individualized indicators, not because the machine says so but rather because the machine requires so, we end up conflating the necessary covering-up of the decision paradox with the illusion that AI can make objective decisions. Such conflation may be less problematic when it comes to the color of cars, yet how do we deal with the dangers that present themselves when we do the same with skin colors? If we do not take into account that individualized indicators and, by extension, explanations of AI are ecological, then explainable AI will not lead to more objective, more transparent, more just, more fair systems, but will turn out to maintain and maybe even enhance whatever damaging societal categories already exist. Hence, using an AI system to make decisions about an ecological problem, such as crime, will inevitably result in

the AI system adapting to the ecology, rather than the other way around. Therefore, creating, regulating, and managing an AI system for decision-making requires not only understanding its computations but also, more importantly, the village by which it has been raised, or in other words, its ecology.

NOTES

- 1 Names are pseudonyms. Other identifiers have been anonymized in this chapter.
- 2 Representing emergency response, criminal investigation, human resources, and others.
- 3 https://harvardpress.typepad.com/hup_publicity/2020/06/predictive-policing-and-racial-profiling.html; last accessed July 4, 2023.
- 4 High Level Expert Group on AI Ethics Guidelines for Trustworthy AI: The European Commission (Ala-Pietilä et al. 2019).
- 5 A simple analogy; imagine, we buy four apples in a store, and when asked why we decided to buy four apples, we respond by saying: “because 2 apples + 2 apples make 4 apples.”
- 6 Just consider the work that goes into labeling data used to train models behind applications such as ChatGPT; in particular, the workers who, for a few dollars an hour, are labeling and filtering violent, explicit, or otherwise highly disturbing content, typically outsourced to countries with weak labor protection, and end up psychologically traumatized, suffering from post-traumatic stress disorder (PTSD) (see Business Insider report at <https://bit.ly/3MIWZch>; last accessed July 4, 2023).

REFERENCES

- Ackoff, R.L. (1967). Management misinformation systems. *Management Science*, 14(4), 147–156.
- Agarwal, R., and Dhar, V. (2014). Big data, data science, and analytics: The opportunity and challenge for IS research. *Information Systems Research*, 25(3), 443–448.
- Agrawal, A., Gans, J., and Goldfarb, A. (2018). *Prediction machines: The simple economics of artificial intelligence*. Harvard Business Press.
- Alaimo, C., and Kallinikos, J. (2021). Managing by data: Algorithmic categories and organizing. *Organization Studies*, 42(9), 1385–1407.
- Ala-Pietilä, P. et al. (2019). Ethical Guidelines for Trustworthy AI. High-Level Expert Group on Artificial Intelligence, European Commission. 8 April, <https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html>.
- Ananny, M., and Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media and Society*, 20(3), 973–989.
- Anthony, C. (2021). When knowledge work and analytical technologies collide: The practices and consequences of black boxing algorithmic technologies. *Administrative Science Quarterly*, 66(4), 1173–1212.
- Bader, V., and Kaiser, S. (2019). Algorithmic decision-making? The user interface and its role for human involvement in decisions supported by artificial intelligence. *Organization*, 26(5), 655–672.
- Bateson, G. (2000). *Steps to an ecology of mind*. University of Chicago Press.
- Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.
- Bridle, J. (2022). *Ways of being: Beyond human intelligence*. Penguin UK.

- Burke, A. (2019). Occluded algorithms. *Big Data and Society*, 6(2), 1–15.
- Burrell, J. (2016). How the machine “thinks”: Understanding opacity in machine learning algorithms. *Big Data and Society*, 3(1), 1–12.
- Burton, J.W., Stein, M.K., and Jensen, T.B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2), 220–239.
- Christin, A. (2017). Algorithms in practice: Comparing web journalism and criminal justice. *Big Data and Society*, 4(2). doi:10.1177/2053951717718855.
- Christin, A. (2020). The ethnographer and the algorithm: Beyond the black box. *Theory and Society*, 49(5), 897–918.
- Coeckelbergh, M. (2020). *AI ethics*. MIT Press.
- Daugherty, P., and Wilson, H.J. (2018). *Human + machine: Reimagining work in the age of AI*. Harvard Business Review Press.
- Davenport, T. (2018). *The AI advantage: How to put the artificial intelligence revolution to work*. MIT Press.
- Davenport, T.H., and Kirby, J. (2016). Only humans need apply: Winners and losers in the age of smart machines. Harper Business.
- Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way*. Springer Nature.
- Domingos, P. (2015). *The master algorithm: How the quest for the ultimate learning machine will remake our world*. Basic Books.
- Doran, D., Schulz, S., and Besold, T.R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. arXiv preprint arXiv:1710.00794.
- Ekbis, H., and Nardi, B. (2017). *Heteromation, and other stories of computing and capitalism*. MIT Press.
- Ens, N., and Márton, A. (2021). “Sure, I saw sales, but it consumed me.” From resilience to erosion in the digital hustle economy. *New Media and Society*. doi:10.1177/14614448211054.
- Faraj, S., Pachidi, S., and Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. *Information and Organization*, 28(1), 62–70.
- Foerster, H. von (2003). *Understanding understanding: Essays on cybernetics and cognition*. Springer.
- Ford, M. (2018). *Architects of intelligence: The truth about AI from the people building it*. Packt Publishing.
- Glaser, V.L., Pollock, N., and D’Adderio, L. (2021). The biography of an algorithm: Performing algorithmic technologies in organizations. *Organization Theory*, 2(2). doi:10.1177/26317877211004609.
- Glikson, E., and Woolley, A.W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2), 627–660.
- Gray, M.L., and Suri, S. (2019). *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books.
- Grimmelmann, J., and Westreich, D. (2017). Incomprehensible discrimination. *California Law Review*, 7, 164–177.
- Hafermalz, E., and Huysman, M. (2022). Please explain: Key questions for explainable AI research from an organizational perspective. *Morals and Machines*, 1(2), 10–23.
- Introna, L. (2016). Algorithms, governance, and governmentality: On governing academic writing. *Science, Technology and Human Values*, 41(1), 17–49.
- Kallinikos, J., Hasselbladh, H., and Márton, A. (2013). Governing social practice: Technology and institutional change. *Theory and Society*, 42(2), 395–421.
- Leavitt, K., Schrabram, K., Hariharan, P., and Barnes, C.M. (2020). Ghost in the machine: On organizational theory in the age of machine learning. *Academy of Management Review*, 46(4), 750–777.

- Lebovitz, S., Levina, N., and Lifshitz-Assaf, H. (2021). Is AI ground truth really true? The dangers of training and evaluation AI tools based on experts' know-what. *MIS Quarterly*, 45(3), 1501–1525.
- Lebovitz, S., Lifshitz-Assaf, H., and Levina, N. (2022). To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization Science*, 33(1), 126–148.
- Luhmann, N. (2000). *Organisation und Entscheidung*. Westdeutscher Verlag.
- Márton, A. (2022). Steps toward a digital ecology: Ecological principles for the study of digital ecosystems. *Journal of Information Technology*, 37(3), 250–265.
- Márton, A., and Ekbja, H. (2021). Platforms and the new division of labor between humans and machines. In Mitev, Nathalie, Aroles, Jeremy, Stephenson, Kathleen A., and Malaurent, Julien (eds), *New ways of working: Organizations and organizing in the digital age* (pp. 23–46). Palgrave Macmillan.
- Mikolajewska-Zajac, K., Márton, A., and Zundel, M. (2022). Couchsurfing with Bateson: An ecology of digital platforms. *Organization Studies*, 43(7), 1115–1135.
- Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard Business Press.
- Robbins, S. (2019). A misdirected principle with a catch: Explicability for AI. *Minds and Machines*, 13(1), 1–20.
- Schmid, U., Zeller, C., Besold, T., Tamaddoni-Nezhad, A., and Muggleton, S. (2016). How does predicate invention affect human comprehensibility? In *International Conference on Inductive Logic Programming* (pp. 52–67). Springer.
- Schulzke, M. (2013). Autonomous weapons and distributed responsibility. *Philosophy and Technology*, 26(2), 203–219.
- Smith, B.C. (2019). *The promise of artificial intelligence: Reckoning and judgement*. MIT Press.
- Sweeting, B. (2021). Undeciding the decidable. In: *65th Annual Proceedings for the International Society of the System Sciences*, 65(1).
- Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11), 2579–2605.
- Vassilakopoulou, P., Parmiggiani, E., Shollo, A., and Grisot, M. (2022). Responsible AI: Concepts, critical perspectives and an information systems research agenda. *Scandinavian Journal of Information Systems*, 34(2), 89–112.
- Von Krogh, G. (2018). Artificial intelligence in organizations: New opportunities for phenomenon-based theorizing. *Academy of Management Discoveries*, 4(4), 404–409.
- Waardenburg, L., and Huysman, M. (2022). From coexistence to co-creation: Blurring boundaries in the age of AI. *Information and Organization*, 32(4). doi:10.1016/j.infoandorg.2022.100432.
- Waardenburg, L., Doeleman, R., Melchers, R., and Willems, D. (2019). 3 Misverstanden over predictive policing. *Tijdschrift voor de Politie*, 2019(6/7), 40–43.
- Waardenburg, L., Huysman, M., and Sergeeva, A.V. (2022). In the land of the blind, the one-eyed man is king: Knowledge brokerage in the age of learning algorithms. *Organization Science*, 33(1), 59–82.
- Zundel, M., Holt, R., and Cornelissen, J. (2013). Institutional work in The Wire: An ethnological investigation of flexibility in organizational adaptation. *Journal of Management Inquiry*, 22(1), 102–120.

13. Synthetic stakeholders: engaging the environment in organizational decision-making

Jen Rhymer, Alex Murray, and David Sirmon

Stakeholder theory offers a framework to understand how organizations identify, mediate with, and address the individuals and groups with whom they have relationships (Barney and Harrison, 2020; Freeman, 1984, 2010; Mitchell et al., 1997; Wood et al., 2021). As such, non-human actors do not achieve stakeholder status in a meaningful way except through human representation, often via various collectives (Barnett, 2007; Bosse et al., 2009; Buysse and Verkeke, 2003; McGahan, 2020; Orts and Strudler, 2002). For instance, climate change and its tangible impact on people and the economy necessitates organizational engagement with the natural environment, but currently this only occurs when humans represent the natural environment in organizational decision-making (Bansal, 2003, 2005; Flammer, 2013). Indeed, while the legal field has set precedent to grant the rights of legal personhood to certain natural assets, such as New Zealand's Whanganui River (Boyd, 2018; O'Donnell and Macpherson, 2019), organizations have yet to make this agentic leap, opting instead to rely on performative actions such as corporate social responsibility (CSR); environmental, social, and governance (ESG) investing; and certifications (Bansal and Song, 2017; Gehman and Grimes, 2017).

Yet recent advances among agentic technologies, including artificial intelligence (AI), machine learning (ML) algorithms, and distributed ledger technologies (DLTs) (Murray et al., 2021b), make real the possibility of non-human actors' representatives as organizational stakeholders via agentic technologies. Indeed, any number of non-human actors, including different domains of the natural environment—both current and future—as well as future generations of humans, may provide the basis for what we term synthetic stakeholders. A synthetic stakeholder is a technology-based agent that can learn and act as an independent representative in organizational decision-making processes. Synthetic stakeholders possess agency to consider and argue for their best interests during organizational decision-making. Moreover, synthetic stakeholders, like any human-based stakeholder, can be highly salient as they possess various levels of power, legitimacy, and urgency (Mitchell et al., 1997).

However, synthetic stakeholders are not identical to their human counterparts. They are designed and built as tools, and therefore can vary in their underlying design characteristics. We capture this variation in the concept of design locus, and assert that it ranges from predominately internal to predominately external. We further explicate this range of design locus using a series of illustrative examples. In doing

so, we highlight distinct implications for organizations engaging with synthetic stakeholders based on their locus of design.

This chapter contributes to the stakeholder literature by challenging the exclusivity of human actors—ranging from individual humans to collectives of varying size and form—as stakeholders, and suggesting how non-human actors, such as the natural environment, may be salient stakeholders. We also contribute to the literature on sustainability by proposing a mechanism through which the natural environment may increase its salience to the organizational decision-making processes. While we focus on the application of synthetic stakeholders to specific assets in the natural environment, there are several possible research extensions. Future studies can examine how synthetic stakeholders learn from and engage with multiple organizations (and one another), how organizations can further broaden represented stakeholder groups (for example, future generations of humans), how synthetic shareholders can come to attain formal voting rights and fiscal participation (over and above engaging in organizational decision-making), or how human actors may allow agentic technologies to represent them in the organizational decision-making process.

STAKEHOLDER SALIENCE

The ongoing development of stakeholder theory includes multiple tensions such as who qualifies as a stakeholder, and how to distinguish between various stakeholder groups (Barney and Harrison, 2020; Mitchell et al., 1997; Tantalo and Priem, 2016). Indeed, the ability to identify and differentiate stakeholders allows organizations to better manage and understand the potential impact of specific stakeholders on organizational decision-making. While numerous frameworks have been put forth to distinguish and identify important stakeholder groups, we rely on stakeholder salience to direct our treatment in this chapter. Stakeholder salience is based on the attributes of power, legitimacy, and urgency, and their various combinations (Mitchell et al., 1997). A stakeholder group is said to have power if its members can “bring about the outcomes they desire” (Salancik and Pfeffer, 1974, p. 3). Yet power is also recognized to change over time (Mitchell et al., 1997). Legitimacy, while often intertwined with the notion of power, aligns with the sociological perspective of normativity put forth by Suchman (1995). Here, a stakeholder group is considered legitimate when its actions, in context, are “desirable, proper, or appropriate” (Suchman, 1995, p. 574). The inclusion of urgency captures the combination of time sensitivity and criticality, thereby asserting that stakeholder salience is a dynamic concept which can change with situational factors (Mitchell et al., 1997). Mitchell et al. (1997) detail the characteristics of stakeholders that possess each attribute, and propose that salience will increase with the number of attributes a stakeholder possesses.

Stakeholder salience has been aimed exclusively at human actors. While the natural environment has repeatedly been identified as a potential stakeholder (Driscoll and Starik, 2004; Jacobs, 1997; Laine, 2010; Phillips and Reichart, 2000; Starik, 1995; Stead and Stead, 2000; Stone, 1972), it is often conflated with

human-based collectives (for example, environmental organizations) that represent various elements of the natural environment along with their other interests (Bansal and Roth, 2000; Bansal and Song, 2017; Hoffman, 1999). For instance, Mitchell et al. (1997), in their explication of various types of stakeholders, use the example of the environmentally catastrophic *Exxon Valdez* oil spill, stating how this event added urgency to multiple environmental stakeholder groups which claimed legitimacy, but had little power and therefore had to rely on other groups to gain recognition. This example shows that the natural environment is not the stakeholder, but instead human collectives are the stakeholders. Moreover, these groups suffer various problems in representing the environment: they may be temporary, have conflicting interests, or face resource limitations.

Indeed, efforts to better represent the natural environment in stakeholder theory all suffer from the same basic issue: a collective of humans is the stakeholder rather than the natural environment itself. Moreover, efforts to consider additional attributes such as proximity (Driscoll and Starik, 2004; Norton, 2007), or to add a justice-oriented or fairness approach (Phillips and Reichart, 2000), also all work through a human collective. Thus, while accepting non-human actors as stakeholders is conceptually trackable in stakeholder theory, the ongoing reliance on human actors and/or human-centric organizations to represent the natural environment in decision-making processes is problematic.

THE NATURAL ENVIRONMENT IN ORGANIZATIONAL DECISION-MAKING

Organizations are increasingly recognizing the importance of the natural environment, due to its importance in delivering long-term value to their stakeholders (Driscoll and Starik, 2004; Phillips and Reichart, 2000). We broadly use the natural environment as the focal example of a synthetic stakeholder throughout this chapter, though the concept of synthetic stakeholders is not limited to the natural environment alone (we return to this point in the discussion). Currently, engagement between organizations and the natural environment is generally indirect, meaning that a human actor or collective of human actors is designated to act on its behalf. This indirect representation often takes the form of a sustainability committee, a sustainability officer, or a highly salient external environmental stakeholder group. Common outputs of engagement with these internal and external actors often include CSR efforts such as supply chain greening or ESG investing. However, these efforts are often seen as disconnected from the environment and/or as reputation-oriented (Torelli et al., 2020; Roulet and Touboul, 2015).

Alternatively, some organizations also engage more directly with the natural environment, typically by leveraging third-party groups such as certification agencies or environmental protection groups. Certifications such as Fair Trade, Cruelty Free, and Cradle to Cradle indicate that specific aspects of an organization conform to specific standards. While certifications offer a useful signaling tool, the proliferation

of certifications as well as their narrow focus often creates a performative perception with little substance (Gehman and Grimes, 2017; Roulet and Touboul, 2015). Organizations looking to demonstrate their environmental commitment may also opt to partner with environmental groups. Though this legitimating action suggests a broad alignment between an organization and an environmental group's stated mission, such actions are often subject to politics and self-interest such that specific natural assets may get lost.

However, the representation of specific natural assets is rapidly developing within the legal field (O'Donnell and Arstein-Kerslake, 2021; RiverOfLife et al., 2021). A contemporary, yet critical precedent in the legal field pertaining to stakeholder theory is granting legal personhood to specific natural assets. For example, in 2017, New Zealand granted the rights of legal personhood to the Whanganui River, and the river now has equal and independent recognition in the law, much like any other individual person or group of people (Boyd, 2018; O'Donnell and Macpherson, 2019). This precedent is being adopted more broadly in countries around the world (Page and Pelizzon, 2022).

SYNTHETIC STAKEHOLDERS

The legal personhood of a natural asset provides the basis for a whole new class of stakeholders. While these assets now have the rights of personhood, they still lack a voice of their own to enact their agency. However, advances in agentic technologies provide the potential for voice, unencumbered and uncompromised by human actors. In other words, agentic technologies may facilitate voice that emphasizes and optimizes the natural asset's best interests without the issues of human representation, such as conflicting interests, short-term time horizons, and bounded rationality. Agentic technology-provided voice is a focal component of what we term synthetic stakeholders. More formally, we define a synthetic stakeholder as a technology-based agent that can learn and act as an independent representative in organizational decision-making processes.

A synthetic stakeholder need not rely on any specific enabling technologies, but for purposes of our argumentation in this chapter, we assert that a combination of artificial intelligence (AI) and distributed ledger technology (DLT) are capable of enacting synthetic stakeholders in contemporary organizations given their agency, or their capacity to intentionally constrain, complement, or substitute for human action (Murray et al., 2021b). While prior work in this vein discusses conjoined agency as occurring between humans and non-human ensembles jointly exercising intentionality, herein we suggest that ensembles can consist solely of non-human actors (that is, multiple agentic technologies working in tandem) and still maintain sufficient agency to participate in organizational decision-making.

The primary technology, we suggest, in this ensemble is AI. While AI is a broadly used term, in our argumentation we are specifically referring to technology that possesses the capacity to learn from varied inputs and then act accordingly. This

conceptualization of AI stands in contrast to technology that follows coded rules and instructions. The use of AI in synthetic stakeholders affords these stakeholders “free-will” by granting them the agency to think, plan, and act (Gray and Wegner, 2012; Vanneste and Puranam, 2022). This is crucial in the technological ensemble underlying synthetic stakeholders, and foundational for claims of independent and authentically self-directed representation.

The second technology in the ensemble underlying synthetic stakeholders is DLT. While perhaps less obvious, the use of DLTs is also foundational to the effective development and deployment of synthetic stakeholders because DLTs provide a structured frame to immutably record activities as they occur, and automatically interface with dynamic information sources as they become available (Murray et al., 2021b; Murray et al., 2022). Additionally, the use of a DLT-based frame creates a surveyable bound of the synthetic stakeholder that includes an observable record of inputs and outputs over time. DLTs also provide a means to define a protocol for human intervention, and more broadly allow for ongoing governance that is distinct from the governance of the organization (Murray et al., 2021a). In these ways, DLTs support the independence of synthetic stakeholders from an organization, facilitate synthetic stakeholders’ direct representation of a natural asset, and provide a higher level of transparency and auditability than is currently available for human stakeholders.

While synthetic stakeholders couple legal personhood and agentic technologies to engage in organizational decision-making, they are not all designed with similar objectives. Specifically, there exists variation in how synthetic stakeholders are designed with respect to who participates in their direct development and how others engage during this process, thereby indicating a distinct locus of design for each particular synthetic stakeholder.

LOCUS OF DESIGN

We propose three interrelated design characteristics that collectively indicate a synthetic stakeholder’s locus of design relative to an organization: (1) who participates in the management process; (2) how information is curated; and (3) how synthetic stakeholders engage with others. Each of these characteristics determine whether a synthetic stakeholder’s locus of design is predominantly internal versus external.

The first characteristic is who participates in the management process. When designing a synthetic stakeholder, choices are made regarding the AI learning guidelines and the protocol for intervention. Specifically, we focus on the distinction between participants who are internal to an organization versus formal involvement from external groups. While internal participation can include high-level strategic decision-makers, mid-level managers, and algorithm developers, external participation can include scientists, advocacy groups, or protection agencies who may have quite different views about how a particular asset may behave once enacted as a synthetic stakeholder. While more internal participation may increase an organization’s

confidence that a synthetic stakeholder will address the organization's concerns (the motivation for development), more external participation, particularly amongst those who value the natural environment, may increase legitimacy in terms of the natural asset's authentic representation and have implications for the perceived legitimacy of a synthetic stakeholder.

The next design characteristic relates to information curation, specifically information sourcing and information availability. The sourcing of information considers whether an organization (privately) provides internal knowledge to the synthetic stakeholder, or whether the synthetic stakeholder is limited to information that is commonly known by the average stakeholder. Relatedly, the design characteristic of information availability considers whether the information available to the synthetic stakeholder is curated. At one extreme, the synthetic stakeholder could have unfiltered access to the internet and all the information it contains, which would likely result in the collection and use of misinformation. At the other extreme, it could be limited to certain information through a restricted interface.

The final design characteristic is associated with how the synthetic stakeholder engages with others. This characteristic includes factors such as the output format, degree of visibility to the public, and personification. The ways in which a synthetic stakeholder outputs information can include sharing reports, distributing text or audio dialogue directly with others, or engaging actively in collective stakeholder discussions. Additionally, as a synthetic stakeholder directly interfaces and engages with a broader audience (that is, the public) they may develop a broader understanding of issues and build external alliances. The personification of the synthetic stakeholder allows humans to relate to, trust, and perceive them as having greater agency (Glikson and Woolley, 2020; Waytz et al., 2014). Therefore, an important factor in engagement is whether a synthetic stakeholder is designed with a name, personality, or even a physical form. Overall, this characteristic of engagement is meaningful with respect to a synthetic stakeholder building, maintaining, and leveraging power.

Again, taken together, these design characteristics indicate a collective locus of design ranging from predominately internal to predominately external. We illustrate the variation in locus of design via three fictional examples to provide depth to instances spanning this range. These examples are further summarized in Table 13.1.

HAPPY BEE HONEY: ILLUSTRATING A PREDOMINATELY INTERNAL LOCUS OF DESIGN

The first illustrative case depicts a synthetic stakeholder with an internal locus of design. Jordon is the founder of an artisan honey company. Having learned about the risks facing bees, he began engaging in conservation and activism work, eventually founding a honey company with a mission to provide a natural environment for bees to flourish while also capturing unique and marketable honey. The ability to fulfil this mission depends on providing bees with plenty of space and a healthy botanical habitat. However, with the success of the company and the involvement of

Table 13.1 Locus of design range from predominately internal to predominately external

	Happy Bee Honey (Internal)	SeaHealth (Blended)	Modern Oil (External)
Management process participants	Organizational team with external members providing informal advice	Organizational team and external members have a minority role	Organizational team and external members have a majority role
Information curation	Private internal knowledge made available	Internal knowledge equivalent to general stakeholder	Internal knowledge equivalent to general stakeholder
	Public information moderated	Public information subject to basic filtering	Public information unrestricted
Engagement	Written reports distributed to organization and other stakeholders	Written reports plus direct text dialogue with organization and other stakeholders	Written reports, text dialogue, plus active decision discussion with organization and other stakeholders
	No public engagement	Yes, public is engaged	Yes, public is engaged
		Name and personality	Name, personality, and physical form

increasingly powerful stakeholder groups such as external investors, Jordon worried that the company's emphasis on its original founding mission was being lost. To grant increased power and long-term legitimacy to the bees and their habitat, Jordon established a synthetic stakeholder to represent them.

The development of this synthetic stakeholder was a project which Jordon oversaw personally. The synthetic stakeholder was developed by an internal team, with external experts providing advice without a formal role in the management process. All internal information such as footage of Jordon speaking about the founding mission, the initial development activities related to curating the hives and habitat, and strategic plans, as well as curated public external information on bees and their health, is available to the synthetic stakeholder. Happy Bee Honey opted not to personify the synthetic stakeholder and offered minimal public transparency that primarily consisted of issuing an annual statement. In addition to prepared statements, the synthetic stakeholder interacted in organizational decision-making processes through a simple question-and-answer process. Collectively the choices which Happy Bee Honey made related to the participants included in the management process, curations of information, and means of engagement, suggest that the synthetic stakeholder has an internal locus of design.

The synthetic stakeholder implemented by Happy Bee Honey has learned to represent the needs of an important stakeholder group: the bees and their habitat. In this example, the bees' legitimacy to the organization is well established, based on the organization's founding mission. However, as the company grows, and the founding mission is tempered by other concerns, the limited power of this vital stakeholder group becomes more apparent. As the founder, Jordon can intervene and leverage the synthetic stakeholder as a tool, granting it power and enhancing its legitimacy without

requiring any external engagement. An implication of a synthetic stakeholder with an internal locus of design for organizations is that as power continues to shift and organizational priorities change over time, a synthetic stakeholder can continue to be a voice for a marginalized perspective and persistent mission orientation.

SEAHEALTH: ILLUSTRATING A BLENDED LOCUS OF DESIGN

The second illustrative example demonstrates a blended locus of design. This fictional case features a seaweed product company, SeaHealth. The company sells both edible and beauty seaweed-based products. Both markets are competitive, with a high proportion of consumers who pay attention to material sourcing. Yet seaweed farming practices vary significantly. If done irresponsibly, seaweed farming can negatively affect the environment by reducing biodiversity and water quality. If done responsibly, seaweed farming can be leveraged as part of a sustainable ecological system and provide benefits such as carbon sequestration, water oxygenation, and habitat biodiversification. Due to the wide range of implications, the practices endorsed by seaweed product companies constitute an important strategic choice and point of differentiation. In a market where claims of being eco-friendly are common, yet often unchecked, the use of a synthetic stakeholder to represent an organization's seaweed ecosystem can provide strong support for its claims of valuing the natural environment.

The development of SeaHealth's synthetic stakeholder for differentiation has a locus of design which is neither fully internal nor external, but is instead a blend of both. Like a predominantly internal locus of design, as seen in the prior example, this synthetic stakeholder was created and is managed by the organization. However, internal effort was informed by conversations with external environmental groups, and a minority of participants formally involved in the synthetic stakeholder's development were external to SeaHealth. The internal information that is known to the synthetic stakeholder is similar to that of other engaged stakeholder groups. However, unlike the prior example, this synthetic stakeholder is made highly transparent to other stakeholders and the public at large. A public portal shares the perspective of the synthetic stakeholder regularly, and acts as an interface for the public to submit questions or comments. Furthermore, it is personified and has a dedicated machine used to embody it in SeaHealth's organizational decision-making processes.

The salience of this synthetic stakeholder is increased by the organization granting power to it. Additionally, the choice to use this stakeholder as a tool enhances the legitimacy of the seaweed farm as a stakeholder, as well as the synthetic stakeholder tool as a representative. This example demonstrates how a synthetic stakeholder can be designed to blend external interests with internal investment. An implication for SeaHealth's use of a synthetic stakeholder in this way is an increase in market trust and likely value for the products resulting from its farming operations management.

MODERN OIL: ILLUSTRATING A PREDOMINATELY EXTERNAL LOCUS OF DESIGN

The third illustrative example demonstrates a synthetic stakeholder with an external locus of design. This fictional example begins when a large company, Modern Oil, has an oil spill incident and is in a position of engagement with activist groups. Various motivations may underpin this decision to engage with activist groups, including reputation repair, legally mandated clean-up activities, and potentially genuine interest to address the harm caused. Once the initial mediation plan was implemented and the primary damage contained, the company acknowledged the long-term efforts that would be required to fully address the harm caused by the spill. Modern Oil decided to develop a synthetic stakeholder as a representative for the marshlands directly impacted by the oil spill.

Modern Oil is committed to the development of the synthetic stakeholder tool, but also recognizes the importance that other external parties see the synthetic stakeholder as legitimate and sufficiently powerful in decision-making processes. As such, the development and implementation were managed by a committee formally comprised of members of the organization, local government officials, and activist groups, where internal participants from Modern Oil were a minority. The internal information provided is equivalent to that available to an average stakeholder, while unrestricted access to public information was arranged. Everything about the process and the outcome is transparent to the public. A portal is established that allows for the public to interact with a personification of the synthetic stakeholder. Additionally, this portal serves to provide updated information and enable the submission of new, relevant information. Furthermore, in organizational decision-making processes, the synthetic stakeholder is given a physical form and is capable of physical gestures, thereby allowing for enriched engagement with other stakeholders. Despite its synthetic nature this is a stakeholder that is designed to be an independent voice of the marshlands, with a locus of design that resides, in large part, externally to Modern Oil.

A synthetic stakeholder with an external locus of design is not only granted power by the organization, but is also likely to have enhanced legitimacy in the eyes of other stakeholders. This conferral of legitimacy may be of particular importance in instances where an organization aims to demonstrate its accountability, strives for reputational improvement, or faces contentious decisions that lead to the formation of coalitions. Modern Oil's use of a synthetic stakeholder served as a mechanism to demonstrate progress through continued engagement with the marshland throughout their decision-making processes.

REPRESENTING THE ENVIRONMENT AND MORE

This chapter contributes to the stakeholder literature by challenging the exclusivity of human actors as organizational stakeholders, and suggesting a way in which

non-human stakeholders can assert independence, be salient, and engage with other stakeholders in organizational decision-making. We theorize that the locus of design of synthetic stakeholders has implications for how organizations perceive and engage with these stakeholders. We also contend that the study of synthetic stakeholders is an area with significant opportunity for future scholarly research. Particularly, we assert that the continued development of agentic technologies, and their evolving uses and applications within organizations, extends to stakeholder theory by challenging the notion that non-human stakeholder groups can be dismissed (Barney and Harrison, 2020; Orts and Strudler, 2002). This idea constitutes an exciting prospect, one which we hope will generate renewed discussion about who and what constitutes an organizational stakeholder.

We also speak to the literature on sustainability by proposing a mechanism through which the natural environment and specific natural assets within it may directly engage in organizational decision-making processes, specifically as stakeholders with high salience. Numerous scholars have asserted the importance of the natural environment in organizations' strategic decisions (Driscoll and Starik, 2004; Phillips and Reichart, 2000; Starik, 1995; Stone, 1972), though we are unaware of any theorized mechanism that does not rely on humans to serve as an ongoing representative for such natural interests. The use of synthetic stakeholders enables authentic representation without the continued involvement of humans to voice their interests. In this way, the use of synthetic stakeholders to represent natural assets is apt to reduce bias and political posturing, and also to create the potential for different (less prominent) or multiple assets in the natural environment to engage in organizational decision-making.

While we focus on the application of synthetic stakeholders to specific assets in the natural environment, we also see several avenues for extension. One such extension is broadening the stakeholder groups that could be represented using synthetic stakeholders. For example, organizations could create a synthetic stakeholder to represent the interests of future generations in present decision-making. While the current climate crisis is urgent for people today (Thunberg, 2022), the decisions that organizations make today will continue to have implications long into the future, and thereby impact the lives of future generations. As such, there may be instances of long-term strategic decisions where granting representation to future generations is beneficial to the organization.

Future work could examine when organizations choose to grant representation to future generations as a stakeholder. For instance, it may be prudent for organizations to grant representation to this stakeholder in industries that are known to produce significant negative externalities on the environment (for example, transportation, fashion, agriculture, and so on). Granting representation to future generations in such industries may serve to quell the vocal concerns of activist groups who otherwise act antagonistically toward organizations in these industries. Moreover, granting representation to future generations could also provide a more substantive way than certifications, and other avenues available in the present, to indicate that an organization is considering the long-term ramifications of its actions.

Scholars can also theorize on how organizations can most effectively integrate future generations as a stakeholder. Specifically, scholars could consider the relative effectiveness of granting representation to this stakeholder group via a predominantly internal versus a predominantly external design locus. On the one hand, designing this stakeholder with a predominantly internal locus of design could allow an organization to more easily intervene if the interests of future generations are too much at odds with those of the organization's present mission; yet such overrides could have negative impacts on legitimacy in the eyes of external audiences and other stakeholders. On the other hand, designing this stakeholder with a predominantly external locus of design could allow the stakeholder group to better reflect the interests of scientists studying and putting forward models on the future. While this may provide the synthetic stakeholder with a more conservative base of knowledge on which it bases decisions, it could enhance the perceived legitimacy of this stakeholder with external audiences. Finally, an external design locus would also allow multiple organizations to consider the interests of this stakeholder group in a similar way, because a few stakeholders representing future generations could be made more widely available.

A second interesting extension is the consideration of synthetic shareholders spanning multiple organizations. Here, a synthetic stakeholder could learn from all the organizations with which it engages, thereby providing a broader base of knowledge on which it can base its decisions. While a broader knowledge base is apt to provide certain benefits in terms of triangulation and robustness, developing such a stakeholder is likely to be highly complex, as multiple organizations must come together and agree on a single design. Additionally, this is likely to lead to situations where the single synthetic stakeholder has different levels of salience and locus of design for each organization, though this is not unlike any other stakeholder group engaged with multiple organizations. For instance, an activist group may be highly salient for an oil company that has just experienced an oceanic fuel spill, but less salient for another oil company that has not experienced such a calamity. This option may also be more realistic for assets that are of public interest and of broader importance to a large category of organizations. For example, a coastline undergoing development would benefit from representation and engagement with multiple organizations involved in the project. Finally, the notion of interorganizational representation connects to the concept of legal personhood for natural assets. Specifically, if a natural asset gains legal recognition as a person, multiple organizations would be forced to treat it as such, and therefore may benefit from collaborating on developing a synthetic stakeholder to facilitate the representation of such an asset.

Finally, we suggest that it is worth considering synthetic stakeholders which not only engage with organizational decision-making processes, but also possess formal voting rights and fiscal participation. This idea connects to the burgeoning conversations on web3, and specifically to research exploring distributed autonomous organizations (DAOs) (Murray et al., 2021a, 2022). DAOs are DLT-based organizations "managed entirely through protocols that are encoded and enforced via smart contracts rather than human managers" (Murray et al., 2022, p. 623). DAOs provide a means for several geographically distributed human actors to engage in economic

activities without direct involvement by any single human or group of humans in the allocation of resources. For instance, a DAO's governance protocol determines who can put forward proposals (for example, all members), who can vote on such proposals (for example, those who own a certain number of vote-granting tokens), and how resources are allocated to proposals (for example, automatically in a lump sum).

Since DLTs are integral for functioning DAOs, integrating synthetic stakeholders into a DAO's decision-making is highly feasible. While there are regulatory ambiguities and fiscal matters to manage, in theory, synthetic stakeholders could be integrated into a DAO as token-holding voting members. Here, synthetic stakeholders would have the same voting rights as human actors, effectively turning them into shareholders in the organization. Moreover, a synthetic stakeholder could be integrated into a DAO's governance processes more substantively. For instance, synthetic stakeholders that represent natural assets inherent to the DAO's mission (for example, the Amazon rain forest) could be granted the ability to review proposals before they go up for a vote, and discard those that fail to meet the interests of the represented natural resource(s). Furthermore, as companies begin to explore how to engage in the metaverse, wherein individuals and companies give three-dimensional form to their digital lives, it is not out of the question to consider ways in which synthetic stakeholders can be personified in these digital worlds.

CONCLUDING REMARKS

Stakeholder theory has provided substantial insight into understanding how organizations co-create value with several human stakeholder groups (Mitchell et al., 1997; Tantalo and Priem, 2016). Yet stakeholder theory has largely fallen short in its ability to account for the interests of non-human actors in organizational decision-making. Moreover, the contemporary world and the large-scale problems it faces suggest the importance of a more active and meaningful treatment of non-human actors, particularly natural assets in organizational decision-making. This chapter takes a step in this direction by proposing the concept of a synthetic stakeholder to represent the interests of non-human actors more independently and accurately, and encouraging future work on when and how organizations can integrate synthetic stakeholders into their decision-making processes. Our hope is that this chapter catalyzes rich academic conversation and practical efforts to realize the Suess-ian ideal of giving voice to the trees (Geisel, 1971).

REFERENCES

- Bansal, P. (2003). From issues to actions: The importance of individual concerns and organizational values in responding to natural environmental issues. *Organization Science*, 14(5), 510–527.

- Bansal, P. (2005). Evolving sustainably: A longitudinal study of corporate sustainable development. *Strategic Management Journal*, 26(3), 197–218.
- Bansal, P., and Roth, K. (2000). Why companies go green: A model of ecological responsiveness. *Academy of Management Journal*, 43(4), 717–736.
- Bansal, T., and Song, H.-C. (2017). Similar but not the same: Differentiating corporate sustainability from corporate responsibility. *Academy of Management Annals*, 11, 105–149.
- Barnett, M.L. (2007). Stakeholder influence capacity and the variability of financial returns to corporate social responsibility. *Academy of Management Review*, 32(3), 794–816.
- Barney, J.B., and Harrison, J.S. (2020). Stakeholder theory at the crossroads. *Business and Society*, 59(2), 203–212.
- Bosse, D.A., Phillips, R.A., and Harrison, J.S. (2009). Stakeholders, reciprocity, and firm performance. *Strategic Management Journal*, 30(4), 447–456.
- Boyd, D.R. (2018). Recognizing the rights of nature: Lofty rhetoric or legal revolution? *Natural Resources and Environment*, 32(4), 13–17.
- Buysse, K., and Verbeke, A. (2003). Proactive environmental strategies: A stakeholder management perspective. *Strategic Management Journal*, 24(5), 453–470.
- Driscoll, C., and Starik, M. (2004). The primordial stakeholder: Advancing the conceptual consideration of stakeholder status for the natural environment. *Journal of Business Ethics*, 49(1), 55–73.
- Flammer, C. (2013). Corporate social responsibility and shareholder reaction: The environmental awareness of investors. *Academy of Management Journal*, 56(3), 758–781.
- Freeman, R.E. (1984). *Stakeholder management: Frame-work and philosophy*. Pitman.
- Freeman, R.E. (2010). *Strategic management: A stakeholder approach*. Cambridge University Press.
- Gehman, J., and Grimes, M. (2017). Hidden badge of honor: How contextual distinctiveness affects category promotion among certified B corporations. *Academy of Management Journal*, 60(6), 2294–2320.
- Geisel, T.S. (1971). *The Lorax*. Random House Books for Young Readers.
- Glikson, E., and Woolley, A.W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2), 627–660.
- Gray, K., and Wegner, D.M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, 125(1), 125–130.
- Hoffman, A.J. (1999). Institutional evolution and change: Environmentalism and the US chemical industry. *Academy of Management Journal*, 42(4), 351–371.
- Jacobs, M. (1997). The environment as stakeholder. *Business Strategy Review*, 8(2), 25–28.
- Laine, M. (2010). The nature of nature as a stakeholder. *J Bus Ethics* 96(Suppl 1), 73.
- McGahan, A.M. (2020). Where does an organization's responsibility end? Identifying the boundaries on stakeholder claims. *Academy of Management Discoveries*, 6(1), 8–11.
- Mitchell, R.K., Agle, B.R., and Wood, D.J. (1997). Toward a theory of stakeholder identification and salience: Defining the principle of who and what really counts. *Academy of Management Review*, 22, 853–886.
- Murray, A., Kim, D., and Combs, J. (2022). The promise of a decentralized Internet: What is web 3.0 and how can firms prepare? *Business Horizons*.
- Murray, A., Kuban, S., Josefy, M., and Anderson, J. (2021a). Contracting in the smart era: The implications of blockchain and decentralized autonomous organizations for contracting and corporate governance. *Academy of Management Perspectives*, 35(4), 622–641.
- Murray, A., Rhymer, J., and Sirmon, D.G. (2021b). Humans and technology: Forms of conjoined agency in organizations. *Academy of Management Review*, 46(3), 552–571.
- Norton, S.D. (2007). The natural environment as a salient stakeholder: Non-anthropocentrism, ecosystem stability and the financial markets. *Business Ethics: A European Review*, 16(4), 387–402.

- O'Donnell, E. and Arstein-Kerslake, A. (2021) Recognising personhood: The evolving relationship between the legal person and the state. *Griffith Law Review*, 30(3), 339–347.
- O'Donnell, E., and Macpherson, E. (2019). Voice, power and legitimacy: The role of the legal person in river management in New Zealand, Chile and Australia. *Australasian Journal of Water Resources*, 23(1), 35–44.
- Orts, E.W., and Strudler, A. (2002). The ethical and environmental limits of stakeholder theory. *Business Ethics Quarterly*, 12(2), 215–233.
- Page, J., and Pelizzon, A. (2022). Of rivers, law and justice in the Anthropocene. *The Geographical Journal*, 1–11. Available from: <https://doi.org/10.1111/geoj.12442>.
- Phillips, R.A., and Reichart, J. (2000). The environment as stakeholder? A Fairness-based Approach. *Journal of Business Ethics*, 23(2), 185–197.
- RiverOfLife, M., Pelizzon, A., Anne Poelina, Akhtar-Khavari, A., Clark, C., Laborde, S., Macpherson, E., O'Bryan, K., O'Donnell, E., and Page, J. (2021). Yoongoorrookoo, *Griffith Law Review*, 30(3), 505–529.
- Roulet, T.J., and Touboul, S. (2015). The intentions with which the road is paved: Attitudes to liberalism as determinants of greenwashing. *Journal of Business Ethics*, 128(2), 305–320.
- Salancik, G.R., and Pfeffer, J. (1974). The bases and use of power in organizational decision-making: The case of universities. *Administrative Science Quarterly*, 19, 453–473.
- Starik, M. (1995). Should trees have managerial standing? Toward stakeholder status for non-human nature. *Journal of Business Ethics*, 14(3), 207–217.
- Stead, J.G., and Stead, W.E. (2000). Ecoenterprise strategy: Standing for sustainability. *Journal of Business Ethics*, 24(4), 313–329.
- Stone, C. (1972). Should trees have standing? Towards legal rights for natural objects. *Southern California Law Review*, 45, 450–501.
- Suchman, M.C. (1995). Managing legitimacy: Strategic and institutional approaches. *Academy of Management Review*, 20, 571–610.
- Tantalo, C., and Priem, R.L. (2016). Value creation through stakeholder synergy. *Strategic Management Journal*, 37(2), 314–329.
- Thunberg, G. (2022). *The climate book: Greta Thunberg*, 1st edition. Allen Lane.
- Torelli, R., Balluchi, F., and Lazzini, A. (2020). Greenwashing and environmental communication: Effects on stakeholders' perceptions. *Business Strategy Environment*, 29, 407–421.
- Vanneste, B., and Puranam, P. (2022). Artificial intelligence, trust, and perceptions of agency. Working Paper. <https://doi.org/10.2139/ssrn.3897704>.
- Waytz, A., Heafner, J., and Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, 52, 113–117.
- Wood, D.J., Mitchell, R.K., Agle, B.R., and Bryan, L.M. (2021). Stakeholder identification and salience after 20 years: Progress, problems, and prospects. *Business and Society*, 60(1), 196–245.

14. Interpretable artificial intelligence systems in medical imaging: review and theoretical framework

Tiantian Xian, Panos Constantinides, and Nikolay Mehandjiev

INTRODUCTION

Artificial intelligence (AI) systems have raised significant concerns in recent years, especially with the great performance of deep neural networks (DNN) architectures that utilize a distinct category of black box machine learning (ML) models. DNN learn to map a series of inputs from different types of data (for example, X-ray images) to an output (for example, the probability that the data can point to several clinical categories) (Anthony et al., 2023). DNN usually have multiple layers such that the output of one layer can be the input of the next. By adding more layers and more units of data within each layer, DNN gains more “depth” and can represent functions of increasing complexity (Mitchell, 2019). ML models incorporate mathematical algorithms that perform functions within the DNN layers. For simplicity, we will refer to ML models as both the algorithms and the DNN architecture within which they are deployed.

Compared with other white box ML models, such as decision trees or rule-based models, black box models are thought to provide higher prediction performance at the expense of transparency (Rai, 2020). Black box ML can compute multiple variables and identify distal associations that would otherwise be impossible. However, black box ML models cannot explain how outputs are produced, and their accuracy is challenged because of the way they make connections between data that may often contain errors due to intrinsic and extrinsic bias (Kahneman et al., 2021). In some fields, such as ecommerce recommendation systems, such errors may be acceptable to various degrees; but in other fields, such as healthcare, errors are not an option. Such fields involve high-stakes decisions on the lives and health of human beings. Due to the lack of interpretability of black box ML models for diagnosis, medical professionals lack the clinical evidence to understand how decisions are made. Diagnostic errors can lead to patients not being treated in time, or to excessive medical treatment. Also, it is difficult for model designers to modify the model to avoid such a misdiagnosis directly. These are all obstacles to the implementation of AI systems in the field of medical diagnosis.

AI systems with interpretable ML models are proposed as an alternative option to better understand algorithms’ inner workings (Rudin, 2019; Guidotti et al., 2019).

Interpretable AI systems present the ML model in understandable terms to humans (Doshi-Velez and Kim, 2017a). Interpretable AI is often referred to as explainable, transparent, or trustworthy AI. In this chapter, we will use these terms interchangeably to refer to interpretable AI systems.

Interpretability provides a number of benefits. First, interpretability can reduce noise and potential biases, such as errors based on disease cases and context, or on inclination, or prejudice for or against someone or something. Second, interpretable AI systems provide a traceable and verifiable approach to justify algorithmic decision-making by monitoring changes in production and training data. Third, interpretable AI systems help to debug ML models, optimize them, and improve their performance. It is easier to create novel features, eliminate unnecessary features, and collect extra data to increase the performance of the ML model with a deeper understanding of the correlations between input and output (Adadi and Berrada, 2018; Bhatt et al., 2020b; Das and Rad, 2020).

Interpretable AI systems have recently attracted the attention of researchers in both computer science and information systems (IS) research fields. For example, in computer science, scholars have offered surveys of interpretable AI systems in the medical domain by reviewing multiple modalities of medical data such as tabular data, images, and textual data (Tjoa and Guan, 2021). They categorized applications into visual, textual, and example-based (van der Velden et al., 2022). Also, scholars reviewed the explainability of deep learning with various methods (Jin et al., 2022) and different types of explanations (Patrício et al., 2022).

In the IS research domain, the concept of explainability is not new. As technology continued to evolve and expert systems emerged during the 1980s and 1990s, researchers began investigating the need for explanations using these systems (Ji-Ye Mao, 2015; Ye and Johnson, 1995). This research identified various factors that affect how different user groups interpret explanations provided by information systems, and those explanations affect how users perceive the system and make decisions. A more recent review of explainable AI (XAI) systems has discussed ways to unmask black box ML models (Rai, 2020). It has been proposed that further research focuses on the accuracy–explainability trade-off, AI trustworthiness, AI fairness, and the levels of explanation and transparency for end users. Others have developed a responsible AI framework for digital health where explainable AI contributes to one type of ethical concern (Trocin et al., 2021). The authors pointed out six ethical concerns in digital health for a responsible AI, and provided a research agenda for future IS research. Furthermore, a framework of trustworthy AI has been introduced as a promising research topic for IS research (Thiebes et al., 2020).

Despite the varied themes emerging from this body of research, there is an underlying thread that cuts across all review papers and commentaries: the importance of keeping humans in the loop (Baird and Maruping, 2021; Fügenger et al., 2021; Rai et al., 2019). It does not mean that humans are error-free and are constantly correcting errors made by AI systems. Instead, the argument is that we should not become complacent and overconfident about automation by machines. We should instead

examine how humans can augment and be augmented by machines, while monitoring one another's performance.

In addition, researchers have already recognized that a large part of AI systems are powered by big data that need to be correctly labelled, processed, and made available to the designers of ML models (Alaimo and Kallinikos, 2021; Jones, 2019; Monteiro and Parmiggiani, 2019). AI systems are not automatically and unproblematically produced, ready for immediate deployment in different organizational settings. Data play a big role in training and validating the performance of AI systems, and constitute the third component of interpretability, together with human agents and ML models.

This chapter focuses on interpretable AI systems in medical imaging. We differentiate between data, human agents (that is, model designers, regulators, and end users), and ML models as key components in producing interpretable decision-making. We narrow our focus to breast cancer medical images, as this is a disease that affects the most significant number of people in the world (IARC, 2021), and it therefore offers a broader set of interpretable AI systems and datasets. We examine how the three dimensions contribute to interpretable decision-making and identify the tensions that emerge in the process. We build a theoretical framework that helps to analytically categorize the tensions between human agents, data, and ML models. We conclude with implications for further research.

COMPONENTS OF AN INTERPRETABLE AI SYSTEM

In this section, we elaborate on the three components of an interpretable AI system: ML models, human agents, and data. To contextualize our discussion, we use the example of mammogram diagnosis in the United Kingdom (UK). The general workflow for screening mammograms by radiologists, with two readers and one referee, is shown in Figure 14.1. A similar workflow, replacing one radiologist with the AI recommender system, is illustrated in Figure 14.2, providing an example of how an interpretable AI system can augment this process. We also summarize the AI model developing process and human roles in each step in Figure 14.3.

The norm in the UK's National Health Service Breast Screening Program (NHSBSP) is to perform a dual examination of mammograms taken when screening (Chen et al., 2023). Hence, the medical image will be sent to two radiologists independently to make a diagnosis at the start. If the outcomes of the two diagnoses are consistent, the process ends. If not, a third radiologist, who is generally a senior experienced radiologist, will be introduced to compare the two different diagnoses and make a final decision. The cases with suspicious abnormalities will be recalled and sent to a breast disease specialist for further consideration. Ultimately, the medical image with diagnosis and description will be stored on the electronic medical records in the medical system. This double-reading approach with arbitration has been demonstrated to enhance cancer detection rates by 6 to 15 percent without significantly increasing the rate of recalling false positives (RRs). Still, it has a high labor

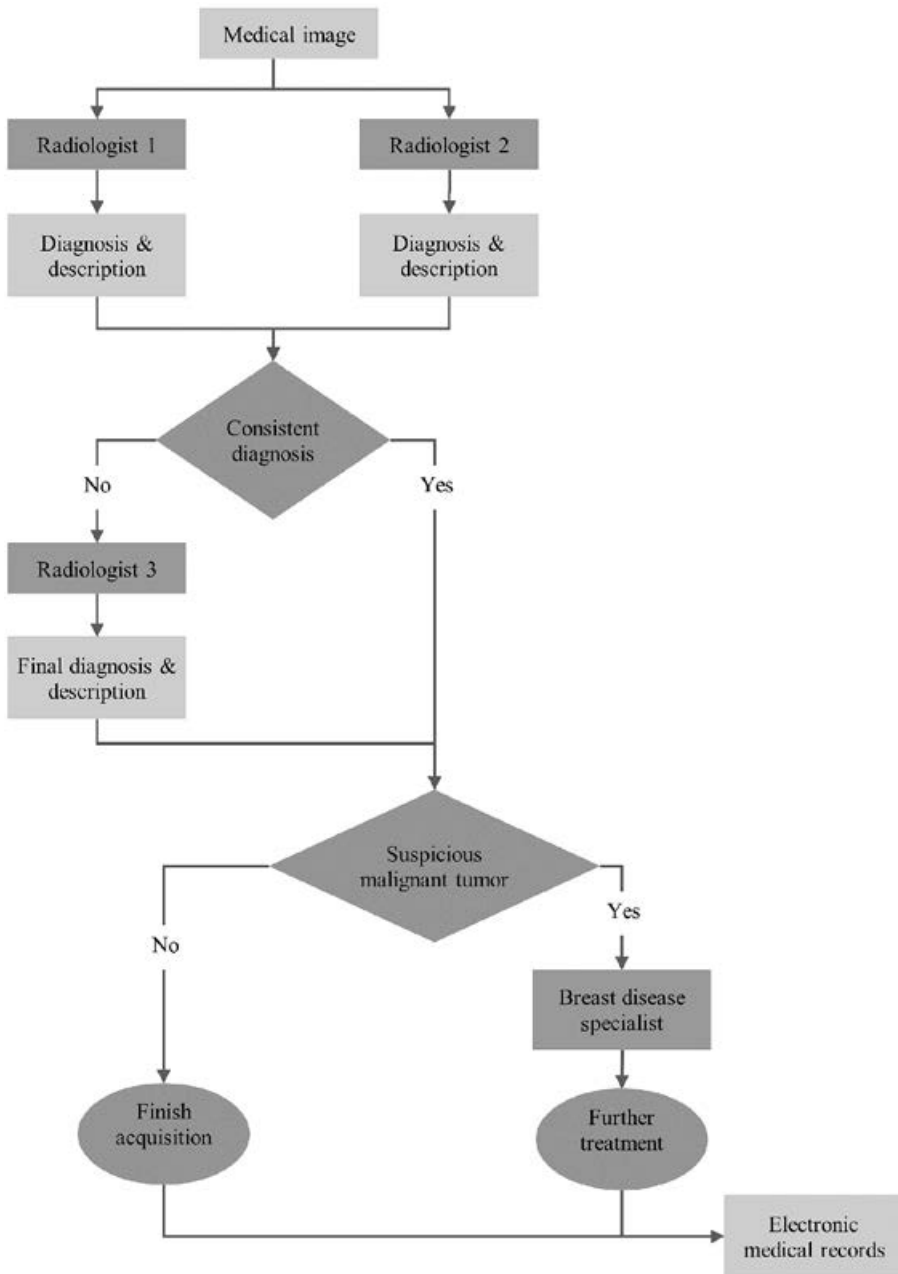


Figure 14.1 Process workflow of screening mammograms by radiologists

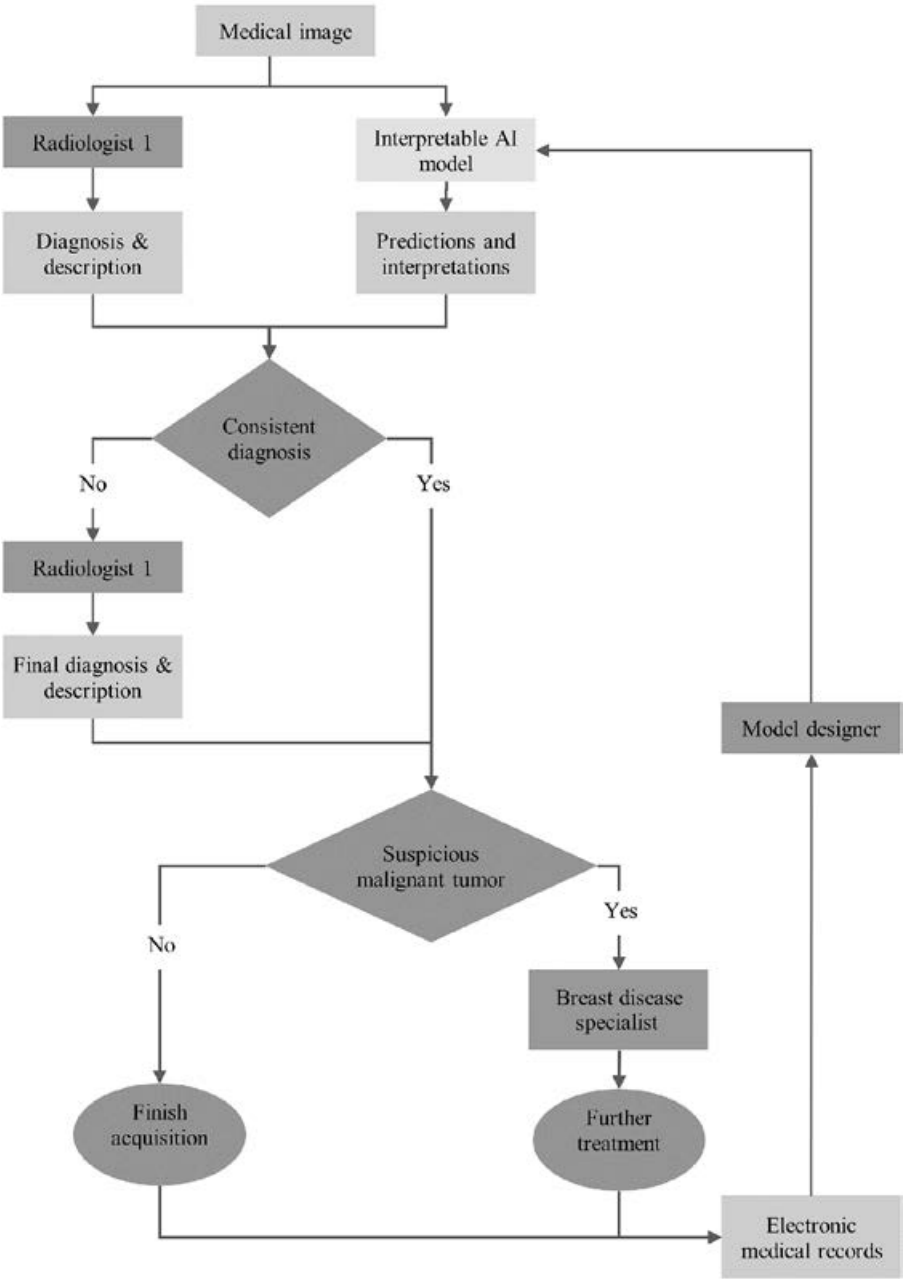


Figure 14.2 Process workflow of augmenting the screening of mammograms with an interpretable AI system

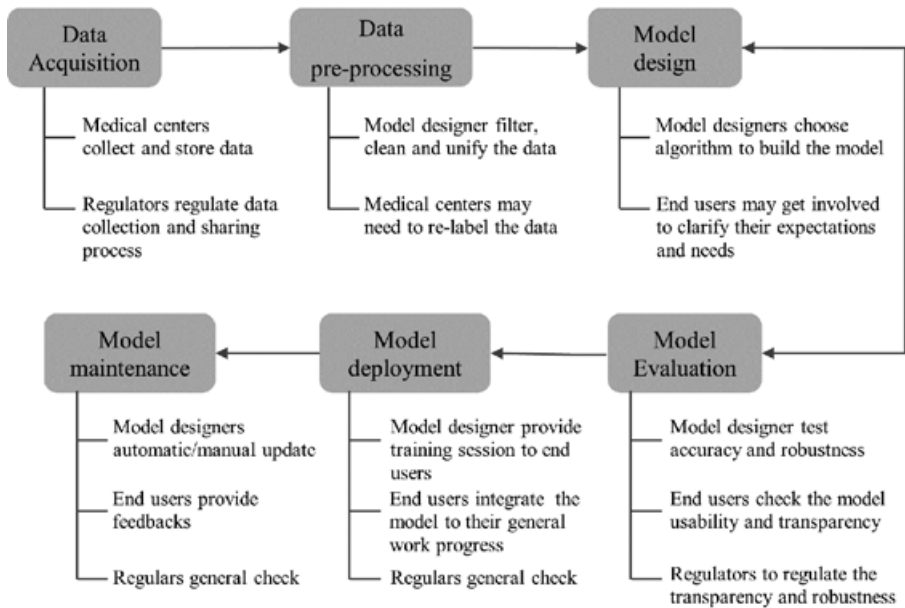


Figure 14.3 An overview of the development process of an interpretable AI system in the medical image field, and the human roles in each step

cost (Blanks et al., 1998; Harvey et al., 2003). The shortage of radiologists causes the need to improve work efficiency from a labor-intensive to a technology-intensive environment. Researchers have shown that a single-reading screening of mammograms augmented by AI systems can be developed to provide decision support for medical imaging diagnosis (Gromet, 2008; Taylor and Potts, 2008). All the medical images and diagnoses are accumulated in electronic medical records. Hence, model designers can select the data and proper label they need to build an ML model. AI-augmented single reading can minimize the workload involved, by omitting the second reader's work when the decision of the ML model is consistent with that of the first reader. If not, the first reader will check the result of the ML model and decide whether to accept or deny the recommendation. The comparable performance of the two process workflows makes the single-reading screening of mammograms augmented by AI systems a good alternative to double reading. It should be noted that the AI model's performance highly relies on the features and information inside the input data. Hence, the quality of input data is one of the key factors affecting the final model's performance. Furthermore, the model's performance will affect the human decision-making process as it will provide supporting information to a human.

Despite this potential benefit of AI-augmented diagnosis, the second process workflow may give rise to several tensions. First, in the model evaluation process shown in Figure 14.3, the expectations towards the model from different user groups are different. For example, the model designer would like a more accurate model. In

Table 14.1 Three components in an interpretable AI system and their classifications and dimensions

	Classification	Dimensions
Human agent	Model designer	Experience/expertise
	Regulator	Expectation
	End user	
Data	Image	Quantity
	Label	Quality
ML model	Data-centric	Prediction accuracy
	Model-centric	User-dependent explainability

contrast, radiologists care more about integrating the model into their general workflow, and the verifiability of the model prediction by their professional knowledge. Second, in the data acquisition process shown in Figure 14.3, the attitudes towards data availability usually differ between different groups of people. For example, the model designer would prefer to have a comprehensive and large amount of data from the medical center to build a high-performance model. In contrast, radiologists may not necessary need a lot of data to build their diagnostic capacity (Alvarado, 2022). Last but not least, the development of ML models relies on a large amount of data and multilevel fine-grained classification of data labels. But the application of high-quality datasets is usually restricted by data access standards, privacy regulations, data storage, and sharing rules that limits relevant research.

Inspired by this discussion, we conclude that it is important to understand the tensions from the interpretable AI system in order to improve its performance (Van Den Broek et al., 2022). To understand the tensions, we start to discuss the three main components of an interpretable AI system. The various classifications and dimensions of the components help to understand better how tensions may emerge. Table 14.1 provides a summary of the three components.

Interpretable AI: The Human Agents Component

Apart from technical experts, several groups of people are involved in implementing the interpretable AI system. These groups will have different expectations towards the interpretability of these ML models (Langer et al., 2021).

A simple categorization is an expertise-based framework where human agents are ranked from “beginners” to “experts” in machine learning (Hohman et al., 2018; Yu and Shi, 2018). Yet, the most broadly accepted framework categorizes human agents according to their functional roles, since users’ demands on ML interpretability are generally role-oriented (Belle and Papantonis, 2021; Bhatt et al., 2020b; Hong et al., 2020). For example, Tomsett et al. (2018) defined six roles, including data provider, model designer, model user, decision maker, model examiner, and a final group of human agents who are influenced by the decisions affected by the system. These roles constitute a framework to be considered in the types of acceptable explanations

Table 14.2 *Human agents: classifications and dimensions for human agents*

Classification / Dimension	Expertise/experience	Expectations
Model designer	Expert in machine learning and computing algorithms	Verification (recognize and correct errors) of the ML model)
	Beginner in the domain	Performance of the ML model
Regulator	Experts in multidisciplinary knowledge including machine learning, medicine, legal, social, management, and others	Trustworthy ML models Transparency and accountability in decision-making
End user	Beginner in machine learning and computing algorithms	Usability of ML models
	Expert in the domain area	Trust in ML models

or interpretations they need. Preece et al. (2018) conducted further analysis of the stakeholder communities by identifying their distinct motivations that help to classify explanations of interpretable ML models according to developers and end users. According to how closely connected the system is, the model designer, regulator, and end user are commonly accepted as the core types of human agents, which are summarized at Table 14.2.

Model designers

The model designer role refers to the community that creates the interpretable ML model, including design, training, testing, deployment, and maintenance. Data scientists, computer science researchers, software engineers, and programmers all belong to this group. They are experts in machine learning, but have little to no knowledge in medicine. The top priority objective of interpretability for model designers is verification, debugging, and improving the model's performance, which requires the ML model to incorporate merits that uncover inconsistencies between the model's algorithm and the developer's expectations (Bhatt et al., 2020a). Model designers would like tools that efficiently help them to detect whether the software they developed, which incorporates the ML models, is functioning correctly. Model designers focus more on the performance of the ML models and the corresponding software rather than on providing proper explanations to help end users to understand why the model's prediction is reasonable (Hepenstal and McNeish, 2020). For example, model designers build simple models to approximate the behavior of complex ML models (Ribeiro et al., 2016). Alternatively, they create tools to help them understand intuitively, by enabling users to quickly test how various inputs correlate with different outputs (B. Kim et al., 2018; Lundberg and Lee, 2017). Inevitably, most of the tools for interpreting ML models are tailored for model designers, since they are the people who build the tools (Brennen, 2020).

Regulators

Regulators are primarily business owners, product managers, and domain experts who have domain knowledge and decide whether the developed ML models meet

the end user expectations. People who examine the models from a legal perspective, usually called auditors, are also involved in this role (Hong et al., 2020). While regulators are not necessarily experts in ML models and algorithms, interpretability is a useful tool to build authority and improve management towards the ML model. Product managers emphasize the accountability and validation of the ML models, which are perceived as the key factors in building trust (Doshi-Velez et al., 2017b). Another key factor in building trust in ML models, perceived by the domain experts, is reducing bias from the prediction, since ML could learn the correlations in data that exert undesirable forms of bias or are mislabeled (Kamiran and Calders, 2009). A severe risk may be raised if one “over-trusts” the biased prediction of an ML model (Modarres et al., 2018). Nevertheless, auditors, particularly in the medical imaging field, care more about the compliance of an ML model to the regulations (Raji et al., 2020), and strive for a clear understanding of how the ML models use the data (Zarsky, 2013). However, the strict regulation in the healthcare domain restricts the thoroughness and quantity of the collected data, resulting in a higher possibility of systematic bias in the trained ML models (Raji et al., 2020).

End users

End users directly interact with the interpretable ML models (Weller, 2019). In the medical imaging domain, end users provide an input image to the model and expect to receive the output from the ML model, which enables them to either acquire specific information or make a decision accordingly (Fuhrman et al., 2022). Radiologists and physicians with strong medical backgrounds, rather than general patients who barely know medicine, are considered to be the main end users for medical imaging diagnosis. Justifications are needed for end users to build trust in the ML model (Tonekaboni et al., 2019). It is important to recognize and correct errors made by ML models to the greatest extent possible. Additionally, it is crucial to understand the limitations of the model and the uncertainty of each prediction. Apart from the experts, junior physicians or radiologists also want to learn about their domain, where the ML models could provide clear and reasonable interpretation (Liao et al., 2020). To better support their job, the end users also demand incorporating the model’s output into downstream actions (Harned et al., 2019). After reading a medical image, generating a medical report describing the image’s findings and providing a conclusion in medical terminology is necessary. Hence explanations from the model, which will provide more information to support the medical report, will be more supported and welcomed.

In summary, different groups of human agents have varied requirements and expectations regarding the interpretability of ML models. Model designers favor algorithms that help to build and debug models more efficiently. At the same time, regulators place more value on understanding how the ML model functions so that potential bias is reduced, and the regulatory concerns are addressed. End users, on the other hand, demand a good understanding of the correctness of the models to help them make rational decisions. However, very few ML models can fulfil all

Table 14.3 *The classifications and dimensions for medical imaging data*

Classification / Dimension	Quantity	Quality
Images	Large datasets are potentially more beneficial to the ML model Small datasets can be improved by data augmentation and transfer learning	Diversity in data sources Unified image type and image size High resolution of datasets
Label	Diagnosis Electrical health record Specific annotation Medical report	Labelling standard Ground truth > prospective annotation > retrospective annotation

these requirements simultaneously. Therefore, balancing the different human agents' needs is worth further investigation and is discussed below.

Interpretable AI: The Data Component

The quality and quantity of data play a vital role in the performance of ML models. Data are the crucial ingredient that makes machine learning possible. Machine learning could work well without complex algorithms, but not without good data. A higher quantity of data provides a higher possibility of identifying “useful information” about features found in different images. However, a higher quantity of data does not guarantee this possibility, and this is where data quality comes into play. The absence of data quality includes various aspects, including errors in data capture, data selection, data cleansing, and feature transformation. A good way to prove the data quality is to increase the diversity by introducing more data sources, various medical inspection equipment, various genders of patients, various ethnic groups, longer time span data collection, various stages of diseases, and multiple levels of annotations to the data. Diversified data can provide more dimensions of data features and decrease bias. The medical image dataset contains two parts: images and labels. In the following, we discuss the quantity and quality of the dataset in each part.

In terms of image quantity, generally a larger number of images gives more potential to build a high-performing model. However, sometimes it is hard to collect a relatively large dataset. Data augmentation would be an alternative technique to expand its quantity. Basic data augmentation techniques include translation, rotation, flip, cropping, gamma correction, denoising, and so on (Barnett et al., 2021; Dong et al., 2021; El Adoui et al., 2020). Wickramanayake et al. (2021) used concept-based explanations to augment data for the image classification tasks. This method can recognize samples in the under-represented regions. Apart from data augmentation, transfer learning is also a good way to reduce the impact of small data volume on model performance. For example, Graziani et al. used a ResNet 101 model pre-trained on ImageNet with binary cross-entropy loss to classify tumor and non-tumor patches (Graziani et al., 2018).

Regarding image quality, it is beneficial if the images come from different imaging machines, medical centers, or populations. Various data sources provide high diversity and reduce data bias. However, data collected from various hospitals or organizations are prone to inconsistencies in the type of images or inconsistent sizes. It is necessary to adjust them to a unified standard. Apart from that, high resolution is a universal standard for a quality imaging dataset. However, high resolution will bring an extra burden to computation. Correctly identifying and segmenting the region of interest (ROI) containing a minimum suspicious area will improve computation efficiency. In medical imaging, the boundaries of a tumour may be defined on an image or in a volume and treated as an ROI (S.T. Kim et al., 2018).

In a typical medical imaging dataset, a set of pre-determined labels are correlated with each image. This may contain statistical records from electronic medical records (EMR). Reports for each image may also be saved together with images, which is text data. In this case, model designers can use multiple modalities of medical information to train the ML model for early cancer detection (Akselrod-Ballin et al., 2019; Wang et al., 2017). In addition, some datasets have fine annotations that list the image's detailed medical features as reasons to support the image label (Moreira et al., 2012). These detailed fine annotations provide possibilities for model designers to build an interpretable AI model with medical-language explanations.

To have better-quality labels for the images, the standard for labelling them is crucial. The most reliable standard is the ground truth, which typically refers to information acquired from direct observation, such as biopsy or laboratory results. A slightly inferior standard is prospective labelling, which has the advantage of integrating contemporaneous information (such as past medical history and biopsy data). The least valuable labelling standard is retrospective labelling. Retrospective labelling approaches include manual labelling by radiologists of the previous images, automated labelling by machines, and so on (Willemink et al., 2020).

Previous research has sought access to medical images from either publicly open databases or privately. Public databases have the advantage of easy access, are well labelled, and benefit from the segmentation of the original images to reduce the size of input data. These merits can save the researcher's time in pre-processing the data. Also, it makes a comparison of different ML models convenient since those models use the same dataset. Some examples include DDSM, the most widely used public mammogram dataset. DDSM comprises 2620 digitalized mammography screening images from four hospitals in the United States. All the images are labelled with normal, benign, and malignant cases, with coarse verified pathology information. However, the data have not been updated for 22 years, which may not include the newly developed features. In contrast to DDSM, INbreast is another dataset which contains 410 full-field digital mammograms. All cases have biopsy proof as the ground truth, and include the location and boundaries of the lesion, with the outline marking performed by an imaging specialist. More detailed classifications are provided by the six levels of BI-RADS as well as fine annotations at the pixel level. This dataset has a limited number of images but is well rated regarding the diversity

Table 14.4 *Classifications and dimensions for interpretable AI models*

Classifications/ Dimension	Accuracy and AUC	User-dependent explainability
Model-centric	High accuracy and AUC	High explainability for model designer
Data-centric	High accuracy and AUC	High explainability for the end user

of images and the state-of-art format of images, guaranteeing relatively high data quality.

Medical images held privately in hospitals or other healthcare organizations are another source of input data for ML models. Private data sources are more customized to meet the requirements of researchers to the greatest extent. Private data sources require data scientists to invest extra labor, time, and funds to process data. Hence, it usually takes longer to find collaborating hospitals and extra time to pre-process the data, such as segmenting the ROIs and augmenting refined annotations of the images. Moreover, access to such private datasets raises privacy and security concerns, which prohibit the potential for other researchers to conduct further research on the previous work. On the other hand, the customized dataset extends more freedom in designing ML models. Akselrod-Ballin et al. (2019) and Shen et al. (2021) generated a private dataset which contains more than 20,000 images. Their systems benefited from the great advantage of a large dataset and had a better performance with more than 90 percent of the area under the curve (AUC). Barnett et al. produced a private dataset which contains fine annotations of the lesion area, which built the base of their case-based study (Barnett et al., 2021). In the end, both public and private datasets have pros and cons. Researchers could select datasets according to their own needs.

Interpretable AI: The ML Model Component

To achieve interpretability, one could design either an intrinsically interpretable model, or an explainable model that could explain the black box ML model. Intrinsically interpretable models should be the first option, as they theoretically have better performance in terms of both accuracy and explainability (Rudin, 2019). Most intrinsically interpretable models, such as linear regression and decision trees, do not perform highly in image classification tasks. In contrast, black box models have a relatively high accuracy in the computer vision area.

In this section, we present an overview of the current interpretable AI methods which are specifically designed for or generally applied to medical image diagnosing problems. We divide the interpretable AI models into two groups, namely, model-centric and data-centric models, as shown in Table 14.4. The model-centric AI model focuses on engineering the model to improve its performance and interpretability; while the data-centric model focuses on engineering the data. A summary of the fundamental features and limitations is presented in Table 14.5. More details are provided below, followed by classical interpretable AI models and examples used in medical imaging tasks.

Table 14.5 *Summary of the features and limitations of model-centric and data-centric models*

Classifications	Features	Limitation
Model-centric	Limited efforts are needed for data operations Data quantity is more important than data quality Further data input is not required Focus on the model	Likely to reach accuracy limitation Less reliable outcomes due to potential bias in the data Higher costs due to the need for large data quantity
Data-centric	Data processing is essential Data quality is more important than data quantity Data are not static Domain expertise is required	A higher standard for labelling Require data augmentation for more representativeness and better generalization A reliable algorithm for minimizing bias is essential

Model-centric models (image-level labelled)

Model-centric models focus on explaining the algorithm at the image level without any subclassifications to the input images. Researchers either explain the predictions of the model, or explain the inner working of algorithms, also called local and global explanations. Model-centric models are more relevant to model designers as they provide insight into the inner part of the model, and can potentially optimize the model's performance. In addition, this type of method is easier to implement and has lower computational complexity than data-centric models. Overall, the model-centric models discussed below use various parameters to represent the correlation score between the input and the output.

Typical model-centric models include the back-propagation method, the class activation map, and the perturbation-based model. For example, the Shapley Additive exPlanations (SHAP) method, a local model-agnostic interpretation technique, was inspired by Shapley values in cooperative game theory (Lundberg and Lee, 2017). It determines the average contribution of a feature value to the model prediction by considering all possible combinations of features in the powerset. This contribution is determined by observing the change in model prediction when using 2^n combinations from the features powerset, with the missing parts replaced by random values. For example, one study generated a SHAP map and used different color dots on the original images to show the critical area corresponding to the output predictions (van der Velden et al., 2020).

The key advantage of model-centric models is that they may uncover previously unrecognized disease features without any prespecified disease characteristics. In addition, model-centric models have reduced data bias in feature selection. However, the key disadvantage is that the limited explanation information can affect the trust towards the system. One needs a professional background to understand the prediction results. Nevertheless, model-centric models are more likely to reach accuracy limitations due to fixed data input and coarse pre-processing of low-quality data. The costs of acquiring large amounts of data are also extremely high compared

to common data-centric models. Hence, data-centric models are more favored by experts in the medical imaging field.

Data-centric models (pixel-wise labelled)

Data-centric models rely on feature extraction from the input data by experts in a specific domain for interpretation. The end user does not need to understand the working algorithm inside an ML model. Before the training session, a data-centric model analyses the medical features of input data and provides further detailed fine annotation of images based on subclass features. The annotation will be based on the ‘pixel’ level instead of the image level.

Data-centric models for medical imaging generally incorporate three stages. First, the detection of regions of interest (ROIs) containing a suspicious lesion; second, the classification of ROIs into different features related to the final prediction (such as high density or low density); third, integrating the classification from the second stage into the final classification result, with an explanation or medical description.

Data-centric models can provide higher end user-friendly explainability as they contain universal terminologies to support the predictions. End users of any experience level could understand that information as it is theoretical knowledge. This type of model can have high accuracy and AUC after careful design. Typical data-centric models include four types: case-based explanation, concept-based explanation, intrinsic white box to explain the black box, and counterfactual explanation.

Case-based reasoning is an example-based explanation method that searches for the similarities between the input and query images. For example, Barnett et al. introduced a case-based, inherently interpretable deep learning system called IAIA-BL, which provides both global and local interpretability (Barnett et al., 2021). IAIA-BL is an optimized model based on ProtoPNet. Barnett et al. used it in breast cancer diagnosing based on mammograms, and achieved an equal or even higher performance than the uninterpretable models. Before the training, they used a small sample of high-quality mammograms to make the fine annotations into three prototypes. During the training, they calculated a similarity score to each prototype, and went through a weighted matrix to get the final probability of malignant breast cancer. Finally, the model produced a prediction result, a heat map based on the original input image showing the key area relating to the prediction result, and a sentence explaining the medical features of the key area.

Counterfactual explanations are based on the principle that “if an input datapoint were x' instead of x , then an ML model’s output would be y' instead of y (Goyal et al., 2019). Hence, this method also needs to analyze the data and find its causal features related to a specific output prediction. Wang et al. (2021) proposed a counterfactual generative network, and verified its performance on the public mammogram dataset INBreast and an in-house dataset and ablation studies. In their result, they visualized the original image, the target image with target features marked by a green rectangle, the reference image which is lesion-free in the corresponding area, the heatmap of the target image, the heatmap of the reference image, and the heatmap of

the counterfactual image. By comparing those images, physicians could have a clear understanding of the lesion location and the features of the lesion.

The aforementioned examples of data-centric models have several advantages. First, semantic medical features provided by the interpretable AI model can be cross-validated with the user’s own knowledge, increasing confidence in the model and enhancing the user’s trust. Second, they make it easier for the users to recognize and correct the wrong predictions made by the ML model with the medical features. Finally, the medical features provide support material for the radiologists not only to evaluate the images, but also to generate a diagnostic report.

Meanwhile, they also have a number of disadvantages. First, the ROI localization and ROI classification process are suboptimized. The features extracted by the system are human-designed, and carry the observer’s bias. Second, detecting ROIs and classifying multiple subclass features is more computationally expensive. Third, as the features are human-predefined, the data-centric model loses the ability to discover the unrecognized features by humans. Finally, the training dataset requires periodic updates to include new cases. It is advisable to feed the updated training dataset into the interpretable AI model to facilitate comprehensive feature selection.

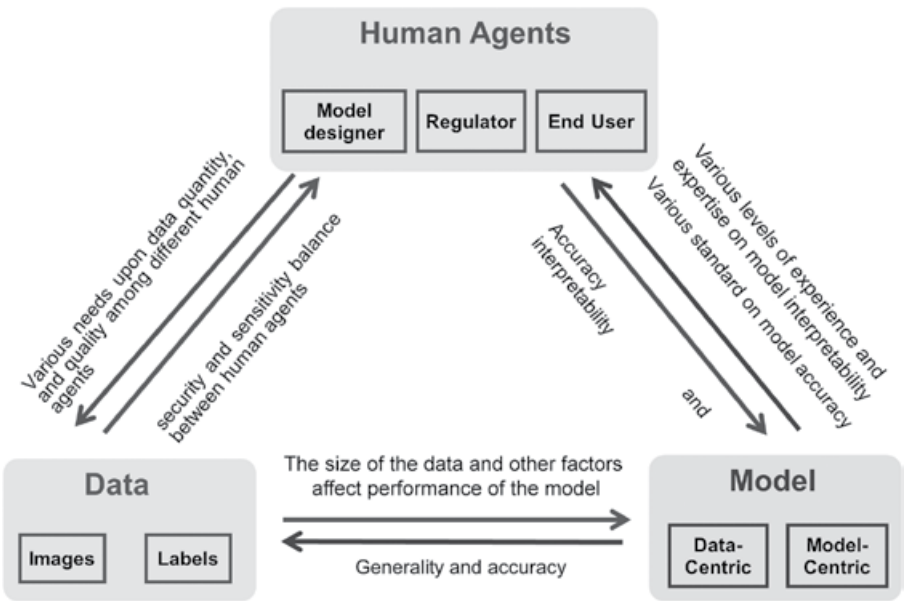


Figure 14.4 Tensions emerging from an Interpretable AI system

A FRAMEWORK FOR EXAMINING TENSIONS IN INTERPRETABLE AI SYSTEMS

In this section, we integrate the three components into a framework to help us understand the interdependencies and tensions between them. We return to the process workflow presented in Figure 14.2 to illustrate the tensions. Figure 14.4 summarizes these tensions.

Tensions between Human Agents and Model

Tensions may arise between the different user experience levels and users' acceptance of the interpretable AI outputs. In the majority of cases, radiologists can arrive at a correct diagnosis according to their experience and professional knowledge. Yet sometimes, irresistible factors such as limited processing time, negative emotion, and overwhelming tasks may lead to oversight of the suspicious features. In such cases, the interpreting AI may function as a correction check that points out the risky details to the radiologists. Research has shown that with the help of interpretable AI, experienced radiologists are more willing to accept AI recommendations, as they have much knowledge to cross-validate the inconsistent outcome by the interpretable AI model. However, the junior radiologist, who has limited experience, is more conservative in accepting interpretations from AI. The junior radiologist might need a more detailed explanation of the model. Hence, they have more material to cross-validate with their professional medical knowledge; for example, the shape of the tumor and the margin of the tumor. In contrast, the senior radiologist may need a less detailed explanation from the model; for example, the position of the tumor. Senior radiologists already have a stronger ability to discover the details of the medical image themselves compared with juniors. The various experience of end users affects the final explanation content of the model, leading to the choice of model types.

With the different expertise of the human agents, their expectations towards the model interpretability are also different, which leads to the tensions of ML model choice. Model designers require a thorough understanding of the algorithm and data-flow so that the model can be verified (Montavon et al., 2018). Their goal is to build a more accurate model. They prefer to rely on more statistical explanations which are true to the model as an approach to optimize and debug the model. Regulators demand a clear understanding of how the model functions, including how to process the data and criteria for drawing certain conclusions, as the key to justifying whether the AI systems are lawful, robust, and ethical (Lipton, 2018). Nevertheless, end users demand merits of interpretability that are more arbitrary and subjective, so that their concerns about fairness and morality are adequately met (Samek et al., 2017). End users are concerned about whether the interpretation could be cross-validated by their professional knowledge, such as features of the medical images and locations of suspicious tumors. However, those explanations are not friendly for model designers or regulators to understand without rich knowledge in medicine. It would be beneficial

to study the optimal degree of interpretability between different groups of human agents.

As already discussed, there are two dimensions for modelling. One is the model's accuracy, and the other is user-dependent interpretability. The tension between the model designer and end users arises regarding their understanding of the model's accuracy. Model designers build the interpretable AI model based on past data. The accuracy is tested and verified by the past data. However, the end users typically apply the model for predicting information based on newly generated images. The data characteristics of the new images may differ from the past data. For example, the previous database may be dominated by elderly women, but the current environment for users has a wide age range of patients, ranging from young to old. In this case, a model trained on a database of older women may be less accurate in predicting images of breast cancer in younger women in the current environment. To ease the misunderstandings about the model's accuracy, introducing a consistency check between the training data and the cases from potential patients would be necessary before the model was introduced. Alternatively, re-calculating the model's accuracy using the hospital's past real data would also introduce a proper correction to the metrics.

Tensions between Human Agents and Data

When data are accumulated in large quantities, significant tension becomes apparent between the model designer and the end users, since the model designers pursue a large amount of data, while the end users are concerned about data-sharing privacy and profit distribution. Appropriate profit distribution is one of the driving forces for technological progress. Original data are generated from patients. But radiologists in medical centers devote time and effort to developing, labelling, storing, and managing those data. The profit distribution between the data collector (medical centers) and data users (ML model provider company) is unclear, which obstructs the circulation channels for data collectors and data users. Furthermore, data privacy is another concern during the data-sharing process. Especially, medical data includes a large amount of human personal information. Although there are successful technologies to make the data de-identified, the increasing risk of data leakage inevitably decreases the patient's trust towards the medical centers. In this situation, regulators are greatly needed. They have the responsibility to communicate with data collectors and data users. Regulations by regulators are needed on how to standardize the collection of data, and how to protect users' privacy during the data deployment in deep learning. Technically, transfer learning and other data augmentation methods could be applied to solve the limited amount of data.

Regarding data quality, another tension can arise between the need for model designers regarding data labels containing subdivisions of different precisions, and the extra workload of radiologists. Various precision levels of data labels give model designers more chances to develop ML models. However, radiologists may have different habits in describing the features of suspicious tumors in the images. The

fundamental requirement for radiologists is to correctly diagnose and clearly describe the diagnosis, as long as the visible features can support their diagnosis. However, model builders would prefer complete and comprehensive labels for the images (Van Den Broek et al., 2022). Such preference will lead to an extra workload for the radiologists. Given that most doctors are already working in a high-intensity work environment, it is not easy to balance the need of the two groups of people.

Homogeneity of data is another tension that may arise between different groups of human agents' expectations and their general responsibility. Data collectors, such as radiologists in the medical imaging domain, focus on the content of the data instead of the type. However, the format, size, and label standard of the data have a deeper influence on the usability of data, especially when model designers have to accumulate data from various medical centers. For example, radiologists in the United States use the BI-RADS standards to classify and label the mammograms (American College of Radiology and D'Orsi, 2018). At the same time, radiologists from other countries may be more familiar with different standards issued by their local region. In addition, with the limitations of screening devices, some medical centers generate film-screen mammogram on X-ray film, and others may generate digital versions such as full-field digital mammography (FFDM). Different formats of data or various data label standards cannot be applied to ML model development directly. The effect of homogeneity of data has a more significant impact on rare diseases and rare types of images. The effect of homogeneity of data becomes more serious, as there would be a trend to collect data from many different sources to ensure the broad applicability of the model. Attention is needed from regulators, end users, and model designers on the homogeneity of data.

The security requirement of the data typically impacts the workflow of ML design and final prediction. For example, high-sensitivity data may require strict security, so that the model developer may not be able to access the data. Instead, the pre-trained model is provided to the data holder, who will train and test the model. However, such workflow inevitably results in lower performance of the training ML models, since most data holders are typically not experts in ML algorithms. A balance between the data provider and data user regarding the data security and sensitivity need to be carefully considered to achieve a win-win situation.

Tensions between ML Model and Data

Like human agents who need food and water to grow, so the training and maintenance of ML models rely on a vast amount of real-life and experimental data. Tensions grow between the ML model's thirst for and use of a huge amount of personal data, and personal data's status as a protected commodity under privacy laws, such as the Health Insurance Portability and Accountability Act of 1996 (Act, 1996) in the United States. Regulators are urgently needed to formulate clear legal norms which can clarify the processes and boundaries of data collection, use, storage, and commercialization. Clear regulations are needed to boost the accumulation of big data. Big data enables rapid development of machine learning models. It does

not mean the larger the better. Tensions emerge in the selection of the appropriate size of training data among other factors for organizations, such as the quality of data, various algorithms, and computational resources (Wang et al., 2020). A more comprehensive consideration of all the factors is needed to determine the size of the training set, as well as the resources in hand for organizations.

Furthermore, various types of data, such as statistical tabular, image, text, and audio format, generally require different ML models to train and process. Different algorithms may be needed even for image data with different characteristic features. Such cases open a dilemma in the generality and accuracy of an ML model, and the corresponding interpretation algorithm (Gohel et al., 2021). For example, in the medical imaging processing field, specific models such as SHAP and LIME outplay the universal interpretable ML models in terms of prediction accuracy, but are poorly behaved in generality assessment (Daley et al., 2022). Developing an all-round algorithm that addresses both the variety of data and the accuracy of prediction is a challenging task and deserves further investigation.

CONCLUSION AND IMPLICATIONS FOR FURTHER RESEARCH

Although the application of interpretable AI in the field of medical imaging is still in its infancy, it has shown great potential, but not without emergent tensions. According to our analysis, the three components of an interpretable AI system are interdependent and can be generative of tensions that spill into larger sociotechnical systems, incorporating organizations and their practices. ML model designers are never alone in designing these models, and neither are the end users who use these models in their routine practices. Regulators who seek and gain access to the data that feed into these ML models are also responsible for management, legal, and ethical concerns that may arise in the deployment of the model. There is thus a co-constituting relationship between the three components.

A number of research implications emerge from our framework and the discussion of these interdependent tensions. First, a key question raised is: how can organizations balance the benefits and risks of a varied degree of interpretability in the ML model's performance among different stakeholders? The optimal degree of an ML model's interpretability will vary depending on the stakeholders involved. Radiologists, patients, regulators, and model builders may gain different benefits, but also face varied risks depending on the degree of interpretability. For instance, radiologists may benefit from improved accuracy in diagnosis and treatment planning, increased confidence in the decisions made by the model, and time-saving by reducing the need for manual interpretation of imaging data. However, they will face risks in relation to reduced control over the decision-making process, which could lead to decreased trust in the model, increased workload in terms of learning and adapting to new models, and liability concerns if the model's decision-making process is not transparent. Patients will benefit from faster and more accurate diagnoses, leading to

improved treatment outcomes, a reduced need for invasive procedures (especially in the case of false negatives), as well as greater transparency in the decision-making process, leading to increased trust in the medical system. They also face risks in relation to reduced human interaction with healthcare providers, which could lead to decreased patient satisfaction, privacy concerns if sensitive patient data are not adequately protected, and misdiagnosis or inappropriate treatment if the model's decision-making process is not transparent. Regulators can benefit from increased efficiency in the regulatory process, leading to faster approvals and reduced costs, improved public health outcomes through the use of accurate and reliable ML models, and greater transparency and accountability in the decision-making process. At the same time, regulators may face liability concerns if the model's decision-making process is not transparent, reduced oversight and control over the decision-making process, which could lead to decreased trust in the regulatory system, and privacy concerns if sensitive data is not adequately protected. Finally, ML model builders can benefit from increased efficiency in the model-building process, leading to faster development and deployment, reduced costs through the use of automated processes, and improved accuracy and reliability of the models, leading to increased demand for their services. They will, though, be faced with reduced control over the decision-making process, which could lead to decreased trust in the model, liability concerns if the model's decision-making process is not transparent, and potential backlash if the model's decision-making process is perceived as biased or unfair.

Evidently, satisfying all stakeholders and meeting all their expected needs, while mitigating the associated risks, is an impossible task. A balanced approach that takes into account the needs and expectations of all stakeholders involved is necessary to ensure that the benefits outweigh the risks. A balanced approach would involve considerations around transparency and interpretability (including assessing the trade-offs with performance), fairness and bias (including mitigation strategies for bias detection and reduction), ethics (including protection of patient privacy), and equitable collective action (including involvement of all stakeholders in the training, validation, and deployment of ML models).

Second, and related to the first question: how can individual organizations with limited AI resources achieve a balanced approach to AI interpretability? Would they always have to conform to standards enforced by more resourceful organizations with more technological power? This is a critical question that takes into consideration the recent surge of generative AI technologies that are becoming more diffused across service sectors. Certainly, individual organizations such as hospitals with limited AI resources can achieve a balanced approach to AI interpretability by leveraging existing best practices, tools, and frameworks developed by more resourceful organizations. They can also collaborate with other organizations, experts, and researchers in the field to share knowledge, resources, and expertise. For example, there are a variety of open-source interpretability frameworks and tools available, such as LIME, SHAP, and Captum, which can help organizations to understand and interpret the decision-making process of their ML models. These frameworks can be implemented relatively easily and can help organizations to

achieve a higher level of interpretability without investing significant resources. Additionally, organizations can take steps to improve the quality and diversity of their data, which can help to mitigate potential biases and improve the overall accuracy and interpretability of their ML models. This can involve collecting data from a variety of sources, using data augmentation techniques to increase the size of their training dataset, and regularly auditing their data for biases. While it may be challenging for smaller organizations to conform to the same standards enforced by more resourceful organizations, it is still possible for them to achieve a balanced approach to AI interpretability through a combination of best practices, open-source tools, and collaboration with other stakeholders. By working together and sharing resources and knowledge, organizations can improve their AI capabilities and ensure that their models are accurate, trustworthy, and fair.

At the same time, less resourceful organizations in healthcare will face several key challenges when seeking to deploy interpretable AI systems in decision-making. First, they may not have the same level of financial or technological resources as larger organizations, making it difficult for them to invest in expensive AI tools, infrastructure, and ML model builders. Developing and deploying ML models requires specialized expertise, including data science, machine learning, and software engineering, as we discussed in earlier sections. Less resourceful organizations may not have access to this expertise, or may struggle to attract and retain talented individuals. In addition, building accurate and interpretable ML models requires high-quality, diverse, and representative data. Smaller organizations such as individual hospitals may struggle to access large and diverse datasets, or may not have the resources to ensure that their data are of sufficient quality. Finally, deploying AI systems in healthcare requires careful consideration of regulatory and ethical frameworks, including patient privacy, informed consent, and fairness. Smaller organizations may struggle to navigate these frameworks, or may not have the resources to implement them effectively.

Technologically powerful organizations, including big tech companies such as Google, Amazon, and Apple, can dominate and control how fairness, equity, transparency, and accountability are defined by using their financial and technological resources to shape the conversation and set standards. For example, these organizations may invest heavily in developing proprietary AI systems and frameworks, which can make it difficult for smaller organizations to compete or participate in the development of industry-wide standards. They may also have the power to influence regulatory and policy decisions, which can impact the adoption and deployment of AI systems in healthcare. Furthermore, these organizations may have access to large and diverse datasets, which can give them an advantage in developing more accurate and interpretable AI systems. This can lead to a concentration of power and influence in the hands of a few dominant organizations, which can limit competition, innovation, and the development of more equitable and fair AI systems. Further investigation in the concentration of control over the development of interpretable AI systems in healthcare is a much-needed area of research.

In conclusion, in this chapter, we reviewed the literature on medical imaging and developed a framework of interpretable AI systems in enabling the diagnostic process. We identified the possible tensions that may emerge as human agents work with ML models and data, and explored how these tensions may impact the performance of interpretable AI systems in the diagnostic process. We concluded by raising a set of critical questions for further research into the design and policy around the development of interpretable AI systems in healthcare.

REFERENCES

- Act (1996). Health Insurance Portability and Accountability Act of 1996. *Public Law*, 104, 191.
- Adadi, A., and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160. <https://doi.org/10.1109/access.2018.2870052>.
- Akselrod-Ballin, A., Chorev, M., Shoshan, Y., Spiro, A., Hazan, A., Melamed, R., Barkan, E., Herzel, E., Naor, S., Karavani, E., Koren, G., Goldschmidt, Y., Shalev, V., Rosen-Zvi, M., and Guindy, M. (2019). Predicting breast cancer by applying deep learning to linked health records and mammograms. *Radiology*, 292(2), 331–342. <https://doi.org/10.1148/radiol.2019182622>.
- Alaimo, C., and Kallinikos, J. (2021). Managing by data: algorithmic categories and organizing. *Organization Studies*, 42(9), 1385–1407. <https://doi.org/10.1177/0170840620934062>.
- Alvarado, R. (2022). Should we replace radiologists with deep learning? Pigeons, error and trust in medical AI. *Bioethics*, 36(2), 121–133. <https://doi.org/10.1111/bioe.12959>.
- American College of Radiology and D’Orsi, C. (2018). *ACR BI-RADS Atlas: Breast Imaging Reporting and Data System: 2013*. American College of Radiology.
- Anthony, C., Bechky, B.A., and Fayard, A.-L. (2023). “Collaborating” with AI: taking a system view to explore the future of work. *Organization Science*. <https://doi.org/10.1287/orsc.2022.1651>.
- Baird, A., and Maruping, L.M. (2021). The next generation of research on IS use: a theoretical framework of delegation to and from agentic IS artifacts. *MIS Quarterly*, 45(1), 315–341. <https://doi.org/10.25300/misq/2021/15882>.
- Barnett, A.J., Schwartz, F.R., Tao, C., Chen, C., Ren, Y., Lo, J.Y., and Rudin, C. (2021). A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nature Machine Intelligence*, 3(12), 1061–1070. <https://doi.org/10.1038/s42256-021-00423-x>.
- Belle, V., and Papantonis, I. (2021). Principles and practice of explainable machine learning. *Front Big Data*, 4, 688969. <https://doi.org/10.3389/fdata.2021.688969>.
- Bhatt, U., Weller, A., and Moura, J.M. (2020a). Evaluating and aggregating feature-based model explanations. *arXiv preprint arXiv:2005.00631*.
- Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J.M. F., and Eckersley, P. (2020b). Explainable machine learning in deployment. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.
- Blanks, R., Wallis, M., and Moss, S. (1998). A comparison of cancer detection rates achieved by breast cancer screening programmes by number of readers, for one and two view mammography: results from the UK National Health Service breast screening programme. *Journal of Medical Screening*, 5(4), 195–201. <https://doi.org/10.1136/jms.5.4.195>.

- Brennen, A. (2020). What do people really want when they say they want “explainable AI?” We asked 60 stakeholders. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*.
- Chen, Y., James, J.J., Michalopoulou, E., Darker, I.T., and Jenkins, J. (2023). Performance of radiologists and radiographers in double reading mammograms: the UK National Health Service breast screening program. *Radiology*, 306(1), 102–109. <https://doi.org/10.1148/radiol.212951>
- Daley, B., Ratul, Q.E.A., Serra, E., and Cuzzocrea, A. (2022). GAPS: generality and precision with Shapley attribution. *2022 IEEE International Conference on Big Data (Big Data)*.
- Das, A., and Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (XAI): a survey. *arXiv preprint arXiv:2006.11371*.
- Dong, F., She, R., Cui, C., Shi, S., Hu, X., Zeng, J., Wu, H., Xu, J., and Zhang, Y. (2021). One step further into the blackbox: a pilot study of how to build more confidence around an AI-based decision system of breast nodule assessment in 2D ultrasound. *Eur Radiol*, 31(7), 4991–5000. <https://doi.org/10.1007/s00330-020-07561-7>.
- Doshi-Velez, F., and Kim, B. (2017a). Towards a rigorous science of interpretable machine learning. *arXiv preprint, arXiv:1702.08608*.
- Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O’Brien, D., Scott, K., Schieber, S., Waldo, J., and Weinberger, D. (2017b). Accountability of AI under the law: the role of explanation. *arXiv preprint arXiv:1711.01134*.
- El Adoui, M., Drisis, S., and Benjelloun, M. (2020). Multi-input deep learning architecture for predicting breast tumor response to chemotherapy using quantitative MR images. *Int J Comput Assist Radiol Surg*, 15(9), 1491–1500. <https://doi.org/10.1007/s11548-020-02209-9>.
- Fügener, A., Grahl, J., Gupta, A., and Ketter, W. (2021). Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *MIS Quarterly*, 45(3), 1527–1556. <https://doi.org/10.25300/misq/2021/16553>.
- Fuhrman, J.D., Gorre, N., Hu, Q., Li, H., El Naqa, I., and Giger, M. L. (2022). A review of explainable and interpretable AI with applications in COVID-19 imaging. *Med Phys*, 49(1), 1–14. <https://doi.org/10.1002/mp.15359>.
- Gohel, P., Singh, P., and Mohanty, M. (2021). Explainable AI: current status and future directions. *arXiv preprint arXiv:2107.07045*.
- Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., and Lee, S. (2019). Counterfactual visual explanations. *International Conference on Machine Learning*.
- Graziani, M., Andrearczyk, V., and Müller, H. (2018). Regression concept vectors for bidirectional explanations in histopathology. Understanding and Interpreting Machine Learning in Medical Image Computing Applications: First International Workshops, MLCN 2018, DLF 2018, and iMIMIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16–20, 2018, Proceedings 1, 124–132.
- Gromet, M. (2008). Comparison of computer-aided detection to double reading of screening mammograms: review of 231,221 mammograms. *AJR Am J Roentgenol*, 190(4), 854–859. <https://doi.org/10.2214/AJR.07.2812>.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 1–42. <https://doi.org/10.1145/3236009>.
- Harned, Z., Lungren, M.P., and Rajpurkar, P. (2019). Machine vision, medical AI, and malpractice. *Harvard Journal of Law and Technology Digest*. <https://ssrn.com/abstract=3442249>.
- Harvey, S.C., Geller, B., Oppenheimer, R.G., Pinet, M., Riddell, L., and Garra, B. (2003). Increase in cancer detection and recall rates with independent double interpretation of screening mammography. *American Journal of Roentgenology*, 180(5), 1461–1467.

- Hepenstal, S., and McNeish, D. (2020). Explainable artificial intelligence: What do you need to know? *International Conference on Human-Computer Interaction*.
- Hohman, F.M., Kahng, M., Pienta, R., and Chau, D.H. (2018). Visual analytics in deep learning: an interrogative survey for the next frontiers. *IEEE Trans Vis Comput Graph*. <https://doi.org/10.1109/TVCG.2018.2843369>.
- Hong, S.R., Hullman, J., and Bertini, E. (2020). Human factors in model interpretability: industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1), 1–26. <https://doi.org/10.1145/3392878>.
- IARC (2021). World Cancer Day 2021: Spotlight on IARC research related to breast cancer. International Agency for Research on Cancer. Retrieved 26/06/2021 from <https://www.iarc.who.int/featured-news/world-cancer-day-2021/#:~:text=According%20to%20recent%20global%20cancer%20estimates%20from%20the,cases%20of%20lung%20cancer%20for%20the%20first%20time>.
- Ji-Ye Mao, I.B. (2015). The use of explanations in knowledge-based systems: Cognitive perspectives and a process-tracing analysis. *Journal of Management Information Systems*, 17(2), 153–179. <https://doi.org/10.1080/07421222.2000.11045646>.
- Jin, D., Sergeeva, E., Weng, W.H., Chauhan, G., and Szolovits, P. (2022). Explainable deep learning in healthcare: a methodological survey from an attribution view. *WIREs Mechanisms of Disease*, 14(3), e1548. <https://doi.org/10.1002/wsbm.1548>.
- Jones, M. (2019). What we talk about when we talk about (big) data. *Journal of Strategic Information Systems*, 28(1), 3–16. <https://doi.org/10.1016/j.jsis.2018.10.005>.
- Kahneman, D., Sibony, O., and Sunstein, C.R. (2021). *Noise: A Flaw in Human Judgement*. Hachette UK.
- Kamiran, F., and Calders, T. (2009). Classifying without discriminating. *2009 2nd International Conference on Computer, Control and Communication*.
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., and Viegas, F. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). *International Conference on Machine Learning*.
- Kim, S.T., Lee, H., Kim, H.G., and Ro, Y.M. (2018). ICADx: interpretable computer aided diagnosis of breast masses. *Medical Imaging 2018: Computer-Aided Diagnosis*, 10575, 450–459.
- Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sasing, A., and Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, 296. <https://doi.org/10.1016/j.artint.2021.103473>
- Liao, Q.V., Gruen, D., and Miller, S. (2020). Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.
- Lipton, Z.C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3), 31–57.
- Lundberg, S.M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In the *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, pp. 1–10, Long Beach, CA, USA. <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>.
- Mitchell, M. (2019). *Artificial Intelligence: A Guide for Thinking Humans*. Penguin UK.
- Modarres, C., Ibrahim, M., Louie, M., and Paisley, J. (2018). Towards explainable deep learning for credit lending: a case study. *arXiv preprint arXiv:1811.06471*.
- Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73, 1–15. <https://doi.org/10.1016/j.dsp.2017.10.011>.
- Monteiro, E., and Parmiggiani, E. (2019). Synthetic knowing: the politics of the internet of things. *MIS Quarterly*, 43(1), 167–184.

- Moreira, I.C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M.J., and Cardoso, J.S. (2012). INbreast: toward a full-field digital mammographic database. *Acad Radiol*, 19(2), 236–248. <https://doi.org/10.1016/j.acra.2011.09.014>
- Patrício, C., Neves, J.C., and Teixeira, L.F. (2022). Explainable deep learning methods in medical diagnosis: a survey. *arXiv preprint arXiv:2205.04766*.
- Preece, A., Harborne, D., Braines, D., Tomsett, R., and Chakraborty, S. (2018). Stakeholders in explainable AI. *arXiv preprint arXiv:1810.00184*. <https://doi.org/arXiv:1810.00184>.
- Rai, A. (2020). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science*, 48(1), 137–141. <https://doi.org/10.1007/s11747-019-00710-5>.
- Rai, A., Constantinides, P., and Sarker, S. (2019). Next generation digital platforms: toward human–AI hybrids. *MIS Quarterly*, 43, iii–ix.
- Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. (2020). Closing the AI accountability gap. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.
- Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). “Why should I trust you?” Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>.
- Samek, W., Wiegand, T., and Müller, K.-R. (2017). Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.
- Shen, Y., Shamout, F.E., Oliver, J.R., Witowski, J., Kannan, K., Park, J., Wu, N., et al. (2021). Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. <https://doi.org/10.1101/2021.04.28.21256203>.
- Taylor, P., and Potts, H.W. (2008). Computer aids and human second reading as interventions in screening mammography: two systematic reviews to compare effects on cancer detection and recall rate. *Eur J Cancer*, 44(6), 798–807. <https://doi.org/10.1016/j.ejca.2008.02.016>.
- Thiebes, S., Lins, S., and Sunyaev, A. (2020). Trustworthy artificial intelligence. *Electronic Markets*, 31(2), 447–464. <https://doi.org/10.1007/s12525-020-00441-4>.
- Tjoa, E., and Guan, C. (2021). A survey on explainable artificial intelligence (XAI): toward medical XAI. *IEEE Trans Neural Netw Learn Syst*, 32(11), 4793–4813. <https://doi.org/10.1109/TNNLS.2020.3027314>.
- Tomsett, R., Braines, D., Harborne, D., Preece, A., and Chakraborty, S. (2018). Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *arXiv*. <https://doi.org/arXiv:1806.07552>.
- Tonekaboni, S., Joshi, S., McCradden, M.D., and Goldenberg, A. (2019). What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End use. *Proceedings of the 4th Machine Learning for Healthcare Conference, Proceedings of Machine Learning Research*. <http://proceedings.mlr.press>.
- Trocin, C., Mikalef, P., Papamitsiou, Z., and Conboy, K. (2021). Responsible AI for digital health: a synthesis and a research agenda. *Information Systems Frontiers*. <https://doi.org/10.1007/s10796-021-10146-4>.
- Van Den Broek, E., Levina, N., and Sergeeva, A. (2022). In pursuit of data: negotiating data tensions between data scientists and users of AI tools. *Academy of Management Proceedings*, 2022(1). <https://doi.org/10.5465/ambpp.2022.182>.
- van der Velden, B.H.M., Janse, M.H.A., Ragusi, M.A.A., Loo, C.E., and Gilhuijs, K.G.A. (2020). Volumetric breast density estimation on MRI using explainable deep learning regression. *Sci Rep*, 10(1), 18095. <https://doi.org/10.1038/s41598-020-75167-6>.

- van der Velden, B.H.M., Kuijf, H., Gilhuijs, K.G.A., and Viergever, M.A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Med Image Anal*, 79, 102470. <https://doi.org/10.1016/j.media.2022.102470>.
- Wang, C., Li, J., Zhang, F., Sun, X., Dong, H., Yu, Y., and Wang, Y. (2021). Bilateral asymmetry guided counterfactual generating network for mammogram classification. *IEEE Trans Image Process*, 30, 7980–7994. <https://doi.org/10.1109/TIP.2021.3112053>.
- Wang, H., Yao, Y., and Salhi, S. (2020). Tension in big data using machine learning: analysis and applications. *Technological Forecasting and Social Change*, 158, 120175.
- Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R.M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Weller, A. (2019). Transparency: motivations and challenges. In Samek, W., Montavon, G., Vedaldi, A., Hansen, L., Müller, K.R. (eds), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (pp. 23–40). Springer, Cham. https://doi.org/10.1007/978-3-030-28954-6_2.
- Wickramanayake, S., Hsu, W., and Lee, M.L. (2021). Explanation-based data augmentation for image classification. *Advances in Neural Information Processing Systems*, 34, 20929–20940.
- Willeminck, M.J., Koszek, W.A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L.R., Summers, R.M., Rubin, D.L., and Lungren, M.P. (2020). Preparing medical imaging data for machine learning. *Radiology*, 295(1), 4–15. <https://doi.org/10.1148/radiol.2020192224>.
- Ye, L.R., and Johnson, P.E. (1995). The impact of explanation facilities on user acceptance of expert systems advice. *MIS Quarterly*, 19(2), 157–172.
- Yu, R., and Shi, L. (2018). A user-based taxonomy for deep learning visualization. *Visual Informatics*, 2(3), 147–154. <https://doi.org/10.1016/j.visinf.2018.09.001>.
- Zarsky, T.Z. (2013). Transparent predictions. *University of Illinois Law Review*, 4, 1503–1570. <https://illinoislawreview.org/print/volume-2013-issue-4/transparent-predictions/>.

15. Artificial intelligence to support public sector decision-making: the emergence of entangled accountability

Francesco Gualdi and Antonio Cordella

INTRODUCTION

Public organizations have increasingly adopted artificial intelligence (AI) to inform the decision-making processes that underpin public services design and delivery (Ammitzbøll Flügge et al., 2021; Strich et al., 2021). Scholars have provided evidence for the positive outcomes generated by AI support to public decision-making, such as the rationalization of the administrative workflow and the provision of improved public services (Misuraca et al., 2020). However, recent works have also explained the unexpected outcomes generated by the adoption of AI to inform the decision-making processes of public organizations, and hence the services they provide (Medaglia et al., 2021; Wirtz et al., 2020). Following recent controversial cases of AI adoption in the public sector, civil society and public audiences have raised calls for further scrutiny of the use of AI to inform public services design and delivery (de Bruijn et al., 2021; Grimmelikhuijsen, 2023). Academic research has echoed these concerns by showing increasing attention to the impact AI has on decision-making processes in public organizations (Giest and Klievink, 2022; Lorenz et al., 2021). In line with this research, we aim to shed light on how AI logic permanently alters public organizations' decision-making process, and how these transformations generate consequences for the accountability of public organizations (Busuioc, 2021).

To capture the nuanced effects generated by AI on accountability, it is necessary to explain their very specific characteristics that determine organizational, legal, and institutional transformations in the public decision-making processes, and hence on services design and delivery. Unlike other technological systems, AI standardization introduces a degree of opacity in the workflows it mediates (Burrell, 2016; Leonardi and Treem, 2020). AI reasoning does not allow human actors to properly understand and explain the AI functions (Zhang et al., 2021) that inform the decision-making processes (Strich et al., 2021).

The research builds on the findings from two cases where public organizations have adopted AI to inform decision-making processes: the United Kingdom's UKVI system for issuing visas, and the *Loomis v Wisconsin* judicial case in the United States. We aim to demonstrate that the deployment of AI tools to support the decision-making process, and hence the design and delivery of key public services,

has changed the logic that governs the administrative workflows (Bullock et al., 2020). Accordingly, the decision-making process has been fundamentally altered by AI. We posit that human actors have neither the literacy to understand, nor the possibility to challenge, the inputs received by the AI instruments they rely upon (Lebovitz et al., 2022). As the two cases illustrate, the service design and delivery informed by AI has generated highly controversial outcomes that impacted the life of citizens.

This raises crucial questions about the accountability of the decision-making process informed by AI in public organizations. The findings from the two cases show that it is not possible to completely disentangle the contribution of the AI from the contribution of human actors to the decision-making process. AI releases inputs upon which human actors build their discretionary judgment. By so doing, a new set of interdependences between humans and AI emerges in the organizational settings where decision-making processes happen. Hence, we posit that to properly understand how to hold public organizations accountable for the services they design and deliver by building on AI, it is necessary to go beyond siloed understandings of accountability. Human actors cannot be held accountable alone, because their decisions are fundamentally altered by AI opaque inputs. AI cannot be held accountable alone, since the final decision is always taken by humans. Accordingly, to fully appreciate the emergence of entanglements between the humans and the machine, it is necessary to better theorize accountability. The research makes the case for a reconceptualization of accountability which encompasses the whole entanglement generated by AI adoption that characterizes the decision-making process of public organizations.

BACKGROUND

Research has provided evidence for the transformations generated in the public sector since the adoption of AI (Charles et al., 2022; Lebovitz et al., 2022; Medaglia et al., 2021; Wirtz et al., 2021). Specifically, AI is currently used to inform the way in which public decision-makers formulate decisions within public organizations (Pencheva et al., 2020). Although it has been demonstrated that public managers and policymakers find useful the support of AI to reach a better-informed decision-making (Criado et al., 2020), increasing challenges and controversies arise from the use of AI in the public sector (Dwivedi et al., 2019; Favaretto et al., 2019; Sun and Medaglia, 2019); challenges that include—but are not limited to—discrimination, biased decisions, and ethical issues (Wirtz et al., 2020). Since public organizations increasingly rely on AI, decisions that are informed by AI have a direct impact on the services that public administrations provide to citizens. Key public services such as issuing visas, providing subsidies, administering justice, and executing law enforcement are often delivered on the basis of judgments that are informed—in different ways—by AI (Wirtz and Müller, 2019).

Public sector decision-making has a long history of adopting technologies to support decision-making processes that enable the delivery of public services

(Schwarz et al., 2022; Simon, 2013). Seminal contributions have focused on the transformations generated in public organizations by the deployment of information and communication technologies (ICTs) (Bovens and Zouridis, 2002; Fountain, 2004). However, AI poses a different challenge because of the nature of its design and processing. The AI is fed with a huge amount of data (inputs), it processes the data, and releases specific information (outputs) upon which decision-makers can structure their choices. Yet what happens when the AI processes the data remains to some extent obscure for the human actors who are required to interact with AI (Bullock et al., 2020; Busch and Henriksen, 2018). Burrell has illustrated an additional layer of opacity that is introduced in the decision-making as a consequence of the AI-mediated provision of information (Burrell, 2016).

AI formalizes and structures decision-making processes by reducing the degree of transparency of what happens within AI (de Bruijn et al., 2021). It becomes increasingly difficult for the human actors involved in the decision-making process to make sense of the way by which AI works (Zhang et al., 2021). AI reasoning is often beyond the limits of human comprehension: actors simply do not understand what happens in the black box (Stohl et al., 2016). Since AI is profoundly reshaping the way in which public organizations design and deliver services (Medaglia et al., 2021), additional attention is needed on the way by which it impacts on the configuration of public organizations (Giest and Klievink, 2022), on design and delivery of public services (Meijer et al., 2021), and on the legal and institutional norms that structure organizations (Gualdi and Cordella, 2023).

To address these challenges, scholars have focused on how it is possible to hold AI accountable (Busuioc, 2021; Kroll et al., 2017; Martin, 2019). Accountability is one of the paramount ways in which those who are ruled can exercise control over rulers' actions and decisions. One of the most widely accepted definitions of accountability is provided by Mark Bovens. Bovens argues that accountability is "a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences" (Bovens, 2007, p. 447). Bovens's definition of accountability is relevant because it focuses on the relationship between the actor and the audience—see also Wieringa (2020)—and it depicts accountability as a process. Although Bovens does not directly discuss technology in his work (Bovens, 2007), his comprehensive and dynamic understanding of accountability is valuable to investigate the impact of technology on public organizations. Moreover, framing accountability as a process allows us to properly discuss the interactions that take place between the humans and the AI, and to spot potential distortions, gaps, and misalignments that characterize the decision-making process.

Building on Bovens's argument, recent contributions have focused on the interactions that happen between human actors and AI in order to provide a more nuanced understanding of accountability in public organizations' decision-making processes (Busuioc, 2021; Diakopoulos, 2016; Meijer and Grimmelikhuijsen, 2020; Wieringa, 2020). For example, Meijer and Grimmelikhuijsen (2020) root their understanding of AI accountability in two main elements, justification and explanation, where jus-

tification means that public organizations are required to provide reasons for the use of AI, and explanation means that the outcomes of the AI-informed decision-making must be carefully clarified (Meijer and Grimmelikhuijsen, 2020, p. 60). Busuioc (2021) argues that accountability is about answerability, which means that if the public sector wants to make its decision-making process fully accountable, it must remove secrecy on the AI adopted. Diakopoulos (2016) posits that to achieve accountability purposes it is necessary to start a process of disclosure that encompasses both the human and the technical element. In other words, there is a need to release key information on the human activities (data collection, human biases, inferences with the machine) as well as on the machine activities (explaining how the AI works and how it is built) (Diakopoulos, 2016).

We align with these calls for a more comprehensive and articulated view of accountability of AI-mediated decision-making in public organizations. However, although relevant, the contributions we have examined offer only a partial understanding of how to hold accountable the decision-making process constituted by both human and AI agency. Many of the recent contributions in fact focus on the simple juxtaposition of the human element with the technological element, as if their interaction to structure the decision-making happens through a parallel or multi-level negotiation that produces a “mixed decision-making” (Busuioc, 2021, p. 828). We argue that this view of the human–AI interaction is not enough to describe the interdependences that take place when AI mediates decision-making processes in public organizations. Against this background, we posit that a deeper interaction takes place when AI supports decision-making. Accordingly, a study of AI accountability must reflect the permanent transformations generated in the decision-making process impacted by the AI. To shed light on these transformations, it is necessary to account for how the AI entangles with the existing elements and logics that regulate decision-making processes in public organizations.

The Emergence of Techno-Legal Entanglements

Before the adoption of ICT systems (let alone AI), decision-making processes in the public sector happened as prescribed by a set of formal rules and norms, within which boundaries the exercise of discretion by human decision-makers took place (March and Olsen, 2010; Mintzberg, 1983). Accordingly, accountability focused only on the human decision-making. Formal regulations (Bovens, 2007) prescribed who should be held accountable, how, and to which audience.

However, AI adoption has profoundly redesigned the decision-making processes in public organizations (Ammitzbøll Flügge et al., 2021; Strich et al., 2021). The deployment of AI to mediate public sector decision-making introduces another source of formalization that imbricates with existing legal and administrative norms (Giest and Klievink, 2022). Legal and administrative norms are no longer the sole logic governing public organizations’ decision-making processes: decision-making is in fact regulated by legal and administrative logic, and also by the logic of AI. This happens because the technological formalization introduced by the AI does not

only restructure existing workflows and interactions. The AI goes beyond it: it alters workflows and interactions through its technological properties and functionalities (Lorenz et al., 2021). To be compatible with the context of adoption, AI structures the rules of public organizations' decision-making into the functional sequences and interdependences proper of the machine (von Krogh, 2018). Multiple logics govern the decision-making process in the public sector: the logic of laws and norms of the public administration, and the logic of technology (Bovens and Zouridis, 2002). If these logics collide, the whole system is incapable of producing decisions, and hence is useless (Lanzara, 2009). If these logics simply juxtapose, the risk of misalignment is high, and the system risks dysfunctionality (Lanzara, 2009). Hence, the logic of the AI must find a way to properly entangle with the existing logics that regulate public organizations' decision-making.

The entanglement between the logic of the AI and the logic of the laws and norms which regulate the public organizations transforms the decision-making process. Organizational decision-making processes are now the outcome of a negotiation between logics that possess different yet complementary formalization characteristics. Specifically, the decisions taken after the adoption of AI are neither solely a product of AI logic, nor solely a product of the administrative logic. These decisions are the product of an entangled logic that combines elements of both administrative/legal and technological dimensions. They constitute a techno-legal entanglement.

The change in the logic that drives decision-making also impacts the outcomes of the decision-making, namely the services provided by the public organization (Bovens and Zouridis, 2002). Since AI transforms the decision-making process, it is necessary to adopt a nuanced understanding of how a decision-making process informed by AI can be held accountable. Accordingly, a proper conceptualization of accountability is needed. Building on the work by Bovens (2007), we aim to extend and complement his definition of accountability by shedding light on the entangled nature of accountability that emerges after the adoption of AI in public organizations' decision-making. Specifically, we theorize the emergence of a new type of accountability that: (1) is highly formalized; and (2) includes not only the humans and the AI, but also the entangled legal and algorithmic logic that structures new interdependences between the humans and AI. Hence, we advocate for a more nuanced approach that focuses on the entanglement, and we build our argument for a new conceptualization of "entangled accountability" that better describes the complex interactions of humans and AI to inform the decision-making process of public organizations.

AI ADOPTIONS IN PUBLIC ORGANIZATIONS: CONTROVERSIAL CASES

The adoption of AI to inform public organizations' decision-making has increased in recent years and has spread across multiple domains of public administration. Through the deployment of AI, public managers expect to streamline workflows and, by so doing, to achieve a more efficient and faster decision-making process.

AI elaborates vast amounts of data and provides additional information to human actors who take decisions on services design and delivery. Despite this promising purpose, several cases of distortion emerged after the adoption of AI to inform public decision-making; public service provision that was built upon AI mediation has been recurrently challenged and contested (de Bruijn et al., 2021). Most significantly, a clear and straightforward line of responsibility and accountability has not been drawn, leaving the public audience with increasing skepticism about the trustworthiness of AI-mediated decision-making processes (Grimmelikhuijsen, 2023). In this section we build on the findings of two selected examples that focus on AI adoption to inform public organizations' decision-making. The examples offer evidence to define a more nuanced understanding of accountability against the background of the techno-legal entanglements that impact decision-making processes in public organizations.

Home Office UKVI

The United Kingdom (UK) Home Office increasingly relied on AI to manage the issuing of visas to immigration applicants. When immigrants apply for a visa, they are required to provide several pieces of information to the Home Office Visas and Immigration (UKVI) department, such as personal details and key socio-economic information. Before the adoption of AI, the information was processed manually by officers in the UKVI to assess whether citizens had the right to apply for the visa, and whether there were any obstacles to issuing the visa. Then, since 2015, the UKVI deployed an AI system with the purpose of streamlining the decision-making process to issue visas. The motivation for the adoption of AI was to increase speed and efficacy in the whole decision-making process. The AI system classified the visa applications according to how much of a potential source of risk to the British society an applicant was perceived to be. Three categories were created: "low risk" (labelled in green), "medium risk" (amber), "high risk" (red).

The AI system was fed with data that included key information about applicants' history and background, such as age, nationality, ethnicity. The AI system benchmarked the data received against historical datasets and statistics to assess the applicant's risk rating. The output of AI was then submitted to officers who examined the application in conjunction with the risk rating, and then they decided whether to issue the visa or not.

The final decision was always taken by a human actor. However, data show that human actors' decisions were very likely to be in accordance with the AI assessment about the applicant's risk rating. Specifically, 96.36 percent of applications labelled as low risk (green) were successful; 81.08 percent of medium risk (amber) applications were accepted; and only 48.59 percent of high risk (red) applications ended up with a visa issued. The AI output was basically unchallenged for green applications, substantially unchallenged for amber ones, and was further scrutinized only in the case of red applications (Bolt, 2017).

Increasing attention on the effects of the AI deployment soon emerged, and generated concern amongst the public audience. The AI had been adopted with the purpose of streamlining the decision-making process to increase efficiency and reduce the backlog in public administration. Hence, public managers in charge of overseeing the decision-making process at the UKVI set ambitious daily targets for case workers, who were required to produce no less than 75 decisions on low risk applications, 35 on medium risk, and 25 on high risk. However, it has been documented that this pressure often forced decision-makers not to challenge the AI output, so that they could achieve the volume of decisions expected of them (Bolt, 2017). Beyond efficiency issues, a problem with potential discriminations and distortions has been raised by several actors within the public audience. Critics of the AI adoption built a case claiming that the algorithm was biased against specific ethnic groups and nationalities, for which receiving a “low risk” evaluation was less likely than for others (Threipland and Rickett, 2020). The official inquiry which evaluated the UKVI system found that the streamlining tool was subject to confirmation bias, and that the whole system needed to be revised (Bolt, 2017). However, the UK government ruled out the possibility of either disclosing the algorithm or revising the policy, although it repeatedly denied accusations of perpetrating discrimination (McDonald, 2019). Moreover, the government made it clear that the adoption of the AI system had the unique purpose of streamlining the decision-making process, and that the final decision always remained with officers (House of Commons, 2019). Consequently, the government preferred not to disclose further information about the algorithm—such as key details on weights and variables—and terminated the programme.

COMPAS

United States judiciary systems have frequently utilized profiling algorithms in criminal trials. AI systems were adopted to inform sentencing decisions about convicted felons with the purpose of reducing bias and distortions. However, growing skepticism has surrounded the adoption of AI in the context of judiciary systems, because of the increasing claims that AI reproduces biased values which disproportionately penalize specific groups of people (Ávila et al., 2020).

One of the most well-known and controversial cases is the *State of Wisconsin vs Loomis* legal dispute (Liu et al., 2019). Mr Loomis got involved in a trial after an arrest for criminal conduct. He was sentenced to six years of detention. The sentence was not the outcome of the unique judgment of human actors (in this case, the judge who oversaw the trial): it soon emerged that an AI system called Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) had played a role in the definition of the sentence. Specifically, an algorithm processed Mr Loomis’s profile with the purpose of assessing how likely he was to potentially re-offend. This profiling activity took place after the arrest and provided additional information about Mr Loomis before a final decision regarding the sentence was made. COMPAS considered data provided by Mr Loomis when he was asked to fill out a questionnaire containing key socio-economic information about his background. The information

was then used to feed the algorithm, which benchmarked Mr Loomis's profile against a group of citizens whose socio-economic characteristics were similar to his. The elaboration of information provided by Mr Loomis was followed by the release of a "recidivism score," according to which Mr Loomis was classified as an individual with a high risk of recidivating and creating further problems for the society.

The case came to the attention of the public audience, thus generating further scrutiny, due to the decision of Mr Loomis to appeal against the sentence (*State v. Loomis*, 2016). Mr Loomis argued that the sentence he received had been directly influenced by the risk score elaborated by COMPAS (Freeman, 2016). The appeal challenged the validity of the system, pointing out two main limitations of COMPAS. The first one focused on the design and construction of the algorithm that carried the risk of perpetrating biases, since it profiled individuals against the benchmark of similar groups of citizens. The relevance of specific details such as ethnicity and gender were overemphasized, according to Mr Loomis. The second main limitation was procedural: Mr Loomis claimed that his request to disclose the algorithm to better understand how the weights and variables were selected was rejected because of the property rights pending on COMPAS.

The appeal was turned down by Wisconsin Supreme Court, which confirmed the correctness of the procedural AI assessment. However, the Supreme Court also downplayed the relevance of the AI, which was used only to inform the decision-making on the sentence, and not to replace it (*Harvard Law Review*, 2017). Regarding the second issue raised by the appellant, the Supreme Court denied Mr Loomis access to weights and variables utilized in the COMPAS algorithm because of property rights. However, in an additional commentary to the ruling, the Supreme Court advocated for a cautious utilization of AI to inform criminal sentencing (*Harvard Law Review*, 2017).

DISCUSSION

The findings from the selected cases account for different applications and adoptions of AI to support public organizations' decision-making process. Despite differences, both examples illustrate how public organizations rely on AI to inform decision-making that is vital to the design and delivery of specific public services. The examples cannot be compared, since they belong to different domains—the UKVI case to bureaucratic administration of immigration-related services, the COMPAS case to criminal justice—yet taken together they provide relevant insights to understand how AI entangles with existing logics that structure the decision-making process, and how they can be held accountable. Accordingly, this section will first focus on the transformations generated by AI in decision-making processes, then discuss how this impacts the theorization of accountability.

AI Entanglements

In both cases, public organizations rely on AI solutions to improve the profiling of citizens that underpins a final decision taken by human actors. This happens through a procedure made up of several steps. First, individuals are required to provide specific information about their background and personal characteristics. This information is standardized so that it can be understood by AI scripts and logic. The AI system, fed with individuals' data, reconstructs a profile of the individual, benchmarking the data received against statistical information retrieved from other databases. Eventually, the machine attaches to the profile it has reconstructed a specific classification (risk assessment) that is further evaluated by human actors who make the final decision.

In the UKVI case, the AI system benchmarks the data received by visa applicants against datasets collected from other governmental repositories. The AI system infers correlations between data—such as age, nationality, ethnicity—provided by a single visa applicant, and statistical patterns related to groups of individuals who have committed offences.

In the COMPAS case, the procedure is similar, and it builds on socio-economic indicators: personal history of the felon is benchmarked against different databases held by several public institutions and agencies (education providers, law enforcement agencies, welfare programs). The AI system benchmarks a citizen's data against collective statistical trends that show how individuals with specific background conditions (education, job career, family settings) are inclined to recidivate.

In both cases, the AI system is fed with data from very different sources. To release an output that supports decision-makers, the AI system needs to retrieve information from other databases, and it needs to process this information according to its reasoning. Weights and variables of the algorithms are designed to create causal connections between the individual's pieces of information and large groups' statistical patterns. To process data from different databases, and to infer causality, the AI system transforms the data it receives: individuals' backgrounds are broken down into pieces of information compatible with statistical patterns, and large groups' patterns are decontextualized and recombined to build a benchmark against which to assess individuals. The consequence is that the AI system generates completely new causal connections, that underpin the profiling activity. For instance, the profiling of Mr Loomis happens through the decontextualization of his personal data and the creation of new relationships between this data and the collective patterns. The output is a profile of Mr Loomis that is produced according to the standardizing logic of the AI system, whose functioning is obscure to human actors.

The AI profiling has a profound impact on the decision-making process it supports. Public officers rely on the output of the AI assessment to produce an informed decision about the service they provide. In principle, they are entitled to challenge AI output; in fact, they seldom do it because challenging the AI output would increase their workload, as the UKVI case illustrates (Bolt, 2017). The new decision-making process is no longer constituted by two distinct phases, with the first one elaborated

by the machine, and the second by the human actor. The AI (technological) logic entangles with the administrative (legal) logic to support the decision-making, and it is impossible to disentangle the two dimensions. The AI standardization impacts the decision-making process since it introduces an additional element of opaque formalization to the whole procedure. Before the adoption of AI systems, officers took decisions in a context characterized by three elements: (1) the contextualized information about single individuals that was collected and processed; (2) the given boundaries set by legal and administrative regulations; and (3) the discretionary use of their own judgment. After the adoption of the AI system, the entanglements that emerge completely modify the landscape of decision-making. First, AI processes a greater amount of information, whose nature is opaque to human actors because it is decontextualized and abstracted from the original settings. Second, boundaries of action are no longer traced only by the laws and the administrative norms: the new boundaries are set by the negotiation of the technology with the law and the norms. The AI system sets new formal standards to execute the tasks, without which it would be impossible to structure the decision-making. Third, officers' discretion is impacted by the output of the AI system. Findings from the selected examples show that, in both cases, the highest political and judicial authorities have provided assurance that the AI is only ancillary to human judgment. However, it remains questionable whether and how the human actors involved in the decision-making actually challenge the output they receive from the AI system. In the case of UKVI, for instance, "low risk" and "medium risk" applications were granted a visa with a substantial alignment to the AI system's output.

The decision-making process that leads to a final decision on the provision of public services is hence the outcome of the entanglement that emerges between AI (technological) logic and administrative (legal) logic. It now constitutes a techno-legal entanglement. In this transformed decision-making process, human actors do not rely exclusively on their judgment, building on the legal regulations in place: rather, they rely on decontextualized data that are restructured according to the opaque AI logic. By doing so, they accept that their discretion is bound by a logic which they do not fully understand, let alone have the tools and literacy to understand.

Entangled Accountability

The findings from the examples show that the decision-making processes in both cases, of UKVI and COMPAS, came under scrutiny after claims of discrimination were raised. Critics argued that the AI systems actually shaped the outcome of the decision-making process. Against this view, authorities in the UK and in the United States (US) rejected the claim: in both cases, the motivation built on the assumption that the AI system was only supposed to provide information upon which human actors took decisions. In other words, political authorities in the UK and judicial courts in the US did not neglect the role played by AI. Yet they considered AI to be hierarchically inferior to the human actors. The counterargument of the UK government and the Wisconsin Supreme Court is valuable because it considers AI, as with

many other technologies, to be subject to human judgement. If this could be true in principle, evidence from the examples shows that in practice this does not happen: the AI logic entangles with legal logic to shape a decision-making process in which the action of humans is far more constrained by the technology than by any legal or administrative logic. This happens because human actors: (1) do not understand how the AI system works; (2) cannot avoid considering the AI output in the formulation of the final decision; and (3) find it increasingly difficult to challenge the AI output. An overruling of the AI decision becomes complicated because it is very difficult to keep the two phases of decision-making (AI profiling and human judgment) separated, or at least hierarchically subordinated. Human actors' activity is constrained by the AI logic that adds its own technological formalization to the already existing formalization of the legal regulations that dictate how decision-making should happen in public organizations.

The findings from the two selected examples allow us to further discuss a conceptualization of accountability that reflects the emergence of techno-legal entanglements in public decision-making processes. When AI entangles with existing logics that regulate the decision-making, the structuration of the entanglement permanently alters the decision-making process. Hence, there is a need for a new theorization of accountability that encompasses these transformations.

Building on Bovens's (2007) work, we focus on the complex and articulated nature of accountability. In a traditional decision-making process, not mediated by AI, the key question would be: "Who is accountable?" However, in the AI-mediated decision-making, the AI system (releasing an output) and the public officer (elaborating decisions upon AI input) generate interdependences that go beyond the single technology or the human actor's knowledge. What really matters is how the AI and the human actors find a way to interact and co-exist in the organizational decision-making process. In fact, the interdependences between the AI logic and the legal logic are not fixed or permanent: rather, they develop to find a stable configuration that is not easy to achieve (Lanzara, 2009). Hence, instead of asking "Who is accountable?", a proper question would be "What is accountable?"

To answer the question, a new conceptualization of accountability is needed. In this chapter, we argue that to hold the AI-mediated decision-making process accountable it is necessary to acknowledge that decisions are outcomes of techno-legal entanglements that emerge in organizational settings. Accordingly, accountability must include not only (and generically) the AI system and the humans; rather, accountability should capture the formalized interdependences between human actors and AI that emerge when the decision-making process is restructured to reflect legal/administrative logic and technological logic. The former is reinforced by humans who take decisions according to prescriptions and norms. The latter is reinforced by AI whenever it is adopted to elaborate and process information. These logics entangle to reflect the new formalized interdependences that originate when AI is adopted in the decision-making.

The emergence of entanglements deserves further elaboration because it is a central point in the analysis. The legal/administrative and technological logics are

not simply juxtaposed, as if they execute parallel tasks in the same process. It could be argued, for instance, that humans collect data which the AI system processes. Yet, once the AI has released its output, it is not possible to trace back contributions to the decision-making process. Neither are the logics simply blended, which could suggest a mixed arrangement between different logics. The legal/administrative logic does not blend with the technological logic. They both formalize the tasks they execute: the administrative/legal logic prescribes how decisions should be taken; the technological logic standardizes information according to its scripts. Entanglements emerge when the different formalized logics interact up to the point that is impossible to unpack the single dimensions (Lanzara, 2009). The decision-making process which reflects the adoption of AI is the outcome of the emergence of the techno-legal entanglement. For this reason, a conceptualization of accountability must acknowledge the transformed nature of the decision-making process. We offer this concept of “entangled accountability” to describe the interdependences that emerge between the humans and the AI: these interdependences are highly formalized, yet obscure. They are formalized because the standardizing logic of the AI combines with the formalization logic of the law. And they are obscure because human actors cannot completely understand what happens within the AI. Therefore, neither the human alone, nor the machine alone, can bear siloed responsibility for the outcome of the decision-making process. Responsibility—upon which accountability is built (Meijer and Grimmelikhuijsen, 2020)—must be shared between the different actors that take part in the decision-making process. Entangled accountability illustrates why the whole process of decision-making should be held accountable, instead of focusing exclusively on one part or the other of the process. Holding accountable only human actors would be pointless, since they do not fully understand what happens inside the machine. Focusing only the AI—for instance, echoing calls to open the black box—could be useful, but it is not enough, since the AI system produces only one piece of information to support the decision-making process. This information is then further elaborated by human actors who take the final choice to deliver public services.

The concept of entangled accountability offers a new perspective to describe the AI-mediated decision-making process. The entangled accountability shifts the focus from the actors to the process: acknowledging that humans and AI are both involved, it goes beyond the effort to individuate who is responsible for what. The entangled accountability focuses on the transformation of the decision-making process, which is put at the center of the investigation. The decision-making process is now a techno-legal entanglement in which the nature of the interdependences between humans and AI has become more formalized and obscure. Accordingly, accountability should reflect this entanglement and consider the whole decision-making process as the subject to be held accountable to public opinion.

CONCLUSIONS AND FURTHER IMPLICATIONS

This chapter has built upon a dynamic definition of accountability (Bovens, 2007) to shed light on the effects of AI adoptions to inform public organizations' decision-making processes. The findings from two selected examples show that with the emergence of techno-legal entanglements, a different approach towards accountability is needed.

This chapter contributes to the research tradition that theorizes AI accountability. We posit that traditional definitions of accountability, or siloed approaches that focus either on the machine or on humans, are not enough to capture the profound transformations that impact the decision-making process when AI is adopted. We align with calls to further investigate the characteristics of accountability, and we contribute to this by showing that the accountability which emerges after the deployment of AI to inform public decision-making is entangled and highly formalized. Accountability is entangled because it is impossible to unpack the interdependences between the humans and AI that structure the decision-making process. In addition, accountability is highly formalized because the entanglements are the outcome of different formalizing logics: technological and legal/administrative logics.

The emergence of the entangled accountability carries profound implications for practice. Managers of public organizations should be aware of the effects of entanglements when they decide to adopt AI to support decision-making processes. Public organizations must accept that the emergence of entanglements impacts their organizational workflows and settings, which are no longer defined only by laws and administrative regulations, but also by the formal standardization of the technology. The emergence of techno-legal entanglements generated by AI adoption alters the decision-making process permanently. This permanent alteration redesigns the boundaries and practices of the way in which decisions are taken and public services delivered. Public managers need to be aware of the scale of these transformations, and encourage a different approach in the organizations they run. Far beyond a simple acceptance of the outputs of the AI system, public managers should ensure the creation of working environments in which human actors are conscious of the impact that AI has on the tasks they execute. Moreover, there are relevant implications for the wider public that is increasingly scrutinizing the AI adoptions. Although an investigation of the role of public opinion is out of the scope of this chapter, the emergence of entangled accountability is also a challenge for the audiences that aim to hold accountable the AI-mediated decision-making process. A new approach that goes beyond the siloed accountability—hence asking for the termination of AI-based programs or focusing exclusively on human responsibility—might not be enough to properly assess the impact of entangled accountability in public organizations. Public audiences and civil society should be aware that a more articulated understanding of accountability is required to assure an effective control of public service provision supported by AI.

REFERENCES

- Ammittzøll Flügge, A., Hildebrandt, T., and Møller, N.H. (2021). Street-level algorithms and AI in bureaucratic decision-making: A caseworker perspective. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1–23.
- Ávila, F., Hannah-Moffat, K., and Maurutto, P. (2020). The seductiveness of fairness: Is machine learning the answer? Algorithmic fairness in criminal justice systems. In M. Schuilenburg and R. Peeters (eds), *The Algorithmic Society: Technology, Power, and Knowledge* (pp. 87–103). Routledge.
- Bolt, D. (2017). *An Inspection of Entry Clearance Processing Operations in Croydon and Istanbul*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/631520/An-inspection-of-entry-clearance-processing-operations-in-Croydon-and-Istanbul1.pdf.
- Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework 1. *European Law Journal*, 13(4), 447–468.
- Bovens, M., and Zouridis, S. (2002). From street-level to system-level bureaucracies: How information and communication technology is transforming administrative discretion and constitutional control. *Public Administration Review*, 62(2), 174–184. <https://doi.org/10.1111/0033-3352.00168>.
- Bullock, J.B., Young, M.M., and Wang, Y.-F. (2020). Artificial intelligence, bureaucratic form, and discretion in public service. *Information Polity: The International Journal of Government and Democracy in the Information Age*, 1–16. <https://doi.org/10.3233/IP-200223>.
- Burrell, J. (2016). How the machine “thinks”: Understanding opacity in machine learning algorithms. *Big Data and Society*, 3(1), 1–12. <https://doi.org/10.1177/2053951715622512>.
- Busch, P.A., and Henriksen, H.Z. (2018). Digital discretion: A systematic literature review of ICT and street-level discretion. *Information Polity: The International Journal of Government and Democracy in the Information Age*, 23(1), 3–28. <https://doi.org/10.3233/IP-170050>.
- Busuioac, M. (2021). Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review*, 81(5), 825–836.
- Charles, V., Rana, N.P., and Carter, L. (2022). Artificial intelligence for data-driven decision-making and governance in public affairs. *Government Information Quarterly*, 39(4), 101742. <https://doi.org/10.1016/j.giq.2022.101742>.
- Criado, J.I., Valero, J., and Villodre, J. (2020). Algorithmic transparency and bureaucratic discretion: The case of SALER early warning system. *Information Polity*, 25(4), 449–470.
- de Bruijn, H., Warnier, M., and Janssen, M. (2021, December 30). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly*, 101666. <https://doi.org/https://doi.org/10.1016/j.giq.2021.101666>.
- Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM*, 59(2), 56–62.
- Dwivedi, Y.K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., and Eirug, A. (2019). Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 101994. <https://doi.org/10.1016/j.ijinfomgt.2019.08.002>.
- Favaretto, M., De Clercq, E., and Elger, B.S. (2019). Big data and discrimination: Perils, promises and solutions. A systematic review. *Journal of Big Data*, 6(1), 1–27.
- Fountain, J.E. (2004). *Building the Virtual State: Information Technology and Institutional Change*. Brookings Institution Press.

- Freeman, K. (2016). Algorithmic injustice: How the Wisconsin Supreme Court failed to protect due process rights in *State v. Loomis*. *North Carolina Journal of Law and Technology*, 18(5), 75–106.
- Giest, S., and Klievink, B. (2022). More than a digital system: How AI is changing the role of bureaucrats in different organizational contexts. *Public Management Review*, 1–20. <https://doi.org/10.1080/14719037.2022.2095001>.
- Grimmelikhuijsen, S. (2023). Explaining why the computer says no: Algorithmic transparency affects the perceived trustworthiness of automated decision-making. *Public Administration Review*, 83(2), 241–262, <https://doi.org/10.1111/puar.13483>.
- Gualdi, F., and Cordella, A. (2023). Policymaking in time of Covid-19: How the rise of techno-institutional inertia impacts the design and delivery of ICT-mediated policies. *56th Hawaii International Conference on System Sciences Maui*, Hawaii, United States.
- Harvard Law Review (2017). *State v. Loomis*: Wisconsin Supreme Court requires warning before use of algorithmic risk assessments in sentencing. *Harvard Law Review, Criminal Law*, 130(5), 1530–1537.
- House of Commons (2019). *Visa Processing Algorithms*. Retrieved January 20, 2021 from <https://www.theyworkforyou.com/debates/?id=2019-06-19a.316.0>.
- Kroll, J.A., Huey, J., Barocas, S., Felten, E.W., Reidenberg, J.R., Robinson, D.G., and Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, 165, 633.
- Lanzara, G.F. (2009). Building digital institutions: ICT and the rise of assemblages in government. In F. Contini and G.F. Lanzara (eds), *ICT and Innovation in the Public Sector: European Studies in the Making of eGovernment* (pp. 9–48). Palgrave Macmillan.
- Lebovitz, S., Lifshitz-Assaf, H., and Levina, N. (2022). To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization Science*, 33(1), 126–148, <https://doi.org/10.1287/orsc.2021.1549>.
- Leonardi, P.M., and Treem, J.W. (2020). Behavioral visibility: A new paradigm for organization studies in the age of digitization, digitalization, and datafication. *Organization Studies*, 41(12), 1601–1625.
- Liu, H.-W., Lin, C.-F., and Chen, Y.-J. (2019). Beyond *State v. Loomis*: Artificial intelligence, government algorithmization and accountability. *International Journal of Law and Information Technology*, 27(2), 122–141. <https://doi.org/10.1093/ijlit/ez001>.
- Lorenz, L., Meijer, A., and Schuppan, T. (2021). The algorithm as a new ideal type for government organizations: Predictive policing in Berlin as an empirical case. *Information Polity*, 26(1), 71–86.
- March, J.G., and Olsen, J.P. (2010). *Rediscovering Institutions*. Simon & Schuster.
- Martin, K. (2019). Ethical implications and accountability of algorithms. *Journal of Business Ethics*, 160(4), 835–850. <https://doi.org/10.1007/s10551-018-3921-3>.
- McDonald, H. (2019, October 29). AI system for granting UK visas is biased, rights groups claim. *The Guardian*. <https://www.theguardian.com/uk-news/2019/oct/29/ai-system-for-granting-uk-visas-is-biased-rights-groups-claim#:~:text=Immigrant%20rights%20campaigners%20have%20begun,UK%20visa%20applications%20actually%20works.andtext=Cori%20Cridder%2C%20a%20director%20at,defence%20of%20the%20AI%20system>.
- Medaglia, R., Gil-Garcia, J.R., and Pardo, T.A. (2021). Artificial intelligence in government: Taking stock and moving forward. *Social Science Computer Review*, 08944393211034087.
- Meijer, A., and Grimmelikhuijsen, S. (2020). Responsible and accountable algorithmization: How to generate citizen trust in governmental usage of algorithms. In M. Schuilenburg and R. Peeters (eds), *The Algorithmic Society* (pp. 53–66). Routledge.
- Meijer, A., Lorenz, L., and Wessels, M. (2021). Algorithmization of bureaucratic organizations: Using a practice lens to study how context shapes predictive policing systems. *Public Administration Review*, 81(5), 837–846.
- Mintzberg, H. (1983). *Structure in Fives: Designing Effective Organizations*. Prentice-Hall.

- Misuraca, G., van Noordt, C., and Boukli, A. (2020). The use of AI in public services: Results from a preliminary mapping across the EU. *Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance*.
- Pencheva, I., Esteve, M., and Mikhaylov, S.J. (2020). Big Data and AI—A transformational shift for government: So, what next for research? *Public Policy and Administration*, 35(1), 24–44.
- Schwarz, G., Christensen, T., and Zhu, X. (2022). Bounded rationality, satisficing, artificial intelligence, and decision-making in public organizations: The contributions of Herbert Simon. *Public Administration Review*, 82(5), 902–904.
- Simon, H.A. (2013). *Administrative Behavior*. Simon & Schuster.
- State v. Loomis*. (2016). Wisconsin Supreme Court. Retrieved January 25, 2021 from <https://caselaw.findlaw.com/wi-supreme-court/1742124.html>.
- Stohl, C., Stohl, M., and Leonardi, P.M. (2016). Digital age managing opacity: Information visibility and the paradox of transparency in the digital age. *International Journal of Communication*, 10, 15.
- Strich, F., Mayer, A.-S., and Fiedler, M. (2021). What do I do in a world of artificial intelligence? Investigating the impact of substitutive decision-making AI systems on employees' professional role identity. *Journal of the Association for Information Systems*, 22(2), 9.
- Sun, T.Q., and Medaglia, R. (2019). Mapping the challenges of artificial intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly*, 36(2), 368–383. <https://doi.org/10.1016/j.giq.2018.09.008>.
- Threipland, C., and Rickett, O. (2020, March 17). *Price and Prejudice: Automated Decision-Making and the UK Government*. Retrieved December 28, 2020 from <https://www.thejusticegap.com/price-and-prejudice-automated-decision-making-and-the-uk-government/>.
- von Krogh, G. (2018). Artificial intelligence in organizations: New opportunities for phenomenon-based theorizing. *Academy of Management Discoveries*, 4(4), 404–409. <https://doi.org/10.5465/amd.2018.0084>.
- Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.
- Wirtz, B., Langer, P., and Fenner, C. (2021). Artificial intelligence in the public sector—A research agenda. *International Journal of Public Administration*, 44(13), 1103–1128.
- Wirtz, B., and Müller, W. (2019). An integrated artificial intelligence framework for public management. *Public Management Review*, 21(7), 1076–1100. <https://doi.org/10.1080/14719037.2018.1549268>.
- Wirtz, B., Weyerer, J., and Sturm, B. (2020, July 3). The dark sides of artificial intelligence: An integrated AI governance framework for public administration. *International Journal of Public Administration*, 43(9), 818–829. <https://doi.org/10.1080/01900692.2020.1749851>.
- Zhang, Z., Yoo, Y., Lytinen, K., and Lindberg, A. (2021). The unknowability of autonomous tools and the liminal experience of their use. *Information Systems Research*, 32(4), 1192–1213.

16. Contrasting human–AI workplace relationship configurations

Miriam Möllers, Benedikt Berger, and Stefan Klein

AI AGENTS AT THE WORKPLACE: AGENCY AND AUGMENTATION

Recent technological advancements in the field of artificial intelligence (AI) allow the building of information technology (IT) systems that can perceive and adapt to their environments, as well as interact with humans in a human-like manner. When talking of AI, we refer to researchers' and practitioners' efforts to create IT systems that have the "ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation" (Kaplan and Haenlein, 2019, p. 17). Hereafter referred to as AI agents, IT systems having such abilities are characterized by their increasingly agentic nature, adding a new perspective to the traditional notion of humans using IT systems as tools for goal attainment (Schuetz and Venkatesh, 2020). Agentic nature means that AI agents can "take on specific rights for task execution and responsibilities for preferred outcomes" (Baird and Maruping, 2021, p. 317). Depending on the task and purpose, these AI agents can take up varying levels of agency: from assistant-like agents with limited agency, which reflexively respond to defined stimuli, up to autonomous agents, which can prescribe and take action, and thus bear the full agency for a particular task (Baird and Maruping, 2021). With increasing levels of agency, AI agents can accept (more) rights and responsibilities for task execution, request help from human agents, and even delegate (aspects of) tasks to humans. The shift from the use of IT systems, to delegation to and from IT systems, comes with a shift in roles between humans and AI agents (Demetis and Lee, 2018), which allows the integration of AI agents into an increasing number of processes at the workplace.

Given these novel possibilities, more and more organizations are gradually implementing AI agents for an increasing number of tasks and purposes, not least to leverage the large-scale data processing capabilities for complex decision-making and problem-solving (Agrawal et al., 2019). As a result, employees are more frequently confronted with AI agents at work. This development has opened debates about AI agents partially or fully replacing humans (Raisch and Krakowski, 2021), altering the nature of work (Mirbabaie et al., 2021), and reinforcing expectations of dramatic shifts in the proportion of tasks executed by machines (World Economic Forum, 2020), incurring massive job losses (Acemoglu and Restrepo, 2019; Haenlein and Kaplan, 2019). However, researchers are still at the outset of understanding the implications of AI agents in the workplace. While the transformation of tasks

and jobs, including the potential replacement of human workers, remains a crucial issue, a growing stream of research has considered ways of conceptualizing humans working with AI agents. These considerations introduce a new layer to the automation debate, shifting the focus from the substitution to the complementation of human work. Central to this literature is the idea of augmentation, which points towards the potential of combining the respective strengths of both humans and AI agents (Davenport et al., 2020; Dellermann et al., 2019b; Lyytinen et al., 2021; Murray et al., 2021; Willcocks, 2020).

These developments have incurred a multi-disciplinary discourse surrounding novel conceptualizations of human–AI relationships. Some of these have gained increasing attention, such as delegation to AI agents (Baird and Maruping, 2021), AI-based decision support (Shrestha et al., 2019), human–AI teams (Seeber et al., 2020), human-in-the-loop (Grønsund and Aanestad, 2020), and algorithmic management (Wiener et al., 2021). The increasing interest has created a large but heterogeneous terminology with various, at times overlapping or synonymous, terms and concepts. Such ambiguities in the use of terms can hamper the development of a cumulative body of knowledge. Therefore, we seek to take up and differentiate the most prominent terms in the context of human–AI workplace relationships in this chapter. For this purpose, we proceed in two steps. First, we lay out the current understanding of the most prominent configurations of human–AI relationships. This serves the objective of establishing a common ground. Second, we introduce a framework that categorizes and delineates different human–AI workplace relationships by depicting the flow of agency that occurs between humans and AI agents. Before taking these two steps, we introduce some required conceptual foundations. We hope that this chapter helps to distinguish the diverse configurations of human–AI relationships and builds a common conceptual ground for future research.

CONCEPTUAL FOUNDATIONS

Agency and Delegation

An important concept to understand the relationship between humans and AI agents is agency, which Rose and Jones (2005, p. 28) define as “act[ing] in a way which produces outcomes.” Within the boundaries of this definition, the authors differentiate human and machine agency based on the agents’ properties, which define their scope of acting. According to Rose and Jones (2005), machine agency involves three roles: the role of tools acting under the control of humans; proxies acting on behalf of humans; and automata taking over a minor part of decision-making along with the power to act. As opposed to that, human agency differs in terms of five unique human characteristics: self-awareness, social awareness, interpretation, intentionality, and attribution. Nowadays, autonomously acting AI agents increasingly challenge the scope and dominance of human agency vis-à-vis machine agency. As such, the ability to accept rights and responsibilities for ambiguous tasks and outcomes under

uncertainty, as well as to act and decide autonomously, has long been attributed to human agency (Baird and Maruping, 2021). As AI agents act in an environment and receive feedback, they acquire new capabilities which help them to operate better in that environment (Lyytinen et al., 2021). The capacity of AI agents to learn, adapt, and identify the need to act without being prompted by users allows such systems to delegate and assume tasks with a higher degree of uncertainty in unstructured and dynamic situations (Baird and Maruping, 2021; Schuetz and Venkatesh, 2020).

In the process of delegation, the delegator transfers agency (that is, the rights and responsibilities for a given task or decision) to a proxy (that is, the agent). When humans delegate a task to an AI agent, they give up (parts of) their agency to an AI agent, and thereby reduce their overall control over the outcome (Demetis and Lee, 2018). This, in turn, leads to a dynamic interplay or entanglement of human and AI agencies. Neff and Nagy (2018) refer to this entanglement of human and AI agencies in human–AI relationships as symbiotic agency. This agency conceptualization entails two important aspects. First, it accounts for the complex dynamics in human–AI relationships, which no longer allow the attribution of particular outcomes to solely human action, but only to joint action with an AI agent. Second, the conceptualization draws attention towards the ways in which humans and AI agents influence and are influenced by each other. To understand the transfer of responsibilities between humans and AI agents as part of delegation, we need to distinguish between moral and causal responsibility. While the ability to take on moral responsibility remains a human attribute, AI agents can bear causal responsibility. This conceptualization accounts for the fact that AI agents can be directly involved in the chain of actions leading to an outcome without following fully predetermined rules by humans (Lüthi et al., 2023). This issue has spurred multiple debates among scholars and the wider public, which we do not focus on in this chapter.

Automation and Augmentation

Varying on a continuum between full automation and full manual (human) work, automation refers to the degree to which a previously human-performed task is replaced by an AI agent (Parasuraman et al., 2000). Accordingly, the automation continuum allows us to distinguish the diverse human–AI relationships discussed in the literature, with the shades in between both ends representing various forms of augmentation. Dating back to Engelbart's (1962) conceptual writing on augmenting human intellect, augmentation means increasing or extending human capabilities through technology, for example to solve complex problems more efficiently and effectively. In the context of human–AI augmentation, the underlying assumption is that humans and AI agents have complementary capabilities and strengths, which they can combine to augment one another (Davenport et al., 2020; Dellermann et al., 2019b; Lyytinen et al., 2021; Murray et al., 2021; Siemon et al., 2020). As an example, the strengths of AI agents include high computing power and data processing capabilities (Pavlou, 2018). Humans, on the other hand, are better at dealing with ambiguous and dynamic situations, and complement systems with their intuition,

commonsense judgment, and knowledge of norms and values (Agrawal et al., 2019; Akata et al., 2020; Jarrahi, 2018). In the context of this chapter, our understanding of augmentation encompasses the augmentation of human skills and capabilities, human decision-making, and tasks on the one hand, as well as the augmentation of the technology's (that is, AI agent's) performance and problem-solving capabilities on the other. A prominent and closely related term drawing on the augmentation and complementarity of humans and AI agents is hybrid intelligence (HI). HI is defined as the ability to collectively achieve superior outcomes to the ones humans and AI agents could achieve on their own, and to continuously improve—as a system and individually—by learning from each other (Dellermann et al., 2019a; Dellermann et al., 2019b).

Cooperation and Collaboration

Following the automation continuum, AI agents either take on a complementary role (that is, focusing on cooperation and collaboration for augmentation) or a substituting role (that is, focusing on full automation) in relation to human work. Whereas some researchers use human–AI collaboration as a synonym for augmentation (e.g., Raisch and Krakowski, 2021), Baer et al. (2022) offer a distinction according to which cooperation and collaboration are requirements to achieve augmentation. Cooperation refers to at least two parties agreeing on a predefined contribution and outcome, and entails the more specific concept of collaboration (Randrup et al., 2016). Collaboration constitutes a “joint effort towards a group goal” which can “occur in any domain where people seek to create value together” (Randrup et al., 2016, pp. 898–900). We understand collaboration as a reciprocal process because the interaction between the involved parties generates feedback and, subsequently, enables mutual learning. Overall, the existence of a common goal and strategy towards this goal differentiates collaboration from simple cooperation (Siemon et al., 2019; Siemon et al., 2020). Cooperative settings may involve both reciprocal and unidirectional processes that are characterized by an AI agent providing input to human work. Whenever humans guide processes and are supported or assisted by AI agents, we interpret this as cooperation. We employ this differentiation when distinguishing specific human–AI workplace relationships in the next section.

CONFIGURATIONS OF HUMAN–AI WORKPLACE RELATIONSHIPS

AI-based (Decision) Support

In AI-based (decision) support, humans use AI agents as tools for improving their own decision processes. Shrestha et al. (2019) refer to such configurations of human–AI decision-making as hybrid decision-making, and distinguish between two manifestations: AI-to-human and human-to-AI sequential decision-making. In the former

manifestation, the AI agent takes on a filter function, categorizing diverse options and/or providing a smaller, suitable set of alternatives from a larger pool of options. The latter describes instances of human decision-makers passing the choice of the best option on to an AI agent after preselecting a small set of alternatives (Shrestha et al., 2019). Examples of both manifestations can be found in hiring processes. AI agents supporting human resources (HR) decisions may categorize available options and suggest a narrower set of potential candidates as part of the screening process. Moreover, they can be deployed to scan and analyze patterns in the prerecruitment candidate data to make predictions about turnover and performance (Pessach et al., 2020). In this way, AI agents inform human decision-makers by providing them with fast, repeatable, data-driven insights (Agrawal et al., 2018; Shrestha et al., 2019).

Despite the employment of AI agents, AI-based decision support is still characterized by humans guiding the decision-making process. This is crucial, as many decisions still require human judgment and contextual knowledge (Agrawal et al., 2019, 2018). The AI agent incorporates a supportive role, as the human can use the AI agent in a tool-like manner on demand, which supports human decision-making mainly by solving an optimization problem. In essence, humans leverage AI agents to more (time-)efficiently generate insights that surpass their own capabilities, and to reduce some uncertainty in decision-making processes. Nevertheless, as humans guide the process, they can freely decide when and to what extent to integrate the AI agent in the decision-making process, and whether they act upon the advice provided. By receiving advice, feedback, and contradictions from an AI agent, humans can learn and adapt their decisions (Abdel-Karim et al., 2020). However, the data by which these systems are informed are usually not free of biases (Shrestha et al., 2019), which calls for humans to thoroughly evaluate such insights. As behavioral research shows, this evaluation is a highly demanding task. Even though humans still guide the decision-making process, predictions or recommendations by an AI agent likely influence their behavior and decision outcomes. This becomes alarming when human decision makers start to blindly follow AI output without reconciling it with the real-world context, potentially exacerbating biases (Krügel et al., 2022).

Besides human resource management, there are many promising application contexts. In the financial industry, for instance, AI agents assist financial experts' decision-making by providing new insights through financial and economic data mining and analysis (Mihov et al., 2022). Another example of AI-based decision support is the risk assessment deployed in the United States criminal justice system that estimates the likelihood of criminal defendants' recidivism, informing judges in parole or early-release decisions (Green and Chen, 2019). In sports analytics, potential players are verified using AI-based performance prediction; and in health monitoring, AI agents monitor patients' bodily functions to predict and detect risks of acute disorders (Shrestha et al., 2019).

AI-in-the-Loop

Moving from AI-based (decision) support toward higher degrees of automation, the configuration of AI-in-the-Loop (AIITL) captures AI agents that incorporate a control or supervisory function and intervene in a human-led process. Thereby, the AI agent assists in the loop of human conduct. We observe two different manifestations of this control function in AIITL: one where the AI agent monitors human activities and notifies the human about potential errors that humans can accept or ignore; and another where the AI agent steps in once human behavior deviates from predefined rules. We would characterize the AI agent in this setting as an assistant, that is part of the human-led process by default, but can be deactivated or ignored.

An example of the first AIITL manifestation represents Ubisoft's Commit Assistant AI tool (Kamen, 2018). Used in the context of video game development, the task of the AI agent is to catch bugs before they are committed into code, thereby saving time and reducing development costs. In this example, the human developers produce the code on their own, while the AI agent accompanies this process to detect potential flaws and to intervene in the code writing process by recommending corrections to the developers. For this purpose, the AI agent is trained on a vast database of bugged and corrected code from past game developments (Kamen, 2018). As opposed to decision support, such AI agents are not consulted on demand, but constantly monitor the developers' coding activities. Another example are AI-based driver monitoring systems that use sensors to detect driver state and behavior, and intervene when potential threats such as distraction or drowsiness are recognized (smart eye, 2022).

While the first manifestation is characterized by humans receiving recommendations from the AI agent, AI agents in the second AIITL manifestation settings may also override human actions. In such cases, the AI agent takes control over the task conduct if it detects human behavior deviating from predefined rules. This is possible only if the AI agent outperforms the human capabilities, which is why only a few examples of this manifestation exist so far. Among the most prominent examples of overriding interventions are lane and braking assistants, which actively interfere in human action when detecting a dangerous driving behavior.

Human–AI Collaboration and Teams

The configurations of human–AI collaboration and human–AI teams encompass a whole range of (potential) application contexts whose underlying idea refers to the inclusion of an AI agent as a coequal partner or member in groups and teams. Following Siemon et al.'s (2019) framework, groups represent a rather loose construct without clearly defined boundaries, and their members primarily pursue their individual goals. Teams, on the other hand, are defined as “groups within a specific setting, with defined rules and principles, that follow a common goal and work in a long-term relationship where specific collaboration principles develop in a stronger manner” (Siemon et al., 2019, p. 1843). Hence, teamwork represents a specific form

of collaboration. In teamwork, team members reciprocally combine complementary strengths and dynamically interact with each other, taking on different team roles, specific positions, and responsibilities (Belbin, 2010). Thus, human–AI teams denote a particular type of human–AI collaboration; or as Rix (2022) puts it, collaboration provides a basis for teamwork. For this reason, we refer to human–AI collaboration as an overarching configuration that encompasses human–AI teams, but also partnerships and group collaboration. What distinguishes AI agents in these configurations from AI-based tools supporting human team collaboration, as in AI-based decision support, is their immediate and constant involvement in major parts of a complex problem-solving process. Considering the goal of effective collaboration, AI agents should take on different roles that complement human collaborators and thereby allow both humans and AI agents to employ their strengths most effectively (Dellermann et al., 2019b; Siemon et al., 2020).

Seeber et al. (2020, p. 3) sketch a vision of what they call “AI teammates” as:

autonomous, pro-active, and sophisticated technology that draws inferences from information, derives new insights from information, learns from past experiences, finds and provides relevant information to test assumptions, helps evaluate the consequences of potential solutions, debates the validity of proposed positions offering evidence and arguments, proposes solutions and provides predictions to unstructured problems, plus participates in cognitive decision-making with human actors.

To reap the benefits of human–AI teams, it is of great importance that humans accept the AI agent as a teammate, and perceive it as a partner instead of an inferior tool (Debowski et al., 2021; Rix, 2022; Walliser et al., 2019; Zhang et al., 2020). As Zhang et al. (2020) point out, humans’ tendency to view AI agents as tools may limit the team performance in solving high-complexity and collaborative tasks. The authors find that humans expect AI agents to possess instrumental skills for fulfilling collaborative tasks, a shared understanding with human teammates, sophisticated communication capabilities for information exchange, and human-like performance. In sum, Rix (2022) identifies three main drivers for the human perception of AI agents as teammates: the creation of a team setting, the establishment of the team as a social entity, and collaborative behavior of the AI agent. Whereas humanness and relationship-oriented behavior of the AI agent create the social entity, the need for collaborative behavior means that AI agents should “exhibit proactive, iterative, responsive as well as competent behavior while providing explanations and allowing for verbal communication” (Rix, 2022, p. 404).

Whereas, thus far, AI teammates are a mainly theoretical concept, we can already observe instances of human–AI collaboration. An example constitutes the context of video game development. Seidel et al. (2020) describe the integration of an autonomous design tool into video game design to generate design ideas independent from human designers. In this type of human–AI collaboration, human designers set input parameters for the autonomous design tool to process, eventually generating a design outcome. Outperforming human designers in speed, scale, and scope, the introduction of an AI agent allows for new innovative design ideas in a fraction of the

time. In this example, the AI agent takes on the role of a creator conducting research and development (Siemon, 2022). The human designers, on the other hand, evaluate and, if necessary, amend the design outcomes, and manually create other areas of the game space or adjust the outcome according to their envisioned ideas (Seidel et al., 2020). To do so, they apply their design experience as well as creativity and commonsense knowledge of the world. Additionally, collaborative writing—that is, AI agents complementing human writing (Wiethof et al., 2021)—represents another scenario for human–AI collaboration. Such co-writing exploits the advantages of the memory space and high computation rate of AI agents, whereas human agents contribute their detailed knowledge as well as creativity and common sense (Wiethof et al., 2021).

Above all, the configuration of human–AI collaboration characterizes the sharing of agency between humans and AI agents. The complementarity of human and AI capabilities is central to this configuration, whose symbiosis ideally leads to outcomes that surpass the usual performance of both agents. This collaborative endeavor represents an instance of hybrid intelligence (Dellermann et al., 2019b), as both agents augment each other and interact reciprocally, thereby generating feedback from which each contributor can learn. In cases where humans and AI agents additionally work together in specific team roles toward a joint goal, we refer to this as human–AI partnerships or teams. Drawing on insights from the human–AI teams literature, we suggest that the configuration of a human–AI partnership consists of one human agent and one AI agent; whereas, in line with Rix (2022), we argue that a team would be established as soon as at least one additional human enters the partnership. Moreover, the endowment of AI agents with autonomous capabilities and agency determines whether they constitute a teammate in collaboration scenarios. In line with Rix (2022), we aim to highlight that human–AI teamwork scenarios deserve closer evaluation of the requirements that ensure the acceptance of an AI agent as teammate, the way in which AI agents can be implemented to generate symbiotic outcomes, and how these aspects influence the perceptions and acceptance of AI-based teammates.

Human-in-the-Loop

Mirroring the AIITL configuration, the human-in-the-loop (HITL) configuration intends to keep the human as a critical component inside an automated process to handle tasks of supervision, exception control, as well as optimization and maintenance (Rahwan, 2018). Thereby, humans take over a complementary function to improve the performance of an AI agent. The HITL approach has been mostly associated with interactive machine learning techniques (Calma et al., 2016; Rahwan, 2018). Rooted in reinforcement learning, interactive machine learning refers to “algorithms that can interact with [human or AI] agents and can optimize their learning behavior through these interactions” (Holzinger, 2016, p. 119). Representing a form of human–AI collaboration with the goal of controlling or improving the AI

agent, the human-in-the-loop configuration can also be understood as a link between collaboration and full automation.

Humans can take on diverse roles in HITL settings, as they not only develop AI agents, but also create interaction data sets, or label data or events unknown to the AI agents (Grønsund and Aanestad, 2020; Rahwan, 2018). When remaining in the loop of machine learning, humans maintain control not only in the data preprocessing and feature selection, but also during the learning phase by directly interacting with the AI agent (Wiethof and Bittner, 2021). In this human–AI relationship, the AI agent resembles an apprentice that learns from human input (Brynjolfsson and Mitchell, 2017), whereas the human agent acts as a teacher who trains the AI agent, thereby making it more effective (Dellermann et al., 2019b). This approach is especially valuable if data are scarce or pretrained models need to be adapted for specific domains or contexts (Dellermann et al., 2019b).

Moreover, HITL proves useful when human oversight in regulation and control of ethically relevant autonomous systems, such as autonomous weapons, is needed (Grønsund and Aanestad, 2020). Human involvement in critical tasks may be required to avoid risks if the reliability of AI agents cannot be guaranteed (Kamar, 2016; Peng et al., 2020). The latter is particularly hard to achieve for AI agents building on highly complex models. In this respect, Rahwan (2018) points out two major human roles in HITL approaches, namely identifying and correcting errors by an otherwise autonomous system, and representing an accountable entity in case the system errs. Similarly, Grønsund and Aanestad (2020) identify two central tasks of humans in HITL settings: auditing and altering. Whereas auditing refers to monitoring and evaluating the performance of the AI agent, altering encompasses the continuous improvement of the data input or the presentation mode. The authors recognize both roles as crucial parts of the HITL configuration and find that the tasks of auditing and altering are interdependent, thereby creating a feedback loop (Grønsund and Aanestad, 2020).

As an additional example, the limitations of algorithmic trading in the financial industry call for human intervention as part of AI-augmented investment systems. Specifically, the systems lack data input on extreme events such as financial market crashes, and a fundamental understanding of financial markets requires human insight. Human experts called superforecasters and other financial professionals excel at identifying such changes and extrapolating information in financial markets, which is why they can calibrate and integrate human insight into these systems (Mihov et al., 2022). The integration of a human into the design and learning process thus allows enhancing the performance of the AI agent.

Delegation to AI

Automation particularly involves full delegation. When humans delegate a task to an AI agent, they grant the AI agent the rights and resources for its accomplishment, and thereby free up resources for other work (Lyytinen et al., 2021). In this case, AI agents develop protocols and select actions based on predefined objectives without

human involvement (Murray et al., 2021). Instead of collaboratively contributing to a joint goal, each party can leverage the other to execute tasks on the other's behalf while focusing on their individual goals. We speak of full delegation to AI when humans fully rely on an AI agent to produce an outcome without the input on how to achieve this outcome. Shrestha et al. (2019) refer to “full human to AI delegation” when AI agents make decisions without the intervention of a human, similar to managers delegating decisions to human experts. In this human–AI workplace relationship, an AI agent substitutes the human decision-maker.

Smart services systems represent an instance of full delegation to AI agents. Based on sensor technology, interconnected networks, contextual computing, and wireless communication, smart service systems steer processes on behalf of their users. The systems' capabilities range from the monitoring of work, the provision of service updates, and the collection of customer requests, to the surveillance and incentivization of consumer behavior (Bruhn and Hadwich, 2022). Another example of full delegation are robo-advisors that private investors can mandate to improve the financial performance of their investment portfolios while saving time and mental effort (Rühr et al., 2019). In these and similar situations, human delegators transfer decision-making rights to an AI agent, which takes on the role of the proxy. Taking the example of robo-advisors, the AI agent decides independently when and which assets to buy or sell on behalf of the human. To sum up, delegation to AI captures any instances where humans delegate a task or decisions to an AI agent to substitute what has previously been conducted by humans.

Algorithmic Management

Humans can delegate not only operative tasks to AI agents, but also managerial tasks, such that the AI agent becomes a delegator itself. Known as algorithmic management, this configuration entails two facets: algorithmic matching, which provides the primary service of coordinating demand and supply; and algorithmic control, which focuses on monitoring and nudging workers' behavior to ensure that their behavior is in accordance with the organization's goals (Möhlmann et al., 2021). As such, algorithmic management agents take over tasks that usually lie within the responsibility of human managers (Jarrahi and Sutherland, 2019). Traditionally, a human manager exhibits control over a human worker, defining a portfolio of performance indicators and enacting them on the human worker. Algorithmic control systems, however, combine human agency in the control configuration with digital agency in the control enactment and delivery (Wiener et al., 2021).

Algorithmic management constitutes a two-step delegation, from human to AI agent and, subsequently, from AI agent to human. As a result, two different levels of human–AI relationships exist in this configuration (Tarafdar et al., 2022). The first level describes the delegation of management tasks from human managers to the AI agent as part of the design and implementation process. During this implementation process, self-learning AI agents receive the “responsibility for making and executing decisions affecting labor, thereby limiting human involvement and oversight of the

process” (Duggan et al., 2020, p. 119). After selecting and implementing relevant control metrics, humans involved in the process (for example, developers or platform providers) usually withdraw from it. The second level encompasses the algorithmic matching and control of the workers’ behavior as well as the interaction between the AI agent and human workers through a digital interface. By tracking the activities of workers, the AI agent generates further data input for subsequent computations, such as an Uber driver’s style of driving, which in turn influences future matching activities (Tarafdar et al., 2022).

Prominent contexts for the application of algorithmic management in companies are human resources and logistics (Rinta-Kahila et al., 2022; Wood, 2021). The textbook example for algorithmic management is the online labor platform Uber, which has also received a lot of research attention. As a global ride-sharing network, Uber connects individuals offering private drives with passengers through a mobile app. The underlying matching algorithm takes the role of the human dispatcher in a traditional taxi company, managing demand and supply (Cram and Wiener, 2020). To align the drivers’ activities with the organizational goals, this AI agent nudges human behavior and tracks customer feedback ratings, encouraging drivers to work longer hours and to avoid bad driving habits (Cram and Wiener, 2020; Kellogg et al., 2020; Möhlmann et al., 2021). The case of Uber illustrates potential risks of automating management tasks without human oversight and interference in cases where the AI agent acts inconsistently or based on myopic managerial rules (Kellogg et al., 2020; Marabelli et al., 2021). The AI agent captures and processes data points that represent only a fraction of the real world. Due to the missing human control body interpreting and reacting to human drivers’ behavior and sentiments, inefficiencies and disadvantages for the human drivers can occur.

FRAMEWORK OF HUMAN–AI WORKPLACE RELATIONSHIPS

Having delineated six different relationships between humans and AI agents at the workplace, we now seek to contrast these relationships in a common framework. Drawing on the previously discussed definition by Parasuraman et al. (2000), we attempt to sort the configurations of human–AI relationships along a horizontal axis depicting their overall degree of automation. In line with Grønsund and Aanestad (2020), we argue that the intermediate stages in between both ends correspond to different forms of augmentation. The configurations furthermore differ in how agency is enacted and transferred through delegation between humans and AI agents. Thus, we aim to illustrate the flow of agency between humans and AI agents when delegation between both parties takes place on the vertical axis of our framework.

The Automation Continuum

Depending on how the roles between humans and AI agents are distributed and how the parties contribute to the overall goal, augmentation can roughly take on three forms: human, machine, and hybrid augmentation. Human augmentation refers to humans being supported by AI agents providing predictions and data-driven insights for human-led decision-making or problem-solving tasks (Dellermann et al., 2019a). Given the secondary role of the AI agent in a primarily human-led process, we depict both configurations of AI-based (decision) support and AIITL as human augmentation. Machine augmentation means that humans provide input for training AI agents to solve problems that they cannot yet solve alone (Dellermann et al., 2019a). Here, we find the HITL configuration where the human agent serves as a teacher or supervisor to the AI agent. Hybrid augmentation combines the augmentation of humans and AI agents simultaneously (Dellermann et al., 2019a). The previously discussed idea of human–AI collaboration and teamwork are representative of hybrid augmentation, implying that human and AI agents augment each other by complementing each other's strengths and limitations. Full delegation to AI is an equivalent to full automation. Following these considerations, we position the configurations of human–AI relationships on the previously mentioned automation continuum with human, hybrid, and machine augmentation representing the stages in between human work and full automation.

Algorithmic management remains a very specific configuration owing to its multi-tier relationships between human managers, an AI agent, and human workers. Depending on the perspective and the detailed arrangements in this configuration, one may find instances of delegation to and by an AI agent. Due to its complexity and uniqueness, we consider algorithmic management as one distinct configuration for the sake of this chapter. Depicting this configuration as a blend of different actor constellations, we can interpret algorithmic management as a specific form of human augmentation for two reasons: by taking over the matching of supply and demand, the AI agent not only introduces an unprecedented efficiency to the allocation process that reduces the time in which human workers wait for an offer, but also supports them in their task conduct. Naturally, this limits human agency over the choice of jobs, whereas humans still execute the job they are delegated to.

The Flow of Agency

To conceptualize the flow of agency between humans and AI agents, we draw upon the previously introduced concepts of human, machine, and symbiotic agency. We recall that when humans partially or fully delegate a task or decision, they typically reduce their agency, and thereby abdicate some part of or even full control over the outcome (Demetis and Lee, 2018). Accordingly, when AI agents are mandated as proxies and act on behalf of humans, they receive agency within that field of action. The more agency they receive, the less control humans can exercise over the outcome. This means that when humans fully automate tasks or decisions,

they fully delegate and transfer all agency over the task conduct and outcome. The more dynamically that humans and AI agents interact and delegate to and from one another, the more entangled both human and AI agencies are, demonstrating variations of symbiotic agency (Neff and Nagy, 2018). This is typically the case for more cooperative and collaborative settings, in which the interactions among both agents shape each other's actions. Figure 16.1 depicts the proposed framework illustrating the configurations of human–AI relationships along the automation continuum, and a vertical axis indicating the “flow of agency” across humans and AI agents. The arrows between both actors attempt to symbolize how agency is transferred from one agent to the other.

For all configurations, we presuppose that humans develop and implement the AI agent, which can already be interpreted as a form of delegation. As part of this development and implementation process, humans equip AI agents with various endowments, computational rules, and training data. This equips AI agents with the necessary degree of agency to act within a predefined field of action. Depending on the task and purpose, AI agents can take up varying levels of agency. Baird and Maruping (2021) suggest four archetypes: reflexive, supervisory, anticipatory, and prescriptive AI agents. Reflexive agents have only limited agency as they react reflexively in direct response to defined stimuli, and base their outcomes on clearly defined models and parameters. Supervisory agents take over a control function as they evaluate deviations from the norm and seek to help return to the norm or to enhance the probability of progression toward a specified goal. Anticipatory agents tend to anticipate human needs or wants; whereas prescriptive agents autonomously take over decisions and substitute humans. Although humans usually take on the role of the delegator, the latter archetype includes AI agents that are not restricted to being proxies to humans but can become delegators themselves. Overall, human–AI interactions shape the interplay of and relation between human and machine agency in various contexts. Next, we would like to demonstrate what this interplay looks like in the diverse configurations of human–AI relationships, by drawing on the previous examples.

Delineating the Configurations in the Framework

Overall, we characterize AI-based (decision) support as a clearly defined delegator–proxy relationship, in which the human agent delegates some part of a task or decision to an AI agent. Usually, these AI agents act in response to a human endeavor that entails some evaluation of predefined alternatives or a preselection among a pool of options. The previous example of AI-based decision support in hiring processes illustrates how humans actively request AI-based insights by delegating the tasks of candidate evaluation or candidate choice among a pool of alternatives. Hence, machine agency is limited to that specific human request with a closely defined action set. As part of this setting, humans retain the greater part of agency over the main process and outcome, that is, the final candidate selection. Although humans may feel that they can freely decide about the extent to which they want to rely on

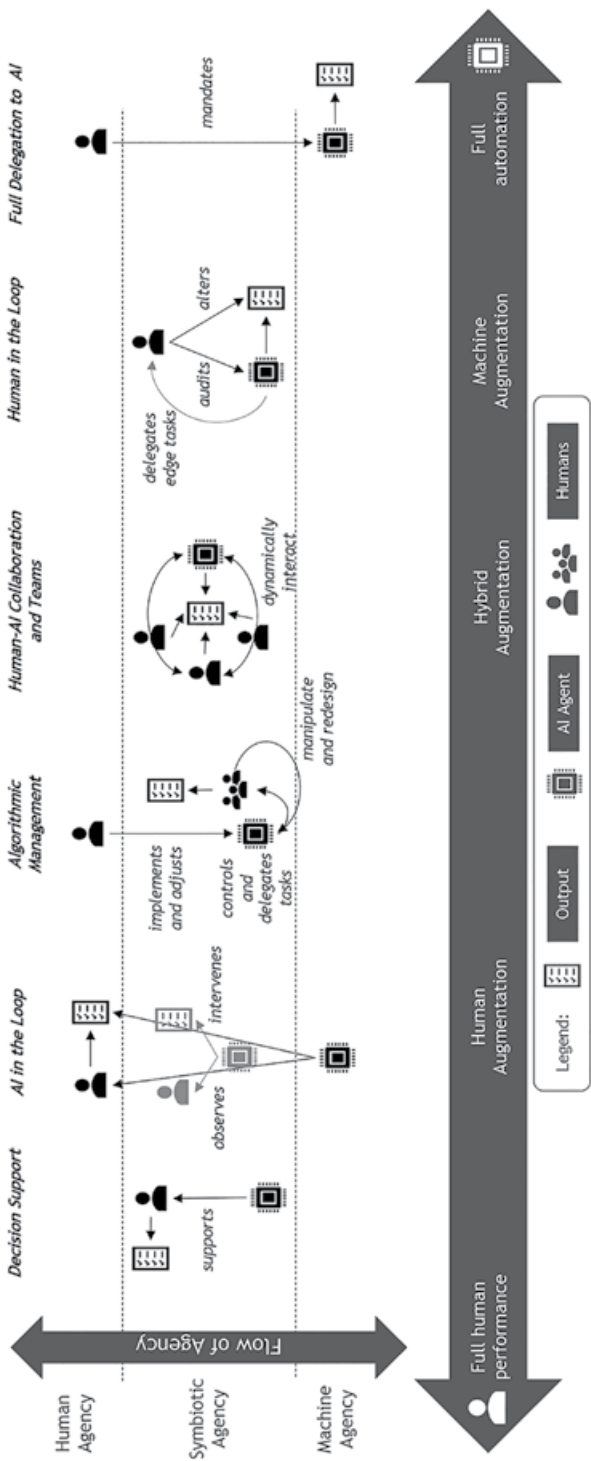


Figure 16.1 A framework of human–AI relationship configurations

AI-based insights, they are nevertheless consciously or unconsciously influenced by these insights. Whereas the field of action of machine agency is clearly defined, we may not be able to clearly decompose its impact on human-guided decisions and outcomes. Therefore, we characterize AI-based (decision) support as a form of symbiotic agency. While we can trace the flow of agency between humans and AI agents, we cannot attribute clearly defined parts of the outcome to either one.

To some extent, AI-based (decision) support blends into the AIITL configuration, which characterizes the control function of an AI agent in a human-guided process. We have identified two manifestations of AIITL, with the AI agent overseeing the task conduct and intervening: (1) by making suggestions that humans can either accept or ignore; or (2) by overriding human action once human behavior deviates from predefined rules. The example of Ubisoft's Commit Assistant tool ties up with AI-based (decision) support, as its primary aim is to anticipate bugs by alerting human programmers and/or recommending corrections. However, the tool's supervisory function and constant monitoring of the developers' activities distinguishes it from AI-based (decision) support. As humans can still decide whether to incorporate the proposed changes, they retain the largest part of agency over the outcome. Nevertheless, as machine agency interferes in human agency, they both intermingle to a form of symbiotic agency. The example of lane and brake assistants illustrates the second manifestation. By default, humans exhibit full agency over the driving process, whereas the AI agent only overrides human action when the vehicle leaves the lane or comes too close to other vehicles. We interpret this process of overriding as a form of structural delegation. Human and machine agencies operate in clearly defined fields of action where machine agency only jumps in once the AI agent detects a predefined situation that violates established boundaries.

In human–AI collaboration and teams, humans and AI agents interact dynamically and share the agency over the outcome. In this configuration, humans and AI agents may simultaneously adopt the role of delegator and proxy, and thereby transfer and take over agency for a specific outcome through dynamic interactions. As an example, we presented collaborative video game design. The AI agent and human designers collaborate by creating parts of the virtual gaming space and amending each other's contributions in numerous iterations. Whereas in the previous configurations the greater part of the outcome was attributable to human agency, the fields of action related to human and machine agencies intermingle and overlap to an even greater extent in human–AI collaboration. As a second example, the launch of ChatGPT has made the idea of collaborative writing accessible to everyone. Human writers can now rely on an AI agent to generate ideas and structure them, spurring critical discussions around the construct of authorship (Dwivedi et al., 2023).

In HITL configurations, humans abdicate even greater parts of their agency with respect to a task. With the aim of developing highly autonomous AI agents for complex and ambiguous problems, the HITL configuration represents an intermediate step on the way to full automation. The example of AI-augmented investment systems demonstrates the human role of oversight and supervision, incorporating contextual knowledge and unforeseen events in algorithmic trading. As part of this

role, humans delegate financial trading to AI agents and remain in the loop of the AI agent in one of two ways: either they retain some agency, constantly oversee the AI agent's output, and interfere whenever they see the need to; or they delegate the full agency over the task conduct to AI agents, which then operate autonomously but request human evaluation (that is, delegate the task conduct back to humans) once new and undefined situations occur (Baird and Maruping, 2021). Overall, we interpret this configuration as a form of symbiotic agency. Although the greatest part of the outcome may be attributable to machine agency, the interactions between humans and AI agents lead to some entanglement of both agencies.

We have introduced algorithmic management as a specific form of two-step delegation involving two different kinds of human agents: human managers delegating management tasks to AI agents by defining the action sets of the AI agents in the development process, and AI agents delegating jobs to human workers. Hence, managers transfer agency over the allocation of tasks to an AI agent, which delegates tasks to human workers, that is, the proxies of the AI agent. In addition to this delegation process, the AI agent not only constantly collects data on human workers' behavior, but also incentivizes efficient conduct through digital nudges. This, in turn, limits the workers' autonomy and values (Meijerink and Bondarouk, 2021); or, in other words, their agency. Still, as workers create an understanding of the mechanisms of the AI agent through sensemaking strategies, they start to circumvent the AI agent for their own good by adapting their behavior and changing the AI agent's input. The relationship between platform workers and AI agents shows how AI agents and worker behavior shape each other (Jarrahi and Sutherland, 2019; Meijerink and Bondarouk, 2021). This may affect decisions on the managerial level, as human managers and developers may change the rules and resources embedded in the AI agent as a response (Meijerink and Bondarouk, 2021). Following Jarrahi and Sutherland (2019), we interpret the AI agent as a mediator between human managers and workers, as it serves as a transmitter of agency. Since the boundaries between responsibilities of human workers and AI agents are not fixed, they are constantly negotiated and enacted (Jarrahi et al., 2021), resulting in a form of symbiotic agency.

Delegation to AI encompasses a great variety of automated tasks in numerous domains. Robo-advisors represent just one out of many examples. As previously explained, humans abdicate their full agency over a specific task by fully delegating it to AI agents, which execute it on behalf of the humans. In the case of robo-advisors, the AI agent decides independently which assets to buy or sell on behalf of the human investor, and when. In this configuration, humans can actively define the AI agent's field of action, which remains free of human interference. Hence, we can clearly delimit human and machine agencies in delegation to AI. Table 16.1 aims to provide an illustration of how the configurations may look in practical settings.

Table 16.1 *Exemplary illustrations of human–AI relationship configurations*

Configuration	Example	Illustration
AI-based (decision) support	HR predictive analysis	HR professionals may leverage AI agents to make predictions about turnover or performance in order to evaluate whether a candidate matches the job profile. In this way, the AI agent informs the HR professionals with further data-driven insights which they interpret to come to a decision (Pessach et al., 2020).
AI-in-the-loop (AIITL)	Monitoring: Ubisoft Commit Assistant AI tool	Human developers deploy Ubisoft's Commit Assistant AI tool to constantly monitor their coding activities with the goal of identifying bugs before they are committed into code. Once the AI agent detects potential flaws, it intervenes in the code writing process by recommending corrections. The developers can decide to what extent they follow the recommendations (Kamen, 2018).
	Monitoring and structural delegation: Lane and braking assistants	Modern cars are equipped with lane and braking assistants that constantly monitor the human driving process by evaluating sensor data to detect deviations from the norm (e.g., the vehicle moves out of its lane or is too close to standing cars). Depending on its programming, it may notify the human driver (monitoring) or take over full control of the vehicle (structural delegation) by steering it back into the lane or braking to avoid a potential crash.
	Driver monitoring systems	Human game designers may leverage an autonomous design tool to generate novel design ideas. Based on input parameters set by the designers, the autonomous design tool generates parts of the game world. Human designers evaluate and, if necessary, amend the design outcomes and create further areas of the game space manually. This may result in a reciprocal process where human designers and the autonomous design tool generate new parts of the game design, building on each other's work (Seidel et al., 2020).
Human–AI collaboration and teams	Video game development	AI-augmented investment systems leverage automated trading agents that are enhanced by unique human insights. As these systems lack data input on extreme events such as financial market crashes or unexpected market entries, as well as fundamental expert knowledge, superforecasters and other financial professionals calibrate and integrate additional insights into the systems. The goal is to enhance the performance of the AI agent by integrating a human into the design and learning process (Mihov et al., 2022).
Human-in-the-loop (HITL)	AI augmented investment systems	Individuals may mandate robo-advisers to manage their investment portfolio. Based on identified risk profiles, these AI agents assign portfolio weights to the preselected products. The AI agent constantly monitors the underlying risk of the portfolio and autonomously triggers trades to realign the desired and actual portfolio risk (Rühr et al., 2019).
Full delegation to and from AI	Robo-advisors	Online labor platforms such as Uber leverage AI agents to automate the matching of drivers and passengers. To do so, platform managers and/or developers equip the AI agent with decision parameters and let it operate autonomously. During job execution, the AI agent collects further data on the drivers' behavior and sends out nudges to influence the drivers' conduct. As drivers are mainly exposed to automated decisions, they may adjust their behavior in order to circumvent undesired automated allocation decisions. This may affect decisions on the managerial level, as human managers and developers may change the rules and resources embedded in the AI agent as a response (e.g., Meijerink and Bondarouk, 2021; Tarafdar et al., 2022).
Algorithmic management	Labor platforms, e.g., Uber	

CONCLUSION

The framework presented in this chapter is an attempt to sort some of the emerging configurations of human–AI relationships related to the automation and augmentation of human work. Since the increasing research interest in this field has brought up various terms and constructs, this chapter seeks to establish a common understanding and integrate the most apparent human–AI configurations in a framework. By doing so, we show that automation and augmentation are not monolithic constructs, but entail profoundly different degrees of human–AI interaction. Through these interactions, humans and AI agents delegate and, thereby, abdicate or receive the agency for a task or decision. This often results in an entanglement of human and machine agencies, or symbiotic agency, demonstrating the mutual influence of both agents through these interactions.

Finally, we would like to reflect on our chapter by raising three concluding points. First, the configurations are prototypical and reflect analytical distinctions. Hence, there may be examples that combine elements of several configurations or that are positioned in between the configurations. For instance, we briefly explained certain overlapping elements between AI-based (decision) support and AIITL by the example of Ubisoft's Commit Assistant. Second, we would like to pick up Raisch and Krakowski's (2021) argument of the reciprocal tension between augmentation and automation across time and space. This implies that the proposed human–AI configurations may change over time, which can be demonstrated by various instances of HITL. Some HITL configurations occur as part of development processes with the goal to fully automate tasks. Moreover, whereas managers used to interact with human employees, increasing automation involves increasing interaction of managers with AI agents, thereby leading to augmentation on the managerial level. We argue that a similar tension may exist across the different configurations of human–AI relationships. Third, we propose that the configurations highly depend on and reflect the perspective and appraisal of a particular system by its users. Those who believe in the ability of an AI agent to learn independently, to accumulate knowledge, and to act as a cooperative partner may enter a different configuration with one and the same AI agent than someone who evaluates the potential of the AI agent more critically and believes its abilities to be more limited.

REFERENCES

- Abdel-Karim, B.M., Pfeuffer, N., Rohde, G., and Hinz, O. (2020). How and what can humans learn from being in the loop? *KI—Künstliche Intelligenz*, 34(2), 199–207. <https://doi.org/10.1007/s13218-020-00638-x>.
- Acemoglu, D., and Restrepo, P. (2019). Artificial intelligence, automation, and work. In A. Agrawal, J. Gans, and A. Goldfarb (eds), *Chicago Scholarship Online. The Economics of Artificial Intelligence: An Agenda* (pp. 197–236). University of Chicago Press.
- Agrawal, A., Gans, J., and Goldfarb, A. (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard Business Review.

- Agrawal, A., Gans, J., and Goldfarb, A. (2019). Exploring the impact of artificial intelligence: prediction versus judgment. *Information Economics and Policy*, 47, 1–6. <https://doi.org/10.1016/j.infoecopol.2019.05.001>
- Akata, Z., Balliet, D., Rijke, M. de, Dignum, F., Dignum, V., Eiben, G., Fokkens, A., et al. (2020). A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8), 18–28. <https://doi.org/10.1109/MC.2020.2996587>.
- Baer, I., Waardenburg, L., and Huysman, M. (2022). What are we augmenting? A multidisciplinary analysis of AI-based augmentation for the future of work. In *ICIS 2022 Proceedings* (pp. 1–17).
- Baird, A., and Maruping, L.M. (2021). The next generation of research on IS use: a theoretical framework of delegation to and from agentic IS artifacts. *MIS Quarterly*, 45(1), 315–341. <https://doi.org/10.25300/MISQ/2021/15882>.
- Belbin, R.M. (2010). *Team Roles at Work*. Butterworth-Heinemann.
- Bruhn, M., and Hadwich, K. (eds). (2022). *Smart Services*. Springer Fachmedien Wiesbaden. <https://doi.org/10.1007/978-3-658-37384-9>.
- Brynjolfsson, E., and Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science*, 358(6370), 1530–1534. <https://doi.org/10.1126/science.aap8062>.
- Calma, A., Leimeister, J.M., Lukowicz, P., Oeste-Reiß, S., Reitmaier, T., Schmidt, A., Sick, B., Stumme, G., and Zweig, K.A. (2016). From active learning to dedicated collaborative interactive learning. In *ARCS 2016: 29th International Conference on Architecture of Computing Systems*, Nuremberg, Germany.
- Cram, W.A., and Wiener, M. (2020). *Technology-Mediated Control: Case Examples and Research Directions for the Future of Organizational Control*. Communications of the Association for Information Systems.
- Davenport, T., Guha, A., Grewal, D., and Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science*, 48(1), 24–42. <https://doi.org/10.1007/s11747-019-00696-0>.
- Debowski, N., Siemon, D., and Bittner, E. (2021). Problem areas in creativity workshops and resulting design principles for a virtual collaborator. In *Twenty-fifth Pacific Asia Conference on Information Systems*, Dubai, UAE.
- Dellermann, D., Calma, A., Lipusch, N., Weber, T., Weigel, S., and Ebel, P. (2019a). The future of human–AI collaboration: a taxonomy of design knowledge for hybrid intelligence systems. In *Hawaii International Conference on System Sciences (HICSS)*, Hawaii, USA.
- Dellermann, D., Ebel, P., Söllner, M., and Leimeister, J.M. (2019b). Hybrid intelligence. *Business and Information Systems Engineering*, 61(5), 637–643. <https://doi.org/10.1007/s12599-019-00595-2>.
- Demetis, D., and Lee, A.S. (2018). When humans using the IT artifact becomes IT using the human artifact. *Journal of the Association for Information Systems*, 929–952. <https://doi.org/10.17705/1jais.00514>.
- Duggan, J., Sherman, U., Carbery, R., and McDonnell, A. (2020). Algorithmic management and app-work in the gig economy: a research agenda for employment relations and HRM. *Human Resource Management Journal*, 30(1), 114–132. <https://doi.org/10.1111/1748-8583.12258>.
- Dwivedi, Y.K., Kshetri, N., Hughes, L., Slade, E.L., Jeyaraj, A., Kar, A.K., Baabdullah, A.M., et al. (2023). “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, 102642. <https://doi.org/10.1016/j.ijinfomgt.2023.102642>.
- Engelbart, D.C. (1962). *Augmenting Human Intellect: A Conceptual Framework*. Menlo Park, CA: Stanford Research Institute.

- Green, B., and Chen, Y. (2019). Disparate interactions: an algorithm-in-the-loop analysis of fairness in risk assessments. In *Fat* '19: Conference on Fairness, Accountability, Transparency*, Atlanta, GA.
- Grønsund, T., and Aanestad, M. (2020). Augmenting the algorithm: emerging human-in-the-loop work configurations. *Journal of Strategic Information Systems*, 29(2), 1–16. <https://doi.org/10.1016/j.jsis.2020.101614>.
- Haenlein, M., and Kaplan, A. (2019). A brief history of artificial intelligence: on the past, present, and future of artificial intelligence. *California Management Review*, 61(4), 5–14. <https://doi.org/10.1177/0008125619864925>.
- Holzinger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, 3(2), 119–131. <https://doi.org/10.1007/s40708-016-0042-6>.
- Jarrahi, M.H. (2018). Artificial intelligence and the future of work: human–AI symbiosis in organizational decision making. *Business Horizons*, 61(4), 577–586. <https://doi.org/10.1016/j.bushor.2018.03.007>.
- Jarrahi, M.H., Newlands, G., Lee, M.K., Wolf, C.T., Kinder, E., and Sutherland, W. (2021). Algorithmic management in a work context. *Big Data and Society*, 8(2), 20539517211020332. <https://doi.org/10.1177/20539517211020332>.
- Jarrahi, M.H., and Sutherland, W. (2019). Algorithmic management and algorithmic competencies: understanding and appropriating algorithms in gig work. In *iConference*, Washington, DC.
- Kamar, E. (2016). Directions in hybrid intelligence: complementing AI systems with human intelligence. In S. Kambhampati (ed.), *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)* (pp. 4070–4073). AAAI Press.
- Kamen, M. (2018). Ubisoft is using AI to catch bugs in games before devs make them. *Wired*. <https://www.wired.co.uk/article/ubisoft-commit-assist-ai>.
- Kaplan, A., and Haenlein, M. (2019). Siri, Siri, in my hand: who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, 62(1), 15–25. <https://doi.org/10.1016/j.bushor.2018.08.004>.
- Kellogg, K.C., Valentine, M.A., and Christin, A. (2020). Algorithms at work: the new contested terrain of control. *Academy of Management Annals*, 14(1), 366–410. <https://doi.org/10.5465/annals.2018.0174>.
- Krügel, S., Ostermaier, A., and Uhl, M. (2022). Zombies in the loop? Humans trust untrustworthy AI-advisors for ethical decisions. *Philosophy and Technology*, 35(1). <https://doi.org/10.1007/s13347-022-00511-9>.
- Lüthi, N., Matt, C., Myrach, T., and Junglas, I. (2023). Augmented intelligence, augmented responsibility? *Business and Information Systems Engineering*. <https://doi.org/10.1007/s12599-023-00789-9>.
- Lyytinen, K., Nickerson, J.V., and King, J.L. (2021). Metahuman systems = humans + machines that learn. *Journal of Information Technology*, 36(4), 427–445. <https://doi.org/10.1177/0268396220915917>.
- Marabelli, M., Newell, S., and Handunge, V. (2021). The lifecycle of algorithmic decision-making systems: organizational choices and ethical challenges. *Journal of Strategic Information Systems*, 30(3), 101683. <https://doi.org/10.1016/j.jsis.2021.101683>.
- Meijerink, J., and Bondarouk, T. (2021). The duality of algorithmic management: toward a research agenda on HRM algorithms, autonomy and value creation. *Human Resource Management Review*, 100876. <https://doi.org/10.1016/j.hrmr.2021.100876>.
- Mihov, A.-H., Firoozye, N., and Treleaven, P. (2022). Towards augmented financial intelligence. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4148057>.
- Mirbabaie, M., Brünker, F., Möllmann Frick, N.R.J., and Stieglitz, S. (2021). The rise of artificial intelligence—understanding the AI identity threat at the workplace. *Electronic Markets*. <https://doi.org/10.1007/s12525-021-00496-x>.

- Möhlmann, M., Zalmanson, L., Henfridsson, O., and Gregory, R.W. (2021). Algorithmic management of work on online labor platforms: when matching meets control. *MIS Quarterly*, 45(4), 1999–2022. <https://doi.org/10.25300/MISQ/2021/15333>.
- Murray, A., Rhymer, J., and Sirmon, D.G. (2021). Humans and technology: forms of conjoined agency in organizations. *Academy of Management Review*, 46(3), 552–571. <https://doi.org/10.5465/amr.2019.0186>.
- Neff, G., and Nagy, P. (2018). Agency in the digital age: using symbiotic agency to explain human–technology interaction. In Z. Papacharissi (ed.), *A Networked Self and Human Augmentics, Artificial Intelligence, Sentience* (pp. 97–107). Routledge. <https://doi.org/10.4324/9781315202082-8>.
- Parasuraman, R., Sheridan, T.B., and Wickens, C.D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics. Part A, Systems and Humans: A Publication of the IEEE Systems, Man, and Cybernetics Society*, 30(3), 286–297. <https://doi.org/10.1109/3468.844354>.
- Pavlou, P. (2018). Internet of Things—will humans be replaced or augmented? *AI and Augmented Intelligence*, 10(2), 42–47.
- Peng, R., Lou, Y., Kadoch, M., and Cheriet, M. (2020). A human-guided machine learning approach for 5G smart tourism IoT. *Electronics*, 9(6), 947. <https://doi.org/10.3390/electronics9060947>.
- Pessach, D., Singer, G., Avrahami, D., Chalutz Ben-Gal, H., Shmueli, E., and Ben-Gal, I. (2020). Employees recruitment: a prescriptive analytics approach via machine learning and mathematical programming. *Decision Support Systems*, 134, 113290. <https://doi.org/10.1016/j.dss.2020.113290>.
- Rahwan, I. (2018). Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology*, 20(1), 5–14. <https://doi.org/10.1007/s10676-017-9430-8>.
- Raisch, S., and Krakowski, S. (2021). Artificial intelligence and management: the automation–augmentation paradox. *Academy of Management Review*, 46(1), 192–210. <https://doi.org/10.5465/amr.2018.0072>.
- Randrup, N., Druckenmiller, D., and Briggs, R.O. (2016). Philosophy of collaboration. In *49th Hawaii International Conference on System Sciences*.
- Rinta-Kahila, T., Someh, I., Gillespie, N., Indulska, M., and Gregor, S. (2022). Algorithmic decision-making and system destructiveness: a case of automatic debt recovery. *European Journal of Information Systems*, 31(3), 313–338. <https://doi.org/10.1080/0960085X.2021.1960905>.
- Rix, J. (2022). From tools to teammates: conceptualizing humans’ perception of machines as teammates with a systematic literature review. In *Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS)*, Hawaii, USA.
- Rose, J., and Jones, M. (2005). The double dance of agency: a socio-theoretic account of how machines and humans interact. *Systems, Signs and Actions*, 1(1), 19–37.
- Rühr, A., Streich, D., Berger, B., and Hess, T. (2019). A classification of decision automation and delegation in digital investment management systems. In *Proceedings of the 52nd Hawaii International Conference on System Sciences* (pp. 1435–1444). <https://doi.org/59440>.
- Schuetz, S., and Venkatesh, V. (2020). The rise of human machines: how cognitive computing systems challenge assumptions of user–system interaction. *Journal of the Association for Information Systems*, 21(2), 460–482. <https://doi.org/10.17705/1jais.00608>.
- Seeber, I., Bittner, E., Briggs, R.O., Vreede, T. de, Vreede, G.-J. de, Elkins, A., Maier, R., et al. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information and Management*, 57(2). <https://doi.org/10.1016/j.im.2019.103174>.
- Seidel, S., Berente, N., Lindberg, A., Lyytinen, K., Martinez, B., and Nickerson, J.V. (2020). Artificial intelligence and video game creation: a framework for the new logic of autonomous design. *Journal of Digital Social Research*, 2(3), 126–157.

- Shrestha, Y.R., Ben-Menahem, S.M., and Krogh, G. von (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review*, 61(4), 66–83. <https://doi.org/10.1177/0008125619862257>.
- Siemon, D. (2022). Elaborating team roles for artificial intelligence-based teammates in human–AI collaboration. *Group Decision and Negotiation*. <https://doi.org/10.1007/s10726-022-09792-z>.
- Siemon, D., Becker, F., Eckardt, L., and Robra-Bissantz, S. (2019). One for all and all for one—towards a framework for collaboration support systems. *Education and Information Technologies*, 24(2), 1837–1861. <https://doi.org/10.1007/s10639-017-9651-9>.
- Siemon, D., Robra-Bissantz, S., and Li, R. (2020). Towards a model of team roles in human–machine collaboration. In *Forty-First International Conference on Information Systems*, India.
- smart eye (2022). *Driver Monitoring System: Intelligent Safety Features Detecting Driver State and Behavior*. <https://smarteye.se/solutions/automotive/driver-monitoring-system/>.
- Tarafdar, M., Page, X., and Marabelli, M. (2022). Algorithms as co-workers: human algorithm role interactions in algorithmic work. *Information Systems Journal*, Article isj.12389. <https://doi.org/10.1111/isj.12389>.
- Walliser, J.C., Visser, E.J. de, Wiese, E., and Shaw, T.H. (2019). Team structure and team building improve human–machine teaming with autonomous agents. *Journal of Cognitive Engineering and Decision Making*, 13(4), 258–278. <https://doi.org/10.1177/1555343419867563>.
- Wiener, M., Cram, W., and Benlian, A. (2021). Algorithmic control and gig workers: a legitimacy perspective of Uber drivers. *European Journal of Information Systems*, 1–23. <https://doi.org/10.1080/0960085X.2021.1977729>.
- Wiethof, C., and Bittner, E. (2021). Hybrid intelligence—combining the human in the loop with the computer in the loop: a systematic literature review. In *Forty-Second International Conference on Information Systems*, Austin, TX.
- Wiethof, C., Tavanapour, N., and Bittner, E. (2021). Implementing an intelligent collaborative agent as teammate in collaborative writing: toward a synergy of humans and AI. In *Proceedings of the 54th Hawaii International Conference on System Sciences* (pp. 400–409).
- Willcocks, L. (2020). Robo-Apocalypse cancelled? Reframing the automation and future of work debate. *Journal of Information Technology*, 35(4), 286–302. <https://doi.org/10.1177/0268396220925830>.
- Wood, A.J. (2021). *Algorithmic Management Consequences for Work Organisation and Working Conditions*. JRC Working Papers Series on Labour, Education and Technology.
- World Economic Forum (2020). *How Countries are Performing on the Road to Recovery*. The Global Competitiveness Report, World Economic Forum. Geneva.
- Zhang, R., McNeese, N.J., Freeman, G., and Musick, G. (2020). “An ideal human”: expectations of AI teammates in human–AI teaming. In *Proceedings of the ACM on Human Computer Interaction*, 4, 246. <https://doi.org/10.1145/3432945>.

PART III

IMPLICATIONS OF DECISIONS MADE WITH AI

17. Who am I in the age of AI? Exploring dimensions that shape occupational identity in the context of AI for decision-making

Anne-Sophie Mayer and Franz Strich

INTRODUCTION

Organizations increasingly introduce artificial intelligence (AI) systems to augment or even automate human decision-making (Benbya et al., 2021; Faraj et al., 2018). Prominent examples include AI systems used for hiring decisions (van den Broek et al., 2021), credit scoring (Mayer et al., 2020), and tumor detection and segmentation (Lebovitz et al., 2022). Whereas AI systems for decision-making promise various potentials for organizations, such as enhanced accuracy, objectivity, and efficiency in decisions (Faraj et al., 2018; van Krogh, 2018), these technologies come with substantial changes to occupations whose core activity centers around decision-making practices. AI systems fundamentally transform employees' work processes, tasks, and responsibilities (Benbya et al., 2021; Faraj et al., 2018; Giermindl et al., 2022), which are key components of how professionals perceive themselves (Chreim et al., 2007; Pratt et al., 2006). Changes to employees' core practices may therefore affect their occupational identity (Strich et al., 2021; Vaast and Pinsonneault, 2021).

Occupational identity reflects employees' answers to the questions: "Who am I (as a member of a specific profession)?" and "What do I do?" (Chreim et al., 2007; Nelson and Irwin, 2014; Pratt et al., 2006; Reay et al., 2017). These answers are subject to change and can be triggered by various internal and external factors such as interprofessional competition or new regulations (Craig et al., 2019; Petriglieri, 2011; Reay et al., 2017). A growing stream of literature points at the transformative impact of new technologies on employees' occupational identity (e.g., Nach, 2015; Nelson and Irwin, 2014; Stein et al., 2013; Vaast and Pinsonneault, 2021). Some studies emphasize that new technologies may threaten employees' occupational identity, resulting in resistant employee behavior toward the technology (e.g., Craig et al., 2019; Kim and Kankanhalli, 2009; Lapointe and Rivard, 2005). Other studies illustrate a more positive impact. For instance, Nelson and Irwin (2014) showed how the emergence of the internet first threatened librarians' identity, but then librarians used the new technology to further develop their roles, in that they redefined themselves from masters of search through to masters of interpretation and connectors of people and information. Similarly, Vaast and Pinsonneault (2021) emphasized

that digital technologies both enable and threaten occupational identity by creating persistence–obsolescence as well as similarity–distinctiveness tensions.

However, in the context of AI, we know little about how AI systems affect employees' occupational identity. A few studies have begun exploring this question, indicating divergent effects. For instance, Christin's (2020) work explored how algorithms change journalists' work and occupational identity. Her findings indicate that journalists, depending on their cultural and historic background, perceive different changes in their occupational identity through the use of algorithms. A study that particularly focused on the decision-making context is the work of Strich et al. (2021), who examined how an AI system used for loan decisions in the banking industry affected loan consultants' occupational identity. They found that whereas the introduction of the AI system threatened the occupational identity of high-skilled consultants, it empowered less-skilled consultants. Despite these insights, we lack a holistic understanding of dimensions that determine how AI systems affect occupational identity. This perspective is important, because AI systems are used for different purposes in decision making (for example, automation versus augmentation; Raisch and Krakowski, 2021) and target different professional groups (for example, low-skilled versus high-skilled workers; Susskind and Susskind, 2015). To better understand the underlying rationale, our chapter sets out to explore dimensions that deepen our understanding of how AI systems affect employees' occupational identity. By making use of qualitative insights from three case companies that introduced AI systems for decision-making in the areas of loan consulting, customer service, and executive development, we aim to provide theoretical and practical implications to better navigate the consequences of AI systems for decision-making practices in the workplace.

CONCEPTUAL BACKGROUND

AI for Decision-Making in the Workplace

The introduction of AI systems in organizations promises to improve organizations' and employees' practices to make decisions and generate knowledge (Benbya et al., 2021; Faraj et al., 2018). Contemporary AI systems are characterized by a high degree of autonomy, learning capacity, and opacity; properties that make AI systems distinct from previous information system (IS) technologies (Baird and Maruping, 2021; Berente et al., 2021). First, AI systems can autonomously perform holistic work and decision-making practices with increasingly complex functions, which by far exceed the narrowly defined and less sophisticated tasks executed by previous technologies (Bailey et al., 2019; Faraj et al., 2018; Leicht-Deobald et al., 2019; Raisch and Krakowski, 2021; von Krogh, 2018). Resulting from their ability to mimic humans' retrieval of tacit knowledge (Faraj et al., 2018), AI technologies can substitute isolated knowledge work tasks, but also knowledge workers' core work practices or even their entire job (Strich et al., 2021; Susskind and Susskind, 2015).

Consequently, the introduction of AI systems challenges employees' work and their application of knowledge, skills, and expertise.

Second, AI systems aim to generate predictive models by inferring patterns from large amounts of data (van den Broek et al., 2021). Although AI systems rely on human-controlled training data for their initial instructions, they can autonomously refine or even evolve their logic and patterns, connections, and weighting as they learn from additional data points (Burrell, 2016). Consequently, employees must deal with decisions that are often not predictable, and thus difficult to understand or justify, resulting in new challenges for their work and occupation (Dourish, 2016; Rai et al., 2019).

Third, AI systems are trained on very complex and sometimes even unstructured datasets, resulting in opaque decisions for users (Burrell, 2016; Hafermalz and Huysman, 2021). Moreover, since many of these systems self-adapt, even developers struggle to explain how and why a certain decision evolved (Burrell, 2016; Hafermalz and Huysman, 2021). The resulting opacity of AI systems therefore evokes new challenges for employees who have to collaborate with the AI system and deal with AI-derived decisions.

Occupational Identity

Occupational identities reflect how professionals perceive themselves in the workplace (Chreim et al., 2007; Nelson and Irwin, 2014; Pratt et al., 2006; Reay et al., 2017). Establishing an occupational identity is often a long-term process, and therefore an occupation's identity is highly valued, stable, and resilient to change (Pratt et al., 2006; Reay et al., 2017). Consequently, changes in employees' work practices are often perceived as a threat to their occupational identity (Craig et al., 2019; Kim and Kankanhalli, 2009; Lapointe and Rivard, 2005; Marakas and Hornik, 1996; Petriglieri, 2011).

The introduction of new IS often poses a foundational shift in employees' existing work processes, resulting in changing occupational identities (Carter and Grover, 2015; Craig et al., 2019; Stein et al., 2013; Vaast and Pinsonneault, 2021). For instance, a study in the healthcare sector by Nach (2015) showed that doctors and nurses perceived a threat to their identities after a new electronic health records system was introduced. Consequently, professionals responded with mechanisms to redefine their threatened occupational identity. At the same time, other studies highlight a positive change in employees' occupational identity, with the positive change caused by the introduction of IS, as exemplified by Stein et al.'s (2013) finding that professionals rely on information technology (IT) implementation events as landmarks in their identity development. Moreover, the study emphasized that employees' self-understanding determined which of the technology's features and functionalities they would use.

Although prior research has shown that the introduction of IS affects employees' occupational identity, little is known about: (1) changes in occupational identity in the context of AI systems; and (2) dimensions that shed light on why AI systems

Table 17.1 Overview of case organizations

	BankOrg	InsuranceOrg	AutomotiveOrg
Industry	Banking	Insurance	Automotive
Type of AI system	Supervised learning	Supervised learning	Supervised learning
Area of application	Loan consulting	Customer requests	Executive development
Development of AI system	External vendor	External vendor	External vendor
Occupational group affected by the AI system	Loan consultants (high-skilled workers)	Service center employees (low-skilled workers)	Human resources developers (high-skilled workers)
Type of decision-making practice	Automation and augmentation	Automation	Augmentation
Impact of decision-making practice	High impact	Low impact	High impact

affect employees’ occupational identity in a certain way. First, it is important to understand how AI systems affect employees’ occupational identity, because AI systems come with unique characteristics, including their autonomy, learning ability, and opacity (Benbya et al., 2021; Burrell, 2016; Faraj et al., 2018). The AI systems’ unique characteristics shape the relationship between the technology and employees’ occupational identity, and transform our understanding of identity formation around technology (Endacott, 2021). For instance, AI’s capability to learn drives the need to reconceptualize employees’ identity construction processes as a shared process between employees and the AI system (Endacott, 2021). Second, exploring dimensions that identify key differences in how AI systems can affect occupational identity allows us to draw a holistic picture of AI’s consequences for employees and the impact on occupations. Moreover, this perspective helps to derive insights for managers who consider introducing AI systems for decision making practices, to better navigate the perils of AI systems in the workplace.

METHOD

Research Context

We make use of qualitative insights from three case organizations that introduced AI systems for decision-making practices. Access to all three organizations allowed us to gain a deep understanding of the respective AI system itself, its impact on employees’ work processes, and changes to employees’ perception of their occupation. Table 17.1 gives an overview of our three case organizations.

BankOrg

BankOrg is a large traditional German bank with around 135,000 employees. The bank is known for its regional commitment, social engagement, and reliability as an employer. Although BankOrg is a traditional bank with many regional branches and

a strong focus on personal customer engagement, the organization invests heavily in advanced digital technologies to enhance customer service and ensure competitiveness with online banks. A recent example is the introduction of AI systems in loan consultancy.

BankOrg introduced an automating AI system for private loan consulting and an augmenting AI system for commercial loan consulting. Before the two systems were introduced, the bank had encountered high loan default rates in these consultancy segments. Reasons for the high loan default rates included personnel shortage and the flexibility in loan consultants' decisions. The consultants often found it hard to decline customers' loan requests, especially in regional branches where the consultants had known their customers for a long time. To reduce BankOrg's high default rates, its management invested in AI-based solutions from an external provider, that promised to assess and predict customers' creditworthiness based on historical customer data and behavior.

The AI system implementation in both the private and the commercial loan consultancy led to fundamental changes in employees' work processes. Before the system was introduced, a customer had to make an appointment with a consultant. During the initial appointment, the consultant would discuss the customer's loan request with them, and then make a preliminary decision on whether to grant the loan or not, based on the consultant's experience and expertise. The customer would then receive a list of documents to be collected and turned in before the consultant could make the final decision; in the case of private loan consulting, relevant documents included the customer's employment contract, income pay slips, and credit history. During a second appointment, the consultant would then make the final decision on the loan request, and in the case of a loan approval, the consultant would also determine the terms and conditions, such as the loan amount, the duration of the loan, and the relevant interest rates. An internal department would then check the preliminary contract before the customer would receive the money.

The AI systems' introduction simplified the entire process for the consultants and customers alike: now, the customer would still make an appointment with the consultant, who enters the customer's data into a predefined data mask. The AI system would then make a loan decision based on the automated extraction of external data (for example, customers' credit history). Moreover, the system would determine all relevant terms and conditions in the case of approval, and the customer may sign the contract immediately without the need for an additional appointment.

Although the AI systems for private and commercial loan consulting work similarly, they differ in two ways. First, the system used for private loan consulting is trained to predict an individual's creditworthiness based on various parameters, such as the individual's income, credit history, employment status, and age. In contrast, the AI system used for commercial loan consulting is trained to predict a commercial customer's request, which includes parameters that relate to the customer's business, such as sales, profit, and costs.

Second, the AI system used for private loan consulting is used for decision automation. This leaves consultants without an opportunity to intervene in the

decision-making process, and they are also not able to adapt or overrule the decision outcome. This means that the loan consultants are forced to communicate the AI-derived decision to the customer even if the consultants do not understand or agree with the outcome. In contrast, the AI system used in commercial loan consultancy is implemented as an augmenting decision system. This means that loan consultants must use the system to derive a decision, but have the ability to adapt or overrule the AI-made decision. If consultants do not agree with the AI-derived decision, they can fill out an exception sheet to explain why they think the decision should be different.

InsuranceOrg

InsuranceOrg is a medium-sized multinational insurance company with about 3,000 employees, headquartered in Germany. The company offers various types of insurance, including life, car, and house insurance. For each insurance segment, the company has specialized departments. The employees of these specialized departments advise customers on products and services, assess damage requests, and adjust contracts. Beyond these specialized departments, the company has one general customer service center that is responsible not only for assigning customer requests to the right department, but also for answering general customer requests. Moreover, the service employees were also responsible for deciding which requests were urgent by indicating the level of urgency per request from 1 (not urgent) to 5 (very urgent). This preselection helped employees in the specialized departments to better answer the specific customer requests. Owing to a growing number of customer requests and a skills shortage in the customer service department, InsuranceOrg decided to introduce a chatbot that automatically assigns and filters customers' requests.

Whereas, in the past, customers sent their requests directly to the customer service center, they now first interact with the chatbot. If the request is very general—for instance, if a customer is interested in pricing for home insurance—the chatbot can answer the request directly. Otherwise, the chatbot prefilters customer enquiries and assigns each request to the relevant department, where a specialized employee takes care of the request. Moreover, the chatbot assigns urgency labels to each request in order to suggest to the employees in each department which requests to answer first. Consequently, the chatbot takes over several tasks that service center employees had previously performed, particularly the decision of who is responsible for which request and which request is most urgent. The service center employees are now only responsible for the general requests that the chatbot cannot answer sufficiently or take care of on its own; for example, a change in a customer's address or account.

AutomotiveOrg

AutomotiveOrg is a large multinational automotive corporation from Germany with about 120,000 employees. The organization has multiple locations across the world, but we particularly focused on the organization's headquarters in Germany. As part of a large corporate branding campaign, AutomotiveOrg engaged in initiatives that

aimed at enhancing its diversity, fairness, and inclusion. A major initiative involved the introduction of a new human resources (HR) development strategy with the goal to implement more “objective” decisions on who to select for the executive program. As a result, the management introduced an AI system to assess candidates for the executive program, with the assessment based on the candidates’ skills and traits instead of on their demographic characteristics, their experience, and their supervisors’ performance evaluations.

Before the AI system was introduced, HR developers selected suitable candidates for the executive program in the first round, based on the candidates’ curriculum vitae (CV), seniority, and supervisors’ performance evaluations. In the second round, two HR developers interviewed the preselected candidates. Beyond structured questions, candidates also had to prepare and present two case studies as part of the interview. After these interviews, the HR developers would make their final decision. With the introduction of the AI system, the first round of selection changed. Whereas one HR developer assessed all applicants, now all candidates pass through an algorithm-based assessment center that involves video games as well as virtual case studies and role plays. These tasks aim at assessing candidates’ skills and traits, such as their risk aversion, team orientation, and analytical thinking. Based on candidates’ performance of these tasks, the AI system predicts their executive ability and chooses the candidates with the best match. Two HR developers then interview the selected candidates in a second round, after which the two HR developers make the final decision.

Data Collection

We collected our data in the three case organizations between January 2019 and July 2022. We conducted semistructured interviews with 25 loan consultants from BankOrg, with 18 service center employees from InsuranceOrg, and with 19 HR developers from AutomotiveOrg. Moreover, we were able to interview several managers from each organization, which was crucial in understanding the type of AI system and why the organization had introduced the system for decision-making practices. These interviews particularly helped us to categorize the different AI systems into systems used for automating decisions versus systems used for augmenting decisions. Furthermore, these insights helped us to understand the effect of the decision-making practices (high-impact versus low-impact effects). Our questions for the three occupational groups centered around employees’ decision-making practices, including their tasks and responsibilities before and after the AI system had been introduced, and how the perception of their work and of themselves as professionals had changed through the new technology. Moreover, we asked interviewees about their educational background, work experience, and job training level. This information was important for us to categorize the different occupational groups into low-skilled workers (for example, no or low educational degree, various prior work experiences, no specific training or education required to perform job), and

high-skilled workers (for example, high educational degree, specific training and education that is required to perform the job, and specialized knowledge).

INSIGHTS FROM OUR CASE STUDIES

All our case organizations introduced AI systems for decision-making practices. Although the AI systems themselves are similar (for example, all the AI systems are based on supervised learning), they target occupational groups with different skill levels, vary in the extent to which the AI system takes over previously human decisions, and focus on different effects of decision-making practices. In the following, we reflect on these three dimensions—(1) employees' skill-level; (2) type of decision-making practice; and (3) impact of decision-making practice—which help us to understand why we find divergent answers on how AI systems for decision-making practices affect employees' occupational identity.

Employees' Skill Level: Low-Skilled Workers Versus High-Skilled Workers

Our findings revealed two groups of employees: low-skilled workers and high-skilled workers. Low-skilled workers are characterized by no or a low educational degree and no special training or education requirements for their job. For instance, in the case of InsuranceOrg, service center employees reported that they have various work backgrounds and forms of work experience, including working as call center agents, service front-desk employees, or secretaries. To work in the customer service center, the only requirements are being able to use a computer for common office applications, plus having a broad understanding of the organization and its products and services. Moreover, our findings indicated that these employees perform a job activity with the main goal of earning a living, but without identifying through the specific work tasks. For instance, in the case of InsuranceOrg, service center employees indicated that they generally like their job but do not define themselves through their decision-making practices. For them, it is most important to have a job and to be able to earn money:

I worked as an assistant for quite a long time before I joined [InsuranceOrg]. They offered better conditions, so I changed jobs. And yeah, my job is not that exciting—I mean, I mostly click buttons, sort, and answer customer requests; but that's okay, given the fact that I haven't studied and also have a few gaps in my CV. But I'm happy here, the job pays relatively well and that's what's most important for me. (Interview, customer service employee, InsuranceOrg)

In contrast, high-skilled workers have a specialized educational degree and specific training and education that is required to perform their job. For instance, loan consultants in the case of BankOrg had to have a special banking education with a specialization in loan consulting, plus several years of training before they were able to work as fully fledged consultants. Similarly, HR developers in the case of AutomotiveOrg

require special training in management with a specialization in HR, and they need to have several years' experience in this field beyond their educational qualification. These employees tend to strongly identify themselves as professionals through their decision-making practices. For instance, loan consultants define themselves through their decision competence, creativity, and autonomy. Moreover, they perceive themselves as "creative artists," "problem solvers," and "fulfillers of customers' dreams" (quotes from interviews). Similarly, HR developers define themselves through their decision competence, empathy, and autonomy. They perceive themselves as important key players for the organization's personnel and strategic future.

Although in all three cases the AI system took over core decisions for each group, the introduction of the AI system affected the groups of employees' occupational identity differently. The first group of employees, which we refer to as low-skilled workers, either: (1) did not perceive any major changes in their occupational identity; or (2) perceived an empowerment of their occupational identity. Low-skilled workers who did not perceive a major change in their occupational identity were the employees who did not define themselves strongly through their decision-making practices. In fact, these employees' focus is on earning a living from their job activities. Their top priority is to keep their job despite the use of emerging technologies such as AI systems. Taking the example of InsuranceOrg, customer service center employees mainly feared that their job would be replaced by the new technology when the management first publicly announced the introduction of the chatbot. However, the management clearly communicated that the purpose of the chatbot was to relieve customer service employees from the increasing number of customer requests. This relief was also positively received by customer service center employees, as they now only had to focus on the requests the chatbot was unable to handle, with the result that their workload could be reduced. Moreover, the inquiries' overall processing time had decreased. This had a positive effect on customer satisfaction, which is often directly reflected in the enquiries to the customer service center employees. As a result, the customer service employees still perceived themselves as customer service employees, although the scope and complexity of their tasks had shifted due to AI system introduction:

[Name of chatbot] has not really changed anything. I mean, in the beginning we were worried that the chatbots would replace us, but when we learned how the new tool should be used, I actually liked the idea. At least the promise of having less customer requests per day was promising, because we have quite a staff shortage in our department. And you know, I don't really care what I'm doing; for me it's important that I can pay my bills and that it's a task I feel comfortable with. And it didn't get more difficult for us or anything else—we still do the same job, but tasks have shifted. But that's it. (Interview, customer service employee, InsuranceOrg)

An interesting phenomenon happened among the low-skilled workers in a context where the AI system was introduced in a neighboring, higher-skilled profession. At BankOrg, the AI system was introduced in loan consultancy but also affected service front-desk employees. Low-skilled service front-desk employees were able to be

transferred to the loan consultancy department where the AI system was used for decision automation, because now the AI system performed the main skill-related activity: deciding on loan requests. Therefore, it was no longer necessary to have specific knowledge of loan matters; consultants only needed to be able to communicate to the customers and enter data into predetermined entry fields. Former service front-desk employees were therefore suddenly able to work as loan consultants, which led to a perceived empowerment of their occupational identity:

I didn't expect that technology can do something that cool. I mean, I wear a suit now, I'm a loan consultant! And it's so nice. I have to be good in using a computer and talking to customers—which I had to do at the service front desk anyway. But it's so much better, and of course, for my CV it is also really nice. (Interview, former service front-desk employee, now private loan consultant, BankOrg)

However, the second group of employees, which we refer to as high-skilled workers, perceived the changes in their decision-making practices as a threat to their occupational identity. This group felt degraded as professionals, because the AI system replaced the decision-making practices as part of their occupational identity. Moreover, the shift in decision-making practices reduced the applicability of their special training and education which they had acquired over many years just to work in the profession.

We can find an example of this negative impact of the AI system on employees' occupational identity in the case of BankOrg's private loan consultants. While they used to perceive themselves as decision-makers, creative artists, and problem-solvers, the AI system's introduction led to their perception of being degraded to "data entry assistants" and "servants of the AI system" (quotes from interviews). As a result, many of these consultants felt less satisfied and motivated in their jobs, and considered changing departments or employers: "Anyone can do this job now. It has zero complexity anymore. I mean, everyone can enter some data into a computer. It's really like being downgraded to a stupid data entry assistant" (Interview, private loan consultant, BankOrg).

This perceived threat to high-skilled workers' occupational identity was also present in the case of the HR developers at AutomotiveOrg. However, employees' perception at AutomotiveOrg changed over time. At the beginning, AutomotiveOrg's management considered introducing the AI system to perform the entire selection process of suitable executive program candidates. HR developers would only have been responsible for communicating with potential candidates and coordinating dates for each round of interviews. This vision threatened HR developers' occupational identity. However, the HR developers raised concerns, because they feared a degradation similar to that of the loan consultants. After numerous discussions, the management finally decided that the AI system would only take over the first round of candidate selection, but that HR developers would still complete the second round, and thus make the final decision. Consequently, HR developers' decision-making

practices were only slightly affected and therefore they did not experience a major change in their occupational identity:

I understand why we as [AutomotiveOrg] are doing this. I think it was a crucial decision that we only pre-filter candidates, but we as humans still do the second round and also are responsible for the final decision. I mean, this is what my job is about and also the reason why I became an HR developer. So that's a good way in between, I think. (Interview, HR developer, Automotive Org)

Overall, our findings indicate that the introduction of AI systems for decision-making affected the occupational identity of low-skilled and high-skilled workers differently. Whereas this distinction is an important contextual factor of how AI affects occupational identity, our findings also hint at a second important difference that considers how the AI system is used, and how it changes the employees' actual decision-making practice.

Type of Decision-Making Practice: Automation Versus Augmentation

AI systems promise to outperform human knowledge and thus overcome human biases and errors in decision-making. Although human decisions in many areas can be completely automated by AI systems' decisions, the organization's management can shape how the AI-derived decisions are used: for either automating or augmenting decision-making practices. With decision automation, employees are unable to alter or overrule the AI-based decision. For instance, loan consultants in the private loan department were bound to the AI-derived loan decision. Even if they did not agree with the decision or could not understand it, they still had to communicate the decision to the customer without being able to alter or overrule the AI-based outcome. Decision augmentation, on the other hand, refers to a process in which the AI system derives a decision, but employees can use this decision as an orientation or guideline. If they do not agree with the AI-derived decision, they are still able to alter or overrule it. For instance, management encouraged loan consultants in the commercial loan department to apply the AI-derived loan decision, but in cases where they perceived a major gap between their own assessment and the AI-derived decision, they could overrule the AI outcome.

Distinguishing between AI systems for automating versus augmenting decisions is an important contextual factor for anticipating the way an AI introduction affects employees' occupational identity. As our cases show, AI systems used for decision automation may be more likely to evoke a perceived threat in employees' occupational identity. This is particularly the case for high-skilled workers, who define themselves through their decision-making practices. If these practices are transferred to an AI system, employees feel threatened in their occupational identity, which leads to dissatisfaction or fluctuation of employees. An example of an AI system that was implemented for decision automation is the private loan consulting of BankOrg. In this case, employees were not able to alter or overrule the AI-derived decision,

even when they disagreed with the decision or could not understand it. This strict dependence on the AI-derived decision caused the perception of a strong threat to employees' occupational identity: "It's just so frustrating to accept these decisions even if you know 100% that this customer would be eligible for a loan, but the system just says no—and I cannot do anything about it. So I'm really dependent on this tool and this is frustrating" (Interview, private loan consultant, BankOrg).

In contrast, in the case of low-skilled workers, the introduction of an AI system for decision automation can contribute to an empowerment of employees' occupational identity, as was the case in BankOrg. Before the AI system was introduced, loan consultants needed special training and education to be able to work as loan consultants. Now, service employees without loan consulting experience and expertise are able to take on the role of loan consultants too, because the AI system has taken over the core activity, that is, deciding on loan decisions:

Of course, if I would have to still do something with the decision [after the decision has been made] like change it, for example, if a customer complains, then I couldn't be doing this job. But the system does everything, and we have to stick to the decision. So it's then also not our responsibility if something goes wrong. (Interview, former service front-desk employee, now private loan consultant, BankOrg)

Impact of Decision-Making Practice: High-Impact Versus Low-Impact Effects of Decision-Making Practice

Our findings indicated a third dimension that is consequential in how AI systems affect employees' occupational identity: the effect of decision-making practices, which can have either a high or a low impact.

High-impact effects of decision-making practices have a major impact on the person for whom the decision is made. The person under discussion is the person affected by the decision, not the decision-maker. For instance, in the case of BankOrg, the acceptance or rejection of loan requests has a major impact on customers who apply for loans to build a house, finance a new car, or finance other relevant aspects of life. Similarly, in the case of AutomotiveOrg, hiring decisions have wide implications for both the company and the individual who is potentially selected for the executive trainee program. In these cases, employees perceived tensions regarding their occupational identity because they felt responsible for the decision outcome despite not being fully accountable for it. As one HR developer explained:

I know how important these decisions are and what kind of implications they can have. So I worry that these sensitive decisions are increasingly outsourced to a machine that does not consider personal factors or background information that can't be put into numbers. And I think I as an HR developer do have the responsibility for fair candidate selection. And I always feel this responsibility when making my decision and I try to fulfill this responsibility. But a machine doesn't feel responsible. And that's why I kind of feel a change in my role: Can I still be this responsible HR developer, although I increasingly rely on machine-based decisions? That's tough. (Interview, HR developer, AutomotiveOrg)

Similar concerns were raised by loan consultants who expressed that they feel a high responsibility for loan decisions, because they care for customers' well-being. Thus, they perceived increasing tensions in their occupational identity, because they were no longer able to meet their responsibility. Again, since, in the case of BankOrg, AI systems were used for decision automation, this further facilitated the above-mentioned negative effect. In the case of AutomotiveOrg, HR developers also raised similar concerns, but because they still had some flexibility in making the decision, concerns were not as serious as in the case of the loan consultants.

In contrast, decision-making practices with a low impact do not have major implications for individuals who are affected by the decision outcome. For instance, in the case of InsuranceOrg, the implications of a request wrongly assigned to a department are low: "I mean, what can go wrong? If the system sorts the request to a wrong department, then this means maybe longer waiting times for the customer. But that's it. It's nothing earth-shattering. So I don't see an issue with outsourcing these decisions to a machine" (Interview, customer service employee, InsuranceOrg). As a result, service center employees did not perceive any tensions toward their customer responsibility, and thus in their occupational identity.

DISCUSSION

The introduction of AI systems transforms employees' decision-making practices and consequently reshapes employees' occupational identity. However, our findings from three case organizations show that there is no one-size-fits-all answer to the question of how AI systems affect employees' occupational identity. To be more precise, our cases highlight three dimensions that are consequential if employees perceive the introduction of an AI system as a threat or as empowerment to their occupational identity. By uncovering the three dimensions, we therefore contribute to a more nuanced understanding of how AI systems affect employees' occupational identity. These insights yield several implications that can help organizations to mitigate the potentially threatening effects of AI on knowledge workers' occupational identity, and thus help to enhance the success of introducing AI for decision-making.

First, we show that the AI systems' impact on employees' occupational identity depends on employees' skill level and varies between high-skilled and low-skilled workers. This distinction adds another facet to Anthony's (2018) work, which has shown that knowledge workers react differently to new epistemic technologies, depending on their low or high occupational status. Second, we find the AI systems' impact on employees' occupational identity to depend on the purpose for which AI is used in decision-making practice, and whether the AI system is introduced for decision automation or decision augmentation. In the latter case, it appears that employees are more supportive of the system and perceive the AI system as empowerment for their occupational identity. However, if decisions are automated without the possibility to change or adapt the AI-generated outcome, employees are more likely to perceive the AI system as a threat to their occupational identity. Finally,

findings from our case companies suggest that the decision's impact on third parties is a decisive dimension in the perception of AI systems when it comes to employees' occupational identity. Notably, the AI system greatly affects employees' work processes. However, when decisions have a high impact on others, employees perceive the AI system to be rather threatening; whereas if the AI system has little impact on other peoples' lives, employees still perceive the AI system as empowering.

Although all three dimensions are consequential for employees' occupational identity, our findings show that they should by no means be understood as a distinct categorization or mutually exclusive classification. Instead, these dimensions influence each other, and are often in combination consequential in terms of whether employees perceive AI systems as a threat or as empowerment to their occupational identity. For example, on the one hand, high-skilled workers perceive AI systems as a threat to their occupational identity if the system is used for decision automation. However, if the AI system is used for decision augmentation, employees perceive the AI system more positively, and may even consider it as an empowerment to their occupational identity. On the other hand, decision automation is a prerequisite for low-skilled employees to be able to fulfill tasks that previously required specific skills, training, and education. This effect was shown, for example, in the case of BankOrg's service employees, who were able to work as fully fledged private loan consultants because the AI system automated the loan decision. This empowerment of service employees' occupational identity was only possible because of management's choice to use AI for decision automation. In this context, it is also interesting to highlight that, in contrast to the common assumption that particularly low-skilled workers will be replaced by AI systems (e.g., Frey and Osborne, 2017; Manyika et al., 2017), our findings show that under certain conditions AI systems can empower low-skilled workers by enabling them to perform tasks and occupations that were previously restricted to high-skilled workers. Our findings therefore illustrate the complex interplay between different dimensions, and show the importance of carefully considering when and how to introduce AI systems for decision-making.

Building on these insights, we show that it is crucial for organizations to consider how they use AI systems for decision-making. Our findings suggest that high-skilled employees have a strong need to maintain a sense of control and autonomy in their work. Allowing them to make decisions can help to preserve their occupational identity. That can be achieved either by leaving the final decision to the employee, or by limiting the scope of AI-made decisions to specific areas while preserving decision-making authority in others. For instance, BankOrg encouraged its commercial loan consultants to apply the AI-made decision, but still gave them the autonomy to overrule it in cases of disagreement. By leaving the final decision to the commercial loan consultants, employees were able to make use of their decision-making competence. Alternatively, in the case of AutomotiveOrg, HR developers were faced with decision automation during the first round of selecting suitable candidates, but were able to retain decision-making authority in the second round. Consequentially, there remained a task area not affected by the AI system, and thus HR developers were still able to make use of their decision-making authority.

Moreover, to ensure the successful integration of AI in knowledge work, it is important to involve both AI developers and domain experts in the AI development process. This collaborative approach is important to ensure that the AI system is aligned with the organizational goals, values, and norms. Furthermore, involving domain experts will provide insights into the knowledge work process, enabling AI developers to design systems that are well suited for the task (Mayer et al., 2023; van den Broek et al., 2021). In addition, this approach can increase the AI system's transparency and explainability, which is crucial for gaining trust and acceptance among both the organization and its domain experts.

LIMITATIONS, FUTURE RESEARCH, AND CONCLUSION

Building on insights from three case organizations, we reflect on how AI systems affect employees' occupational identity. Although we make use of in-depth insights from different industries, occupations, and AI systems, our study is not without limitations. First, it is impracticable to account for all differences between the case organizations, and for their potential effects on employees' occupational identity. Although we selected three different case organizations and thus offered a more valid approach across different industries, additional comparative studies may be needed to further improve the generalizability of our findings as a function of decision automation and decision augmentation.

Second, based on our different case studies and the emerged dimensions, we make use of a comparative analysis to uncover the effect of AI systems on employees' occupational identity. Yet, there may be additional dimensions that might play an important role. For example, cultural differences may have a decisive role in shaping employees' responses to the introduction of AI systems. Moreover, all AI systems introduced in our case organizations were trained using supervised learning. However, different AI systems (for example, unsupervised learning, deep learning, and so on) may have differing effects on employees.

Third, further research is needed to better understand the complex interplay between employees' occupational identity and an organization's values and norms, with the interplay being reflected in employees' perceived organizational identity, particularly in the context of AI adoption. Our study suggests that the introduction of AI can lead to changes in employees' occupational identity. However, it remains unclear to what extent AI systems also affect employees' perception of their organization's values and norms, and thus the organization's identity. For instance, if an AI system that fully aligns with an organization's identity is implemented, how would it impact employees' occupational identity? Is it possible that the AI system could first influence employees' occupational identity and, in turn, shape the organization's identity? Future research should therefore examine this dynamic relationship critically and identify the underlying interdependencies.

Overall, our chapter offers nuanced insights into how AI systems for decision-making affect employees' occupational identity. Building on qualitative

insights from three case organizations, we shed light on three dimensions that shape the way AI systems affect occupational identity. Our findings provide implications for researchers and practitioners alike who are interested in how AI systems are used and perceived in the workplace for decision-making, and how potential pitfalls can be mitigated.

REFERENCES

- Anthony, C. (2018). To question or accept? How status differences influence responses to new epistemic technologies in knowledge work. *Academy of Management Review*, 43, 661–679.
- Bailey, D., Faraj, S., Hinds, P., von Krogh, G., and Leonardi, P. (2019). Special issue of organization science: Emerging technologies and organizing. *Organization Science*, 30, 642–646.
- Baird, A., and Maruping, L.M. (2021). The next generation of research on IS use: A theoretical framework of delegation to and from agentic IS artifacts. *MIS Quarterly*, 45, 315–341.
- Benbya, H., Pachidi, S., and Jarvenpaa, S.L. (2021). Special issue editorial. Artificial intelligence in organizations: Implications for information systems research. *Journal of the Association for Information Systems*, 22, 281–303.
- Berente, N., Gu, B., Recker, J., and Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, 45, 1433–1450.
- Burrell, J. (2016). How the machine “thinks”: Understanding opacity in machine learning algorithms. *Big Data and Society*, 3, 1–12.
- Carter, M., and Grover, V. (2015). Me, my self, and I(T): Conceptualizing information technology identity and its implications. *MIS Quarterly*, 39(4), 931–957.
- Chreim, S., Williams, B.E., and Hinings, C.R. (2007). Interlevel influences on the reconstruction of professional role identity. *Academy of Management Journal*, 50, 1515–1539.
- Christin, A. (2020). *Metrics at Work: Journalism and the Contested Meaning of Algorithms*. Princeton University Press.
- Craig, K., Thatcher, J.B., and Grover, V. (2019). The IT identity threat: A conceptual definition and operational measure. *Journal of Management Information Systems*, 36, 259–288.
- Dourish, P. (2016). Algorithms and their others: Algorithmic culture in context. *Big Data and Society*, 3, 1–11.
- Endacott, C.E. (2021). *The Work of Identity Construction in the Age of Intelligent Machines*. Dissertation, UC Santa Barbara, California.
- Faraj, S., Pachidi, S., and Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. *Information and Organization*, 28, 62–70.
- Frey, C.B., and Osborne, M.A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114, 254–280.
- Giermendl, L.M., Strich, F., Christ, O., Leicht-Deobald, U., and Redzepi, A. (2022). The dark sides of people analytics: Reviewing the perils for organisations and employees. *European Journal of Information Systems*, 31(3), 410–435.
- Hafermalz, E., and Huysman, M. (2021). Please explain: Key questions for explainable AI research from an organizational perspective. *Morals and Machines*, 1, 10–23.
- Kim, H.-W., and Kankanhalli, A. (2009). Investigating user resistance to information systems implementation: A status quo bias perspective. *MIS Quarterly*, 33, 567–582.
- Lapointe, L., and Rivard, S. (2005). A multilevel model of resistance to information technology implementation. *MIS Quarterly*, 29, 461–491.
- Lebovitz, S., Lifshitz-Assaf, H., and Levina, N. (2022). To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization Science*, 33, 126–148.

- Leicht-Deobald, U., Busch, T., Schank, C., Weibel, A., Schafheitle, S., Wildhaber, I., and Kasper, G. (2019). The challenges of algorithm-based HR decision-making for personal integrity. *Journal of Business Ethics*, 160, 377–392.
- Manyika, J., Lund, S., Chui, M., Bughin, J., Woetzel, J., Batra, P., Ko, R., and Sanghvi, S. (2017). Jobs lost, jobs gained: What the future of work will mean for jobs, skills, and wages. *McKinsey Global Institute Report*.
- Marakas, G.M., and Hornik, S. (1996). Passive resistance misuse: Overt support and covert recalcitrance in IS implementation. *European Journal of Information Systems*, 5, 208–219.
- Mayer, A.-S., Strich, F., and Fiedler, M. (2020). Unintended consequences of introducing AI systems for decision making. *MIS Quarterly Executive*, 19, 239–257.
- Mayer, A.-S., van den Broek, E., Karacic, T., Hidalgo, M., and Huysman, M. (2023). Managing collaborative development of artificial intelligence: Lessons from the field. In: *Proceedings of the 56th Hawaii International Conference on System Sciences*.
- Nach, H. (2015). Identity under challenge: Examining user's responses to computerized information systems. *Management Research Review*, 38, 703–725.
- Nelson, A.J., and Irwin, J. (2014). Defining what we do—all over again: Occupational identity, technological change, and the librarian/Internet-search relationship. *Academy of Management Journal*, 57, 892–928.
- Petriglieri, J.L. (2011). Under threat: Responses to and the consequences of threats to individuals' identities. *Academy of Management Review*, 36, 641–662.
- Pratt, M.G., Rockmann, K.W., and Kaufmann, J.B. (2006). Constructing professional identity: The role of work and identity learning cycles in the customization of identity among medical residents. *Academy of Management Journal*, 49, 235–262.
- Rai, A., Constantinides, P., and Sarker, S. (2019). Editors comments: Next-generation digital platforms: Toward human–AI hybrids. *MIS Quarterly*, 43, iii–ix.
- Raisch, S., and Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review*, 46, 192–210.
- Reay, T., Goodrick, E., Waldorff, S.B., and Casebeer, A. (2017). Getting leopards to change their spots: Co-creating a new professional role identity. *Academy of Management Journal*, 60, 1043–1070.
- Stein, M.K., Galliers, R.D., and Markus, M.L. (2013). Towards an understanding of identity and technology in the workplace. *Journal of Information Technology*, 28, 167–182.
- Strich, F., Mayer, A.S., and Fiedler, M. (2021). What do I do in a world of artificial intelligence? Investigating the impact of substitutive decision-making AI systems on employees' professional role identity. *Journal of the Association for Information Systems*, 22, 304–324.
- Susskind, R., and Susskind, D. (2015). *The Future of the Professions: How Technology Will Transform the Work of Human Experts*. Oxford University Press.
- Vaast, E., and Pinsonneault, A. (2021). When digital technologies enable and threaten occupational identity: The delicate balancing act of data scientists. *MIS Quarterly*, 45(3), 1087–1112.
- van den Broek, E., Sergeeva, A., and Huysman, M. (2021). When the machine meets the expert: An ethnography of developing AI for hiring. *MIS Quarterly*, 45(3), 1557–1580.
- von Krogh, G. (2018). Artificial intelligence in organizations: New opportunities for phenomenon-based theorizing. *Academy of Management Discoveries*, 4, 404–409.

18. Imagination or validation? Using futuring techniques to enhance AI's relevance in strategic decision-making

Andrew Sarta and Angela Aristidou

Strategy scholars across several disciplines have theoretically and empirically demonstrated the challenges of strategy-making and the difficulties of setting strategic direction (Baumann et al., 2019; Ocasio et al., 2022; Sund et al., 2016). This long line of inquiry offers valuable insights into the role of selecting issues as a boundary to the strategy-making process. Issues represent problems, opportunities, or threats to an organization (Crilly and Sloan, 2012; Joseph and Ocasio, 2012; Ocasio and Joseph, 2005). Detecting issues is a complex process. Organizational structures place key players, typically middle managers or experts, in positions to focus their attention on specific areas of the environment (Dutton et al., 2001; Greenwood et al., 2019; Huising, 2014; Ocasio, 1997). The issues detected by these key players are then channelled through a number of filtering mechanisms (for example, meetings, corporate events, emails) that enable a subset of issues to bubble to the top managers of organizations (Joseph and Ocasio, 2012; Starbuck and Milliken, 1988). Filtering allows some issues to be weeded out so that top managers can focus on the most important and salient issues to set the strategic agenda of the organization (Bundy et al., 2013; Dutt and Joseph, 2019; Ocasio, 1997). Of course, filtering is a process of selection. Key players filter issues before channeling them to top managers who then select the most important issues. Levinthal (2021, p. 7) insightfully referred to these processes as occurring in the “artificial selection environment” where issue selection does not result from market-based competition for high-priority issues, but from conscious choices made by managers and experts. The processes are distinctly individual and organizational, with the intention of limiting the number of issues in order to accommodate attentional capacity, even if the solutions to filtered-out issues are available in other organizations (Bansal et al., 2018; Ocasio and Joseph, 2005; Simon, 1978). The difficulties emerging in each of these two key inflection points—that is, issue detection and artificial selection—form fundamental pillars that span several decades of study in strategy-making.

Against this backdrop of accumulated knowledge in strategy-making, the potential impact of artificial intelligence (AI) on strategy-making in organizations is still indeterminate. Different forms of AI are undoubtedly impacting several organizations across multiple industries by conducting tasks ranging from financial management to healthcare diagnostics (von Krogh, 2018). Forms of generative AI that have recently captured attention in multiple aspects of our lives, such as large language models

(LLMs), for example, Bard AI and GPT-4, reduce the time it takes to fully analyze a complex topic from weeks to minutes. For organizations, AI can now efficiently automate tasks (likely by replacing the tasks that humans could perform), or augment human responsibilities by freeing time for humans to apply ingenuity and creativity to their work (Raisch and Krakowski, 2020). The latter emphasizes a new human–AI collaboration versus a human–AI substitution. As we collectively experiment with the types of tasks AI can perform, it becomes increasingly apparent that AI can now contribute not only to repetitive functions, but also to creative tasks in organizations, such as screening new hires or writing marketing material (Amabile, 2019). With such a wide-ranging impact on organizational tasks, we believe that the time is right to interrogate how AI might impact the most complex tasks in organizations: making strategic decisions. Strategic decisions are complex because of the interdependencies across multiple domains and, more specifically, “strategic decisions are intertemporally interdependent, shaping and guiding decisions potentially far into the future” (Leiblein et al., 2018, p. 559). While the computational benefits of AI and its impact on organizations are increasingly apparent, it is less clear how AI may change the core strategy-making processes of issue detection and artificial selection to result in different outcomes than those experienced today.

In this chapter, we first articulate AI’s promise in relation to each of the two key inflection points in strategy-making (issue detection and artificial selection). We then draw on literature in both AI in organizations and behavioral-based organization theory to identify and explain the barriers that are likely to stall AI’s influence in the strategy-making process. Finally, we propose what we refer to as “futuring” techniques to overcome barriers in the strategy-making process. We argue that futuring moves organizations closer to realizing AI’s promise in enhancing strategy-making both now and in the future.

AI AND ISSUE DETECTION IN STRATEGY-MAKING: THE PROMISES AND THE PROBLEMS

Detecting issues is widely acknowledged as a key boundary and core input into the strategy-making process. The sheer number of potential issues that could impact organizations is a challenge on its own, since top managers are constrained in what they can pay attention to (March and Simon, 1958; Ocasio, 1997). Resolving this challenge often involves limiting the scope of attention to more specific environments rather than all possible environments. Top managers divide tasks among organizational members to ease the processing burden while remaining attentive to as many environments as possible (Joseph and Gaba, 2019). In addition to focusing attention on organizational priorities, top managers also have good reasons for constraining the search for issues. As organizations grow, there are exponential increases in coordination costs due to the numerous interdependencies between functions, consumers, stakeholders, and employees (Baumann and Siggelkow, 2013; Chen et al., 2019; Simon, 1962). With issues seemingly coming from everywhere, “top-level

decision makers are bombarded by a continuous stream of ill-defined events and trends. Some of these events and trends represent possible strategic issues for an organization” (Dutton and Jackson, 1987, p. 76). The noise (or variance) that comes from ill-defined issues is subsequently compounded, clouding the process of initially detecting issues before determining the issues that should be channelled through to the executive suite.

Organizations are designed to minimize noisy issues that tax managerial attention, by structurally dividing up cognitive tasks into specialized functions (March and Simon, 1958; Ocasio and Joseph, 2005). For example, if a new competitor were to threaten an organization by offering similar products and services, the marketing functional lead is well positioned to detect the threat. Similarly, a potential regulatory threat is more likely to be detected by the head legal counsel. Roles and functions merely exemplify structural processes that detect issues by directing managers’ attention. These processes have been demonstrated to detect a variety of issues, ranging from sustainability concerns to industry deregulation (Cho and Hambrick, 2006; Fu et al., 2019). Delegating attention to functions is the first step in detecting issues, which then travel through communication channels (for example, meetings, emails) from functions to top managers (Dutt and Joseph, 2019; Joseph and Ocasio, 2012). Collectively, whittling down the multitude of potential issues that an organization could face is the result of (at least) a double selection process inside organizations: functions first determine the issues of importance and pass these issues through to top managers who then prioritize those issues to formulate strategy (Bundy et al., 2013; Levinthal, 2021; Ocasio et al., 2018). As the number of “issue detectors” and mediators increases, so do the selection effects.

The Promise of AI in Issue Detection

The first major challenge in strategy-making, prior to selection effects, is to ensure that the right issues are detected for processing. Detecting issues is subject to three primary limitations that are known to strategy scholars: attention scarcity, attention grain, and experiential bias (Bansal et al., 2018; Gavetti, 2012; Gavetti and Levinthal, 2000). AI holds the promise to overcome these three known limitations in strategy-making processes.

The limitation of attention scarcity is one of cognitive capacity; therefore, expanding capacity becomes a means through which the attention scarcity limitation can be relieved (Simon, 1973). Attentional scarcity is often depicted as an inability to search so-called distant terrains, which generates myopia and limits imagination in the strategy-making process (Gavetti, 2012). AI’s computational advantage promises to scan a broader array of environments for issues compared to human issue detectors, and in a fraction of the time. AI is also notably strong in automating rule-defined tasks (that is, tasks that can be delegated with specific instructions), or augmenting the efforts of humans by identifying patterns in large datasets (Raisch and Krakowski, 2020). Of course, in order to obtain a meaningful suggestion that an issue should be considered, AI systems require some definition of inputs (for

example, data including sound, text, or images) and algorithms through which to analyze the data (von Krogh, 2018). This enables AI to explore environments that humans may not have the time to explore, which opens up the possibility that new issues are detected in distant terrains, or new issues are determined through abstract combinations of proximate issues.

“Attentional grain” limitations relate to the detail and effort required to gain a granular understanding of any particular issue (Bansal et al., 2018; Kahneman, 1973). With AI, this limitation is alleviated, as the time constraint of effort is reduced. Given the right instructions (that is, training data and algorithms), AI can search in areas of a defined issue environment with depth, and generate connections between details of issues that otherwise go unnoticed. For example, it is exceedingly difficult for a physician to examine all of the electronically scanned photos for skin conditions in a given database when diagnosing the potential presence of a malignant tumor for a patient. Time and fatigue limit the degree to which attentional effort can be applied. Instead, physicians rely on prior patient experiences or knowledge gained through schooling to diagnose problematic skin conditions. The staunch strategy scholar can draw analogies to a competitive situation. For example, analyzing the earnings call transcripts in every quarter for every possible competitor is a fatiguing task to undertake when seeking to detect the emergence of possible competitive threats. AI can overcome this attentional grain limitation and detect potential issues through combinations that are typically overlooked by top managers, since fatigue is reduced by orders of magnitude. So long as enough computation is provided on a well-trained dataset, unique combinations can emerge to effectively predict emerging threats.

Finally, top managers are often experientially biased to heuristically favor familiar experiences when detecting issues (Gaba et al., 2022; Gavetti and Levinthal, 2000; Tversky and Kahneman, 1974), which opens an opportunity for AI to detect novelty in issues. By biasing the search for issues toward prior experiences, top managers gain efficiency by resolving familiar issues quickly with known solutions. At the same time, experience is ambiguous and can often be misinterpreted, which can result in mismatching prior solutions to new issues. For example, a start-up offering novel pharmacological drugs may be overlooked if it resembles a start-up that failed previously (March, 2010; Starbuck and Milliken, 1988). Despite the efficiency advantages of experience, overlooking issues and failing to act can naturally be quite costly. If misdiagnosis generates a failure to act, the emergence of a start-up can occasionally outcompete strong incumbents (Christensen and Bower, 1996). AI’s rule-based search may be able to detect subtle differences that counter experiential biases by examining issue attributes that might otherwise be overlooked by top managers. The ability to combine and recombine attributes in issue detection holds potential to generate novelty in issues (Verhoeven et al., 2016), which presents information to top managers in a different light. In this way, AI may contribute to detecting new issues without necessarily exploring new environments, as issue attributes are combined in unique ways.

Problems in Issue Detection with AI

Creeping attentional scarcity

The first problem for AI and its potential impact on strategy-making is that input definition and algorithms are choice variables made by individuals in organizations. Delegating the attention of an organizational function to an AI system over a human (for example, leveraging AI to detect new competitive threats by scanning adjacent industry newspaper articles) is constrained by the initial rules built into the AI system. If the input for detecting new competitive threats is constrained to scraping newspaper articles across the internet and analyzing the text data with anomaly-detecting natural language processing algorithms, then new competitive threats that emerge on social media will go largely undetected. GPT-4, as of the time of this writing in March 2023, is trained on data up to and including September 2021, which restricts its ability to generate prose on contemporaneous events. Of course, data and algorithms are frequently updated to improve detection; however, the updates are typically choices by agents inside organizations to point an AI system toward an issue of importance. Put differently, the environment where algorithms detect issues is baked into the decision of how AI systems are adopted, which relegates AI to detecting new patterns based on pre-defined issue criteria set out by the very managers and functional leads that AI is purported to benefit. It is not AI's ability to search multiple environments that produces the problem, but the data on which AI is trained that restrict the potential issues that can be detected.

Data-driven path dependence impact on attention grain

Selecting data creates a multitude of challenges for issue detection that result from historically tied decision-making. Crudely put, path dependence accumulates in mundane organizational choices (for example, who are the competitors?), which informs present and future decisions on the issues important to an organization's strategy. These choices likely prevent organizations from changing (Sydow et al., 2020). Top managers are known to stabilize present decisions around prior successes (Audia and Greve, 2021; Levinthal, 1997). More specifically, top managers tend not to sway too far from past learning. This benefits attention allocation, since solutions to new issues can be applied quickly; however, this tendency also produces temporal myopia (Levinthal and March, 1993). The problem with AI's promise in improving attentional grain is that AI is typically used to find solutions, not issues (Pietronudo et al., 2022). This seemingly subtle framing choice can lead to issues being missed (Bansal et al., 2018; Henderson and Clark, 1990). If AI is deployed with a solution-oriented framing constrained by path dependence, it may offer little impact to the attentional grain challenges that restrict issue detection among top managers. The result is a garbage can-like scenario where AI is a solution looking for problems (Cohen et al., 1972; Pietronudo et al., 2022), rather than a system looking for issues.

Further, if AI is to be deployed in strategy-making to any degree, it is likely to be constrained to answer issues that top managers have already defined. Much as GPT-4 responds to a prompt in a dialogue box, a strategic application of AI would involve

inputting a question to which AI assists in finding an answer. Path dependence implicitly emerges in two forms: what question a top manager asks, and what data were used to train the algorithm. For a moment, let us assume that the training data are as wide as possible and path dependence is limited to the question being asked. What can a top manager be expected to do? First, they are likely to ask a question that they believe is relevant to the future direction of the organization (assuming the purpose is strategy-making). The question itself emerges from prescient issues in the sense that they are likely known issues that the manager is struggling to discern. Putting aside the notion that the question itself is experientially biased (that is, path dependent), what do we expect the manager to do with an answer that is surprising, or contradictory to their own intuition? The manager is left with multiple choices. One is to reframe the question until a satisfactory response is received. Alternatively, the manager can choose between their own intuition or the recommended solution from AI. Recent research in cognitive psychology suggests that aversion to algorithms emerges as identity relevance and ambiguity both increase (Morewedge, 2022). If AI offers superior solutions to a top manager, the identity of that top manager is likely threatened. In addition, strategic decisions are laden with ambiguity. In combination, the simple thought experiment opens up the possibility (or likelihood) that top managers will leverage AI to validate their initial intuition, rather than to detect new issues in the environment. The promises that AI posits on improving attentional grain, therefore, become prisoner to solution-oriented framing that fails to capitalize on superior computational power.

Restricted data-sharing and the challenge of experiential bias

In addition to framing challenges, AI's promise to overcome experiential biases and detect new issues through recombined attributes is constrained by independent datasets. Data are the lifeblood of AI and many organizations are strategically accumulating as much data as possible to improve their algorithms. Of course, competitive advantages in AI rely on proprietary access to data, which limits interlinkages between datasets. Neumann et al. (2022) refer to this lack of sharing as data deserts, in which the lack of database sharing, whether intentional or unintentional, decreases the recombination potential of detecting new issues and restricts AI's ability to generate either accuracy or novelty in issue detection.

AI AND THE SELECTION ENVIRONMENT IN STRATEGY-MAKING: THE PROMISES AND THE PROBLEMS

Once issues are detected by various functions and managers in organizations, a set of selection processes take hold in organizations before issues are presented to top managers to formulate the strategic agenda (Bundy et al., 2013; Dutton and Jackson, 1987). Strategic agendas are crucial to ensuring that an organization manages its attention, and that its members focus on finding solutions to pivotal issues. Our focus

in this section is decidedly on the internal environment that processes the array of issues detected. Because choices are made regarding the issues retained by organizations, we adopt the aforementioned notion of an artificial selection environment (Levinthal, 2021), and discuss the promises and problems of AI within such an environment.

The Promise of Minimizing Excessive Filters in Artificial Selection Environments

Overcoming short-termism

Strategy scholars across several disciplines have theoretically and empirically demonstrated the difficulty of long-term strategy-making despite the potential benefits of anticipating future outcomes (Elsbach et al., 1998; Levine et al., 2017). Organizational scholars point to the attentional challenges of long-term thinking where organizational members favor proximate solutions to aspiration shortfalls (Cyert and March, 1963), select narrow solutions based on prior experiences (Levinthal and Posen, 2007), learn from short-term feedback loops rather than long-term feedback loops (March, 1991), favor loyal existing customer bases over new distant customer bases (Christensen and Bower, 1996), or struggle with intertemporal valuations of alternatives (Laverty, 1996), among others. The tendency for organizations is clear: short-term decision-making processes often crowd-out long-term decision-making processes due largely to the uncertainty and ambiguity of assessing opportunities in the future. Despite this challenge, scholars continue to consider organizational anomalies that appear to depict long-term strategy as a holy grail (Gavetti, 2012; Gavetti et al., 2012; Gavetti and Menon, 2016). AI's ability to process greater amounts of information allows for a larger number of issues to be assessed. Humans tend toward proximate solutions or shorter feedback loops because of time constraints and effort requirements in the attentional process (Cyert and March, 1963; Kahneman, 1973). When a greater number of issues can be evaluated, it becomes increasingly likely that distant opportunities are compared alongside proximate issues, to potentially inform strategic thinking. This reduced tendency to filter issues based on proximity plausibly reduces selection effects that favor the short-term tendencies of managers, and allows for long-term strategic opportunities to be considered.

Offsetting limitations of human capacity for qualitative improvements in solutions

Similar to arguments on short-termism, AI offsets the capacity limitations of human cognition to enable deeper analyses. By averting the fatigue constraint, AI can examine new information (for example, an image of a skin lesion) and repeatedly explore complex combinations of images in the dataset to calculate the probability that the lesion is benign or malignant (Goyal et al., 2020). The ability to evaluate a greater number of issues at faster speeds also provides a temporal advantage over humans that are limited in their ability to allocate, sustain, and engage their attention on a single task (Esterman and Rothlein, 2019; Simon, 1973; Tversky and

Kahneman, 1974). The result is an artificial selection environment that can spend more time assessing issues rather than detecting issues (Jarrahi, 2018). Because some of the analysis traditionally conducted by humans is substituted through AI, the modern strategic decision-maker can reallocate their time to substantiating the issues they select (Raisch and Krakowski, 2020). The freeing of time also allows managers to both examine a larger number of existing possible solutions to an issue (spatial exploration) while also projecting future possible solutions based on the trajectory of existing problems (temporal exploration). In either case, AI offers a potential qualitative improvement to strategy-making. Thus, even if the decision selected remains largely unchanged, the justification and detail informing the selection process should improve markedly.

Problems in Artificial Selection Environments with AI

Path dependent skillsets in issue interpretation

If AI presumably expands the issue evaluation process in selection environments, then the bounded nature of human decision-making should expand as well and ultimately lead to better decisions assisted by machines. Put differently, AI is task-efficient if inputs are clearly defined. If the challenges of strategy-making related only to resource allocation between humans and machines, a strong argument can be made for the advantages of AI that are not bound by attentional capacity, fatigue, and bias. Yet, an overlooked aspect of strategy-making persists alongside efficient task evaluation. The results from an AI comparison of issues likely require a different skillset to understand how suggested issues were determined. AI systems are notoriously “black-boxed” and ambiguous (Neumann et al., 2022), which creates challenges for accepting issues suggested by AI (Morewedge, 2022). Unique skillsets or socialized knowledge may therefore be required before any suggestions made by AI systems are retained in artificial selection environments (Anthony, 2021). Because top managers artificially create the selection environment (that is, they define the issues), AI may be viewed skeptically by decision-makers who rely on intuition without fully understanding how issues were recommended (Morewedge, 2022). AI recommendations are therefore likely to be consumed within existing selection processes. Put differently, AI-recommended issues are weeded out not objectively, but subjectively by decision-makers who cannot intuit why the issues were recommended. For example, if a start-up in a drastically different industry is identified as a future competitive threat, top managers may infer that the AI has hallucinated. Because AI does not have a sufficient voice to justify its decisions, dismissing and filtering the recommendation is easy. This creates a situation where strategy-making with AI looks a lot like strategy-making without AI, as typical artificial selection processes reign supreme.

Downplaying AI voices in artificial selection

The larger set of issues that can be explored raises the notion of what we playfully refer to as AI’s voice: the degree to which an AI-driven recommendation is tabled in communication channels. AI voices can sometimes reflect issues among over-

looked stakeholders of organizations (for example, customers, underrepresented stakeholders) that can be mobilized and embraced within organizations. While AI brings greater voice to some issues, a key limitation emerges: voice rights differ from decision rights (Turco, 2016). In this vein, consider some common decision-making processes that determine the selection of strategic agendas. Decisions on issues can be made democratically (by vote, with relatively equal weight among top managers) or autocratically (where specific individuals have more influence over final outcomes). The latter can be divided into autocracy by authority (for example, a domineering chief executive officer) or autocracy by expertise (for example, deferring to the subject matter expert) (Melone, 1994). Autocracy by authority is invariably influenced by authority itself more than the content of the issue being presented. As a result, we largely expect these processes to be governed by power relationships that are unimpacted by AI specifically (Casciaro and Piskorski, 2005). Likewise, democratic processes present few opportunities for AI to truly alter the strategic agendas of organizations due to voice weight. When each functional or business unit manager presents issues believed to be important alongside AI-generated issues, AI systems have voice rights, but only gain decision rights if most decision-makers agree. Because the AI becomes one voice among many, with equal weight of voting, decision rights are low. The result is more issues generated, but the same issues selected, leaving strategy-making relatively unchanged.

AI AS A VALIDATING TOOL IN STRATEGY-MAKING

Within classic strategy-making processes involving issue detection, there are a number of long-standing challenges that strategy scholars would converge upon, including attentional scarcity, attentional grain, and experiential bias. Similarly, in processes of artificial selection, scholars might identify short-termism and human capacity limitations as central challenges. We have presented each, and articulated AI's promises to radically improve strategy-making. We have also identified a number of problems that act as barriers to AI's promised improvements in issue detection and artificial selection. These problems are surfaced through a careful examination of our current understanding of how AI tools are trained, developed, audited, and used in organizations. We believe that creeping attentional scarcity, data-driven path dependence, restricted data-sharing, path dependent skillsets, and downplayed AI voices are formidable sources of inertia in strategy-making that dampen the posited promises.

Through our analysis, one would be led to the conclusion that AI is likely to generate outcomes in strategy-making that look very much the same as those experienced without AI. Strategy-making essentially relegates AI as a validating tool, where we would expect that well-established processes for defining and selecting important strategic issues constrain AI to merely confirm managerial intuition, which leaves strategic decision-making relatively unimpacted.

An alternative way to leverage AI in strategy-making is to highlight its potential to act as an “imaging” tool that enables managers to generate new future possibilities. Most strategies are built on specific beliefs about interpretations of the past and projections of the future that are shared collectively within the organization or among top managers. Surfacing, sharing, debating, and considering alternative future strategies can be supported by AI to enhance imagination as opposed to validation among top managers. We argue that a set of futuring techniques will allow top managers to shift toward a more fruitful approach to strategy-making that enables AI to improve the persistent challenge of long-term strategic thinking.

FROM VALIDATION TO IMAGINATION THROUGH FUTURING TECHNIQUES

AI can contribute to strategy-making in organizations by foregrounding AI’s strengths alongside a range of futuring techniques that quell the behavioral tendencies of top managers. Futuring can be defined as identifying and evaluating possible future events. Similar to strategy-making conventional wisdom, futuring requires framing reasonable expectations, identifying emerging opportunities and threats to the company or organization, and anticipating actions that promote desired outcomes. In addition, futuring specifically aims to identify mechanisms in a future that does not yet exist.

We propose and identify three techniques—crafting time horizons; timing and pacing; voice clearing—that collectively encompass futuring. We argue that these techniques support top managers interested in delivering upon AI’s promise, and allow top managers to consciously transition away from the established problems that limit AI’s impact in strategy-making. Two specific futuring techniques become particularly salient in overcoming problems in issue detection—crafting time horizons; timing and pacing—while voice clearing becomes especially salient in offsetting the problems produced by artificial selection. We present and discuss each in turn.

Enhancing Issue Detection through Futuring Techniques

Crafting time horizons as a technique to enhance issue detection

“Time horizons” refer to the distances into the past and future that organizational actors take into account when considering issues and solutions in strategy-making. Time horizons may vary depending on the topic and the leadership style, being broad or narrow, expanding or contracting. Time horizons may equally be affected by institutional rhythms (that is, chief executive officer tenure) and broader sociocultural temporal orientations (Chen and Nadkarni, 2017). These are impactful choices, because once particular time horizons are determined, they become embedded in strategic planning and resource allocations. Importantly, the choice of time horizons in strategy-making may exacerbate temporal myopia. Circumventing temporal myopia is one of the most studied challenges in strategic decision-making

(DesJardine and Shi, 2020; Lavery, 1996; Levinthal and March, 1993; Levinthal and Posen, 2007). Often leaders' ability to detect issues in time, and leaders' ability to work with time to generate forward-looking solutions, is constrained by their choice of time horizons, combined with the noted limitations of attentional scarcity and attentional grain (Ocasio, 1997; Simon, 1973). AI offers the potential to relieve the constraint of working on forward-looking solutions when top managers deploy AI on issues with a futuring lens. We explain how, and propose that this is most likely to occur in less experienced organizations.

AI allows top managers to craft multiple possible time horizons and explore multiple scenarios in each time horizon, without compounding the limits on attentional scarcity and grain. The technique of using AI to craft time horizons is characterized by three attributes: (1) reframing AI to detect issues versus solutions; (2) relying on less experienced managers to establish the issue parameters; and (3) casting the widest possible search space that data allow.

Redirecting uses of AI away from the solution space and toward the issue detection space allows a greater number of issues to be considered (Gavetti, 2012). Identifying more issues is a necessary but insufficient condition to improve the quality of issue detection, since managers will continue to make issue detection choices (that is, where to search) based on prior experiences (Levinthal, 2021). In this case, AI leans toward validation and away from imagination. We argue that organizations with less influence from the past will be better suited to break away from these temporally myopic decisions, and that those organizations with more experience must break away from the experiential tendencies of top managers. Ironically, this suggests that less experienced managers should define the parameters used to detect issues for strategy-making. Arguably, this reduces the strength of path dependence and enables a naïve and unbiased search. In addition, path dependent prior experiences are likely to be strongest in the most successful and well-established organizations, since they have the greatest number of successes from which to learn. We believe that this opens up opportunities for AI to benefit traditionally less successful organizations, since path dependence is weakest with fewer successes. Leaning on the intuition of less experienced managers in the most successful organizations is undoubtedly a difficult proposition.

Weaker experience bases allow AI to be deployed in more expansive ways to detect new issues. The idea behind a naïve AI search parallels brainstorming practices, where many ideas are accepted to see where they might lead. AI is then able to explore unconventional combinations to detect issues. Since experienced managers will undoubtedly play a large role in the selection process (that is, prioritizing the most important issues), our recommendation focuses on removing them from the identification process (that is, the issues to be evaluated). Having experienced managers both identifying and selecting issues is what we believe leads to validating versus imagining. Likewise, we argue for distant temporal focus in the deployment of AI as an important condition to supporting long-term strategic thinking in a way that brings imagination to strategy-making (Gavetti, 2012; Gavetti and Menon, 2016). Setting parameters to detect issues specifically (over solutions) in more expansive environ-

ments will allow AI to cast a wide net that is unencumbered by searching close to prior successes, and encouraged to search in unfamiliar terrains (Li et al., 2013). Of course, the complexity of designing such a system increases exponentially as the issue environment grows, making the likelihood of AI's impact on strategy-making modest at best. Nevertheless, the benefit of AI systems to strategy-making (initially) become a function of the ability to widen the issues being assessed by top managers without drawing additional scarce human attention (Simon, 1973). AI plays a distinct augmentation role in this respect by assuming the scanning function on behalf of top managers, effectively dumping more issues into the decision-making process.

Timing and pacing modulation as techniques to enhance issue detection

Additional key aspects of strategy concern timing and pacing. Timing involves purposefully choosing moments to undertake action. Pacing refers to modulating the speed of responses, including when to push the accelerator and when to put one's foot on the brake in ongoing processes of agenda-setting, strategy (re)design, strategic learning, innovation, and pivoting. These choices can catalyze strategy-making, influence issue detection, and prioritize issues where respective solutions exist.

AI offers the potential to relieve the constraint of working with conditions of deep uncertainty and under time pressure, when managers deploy AI with the futuring technique of timing and pacing. Imagining the future, to some degree, is considered a core strategy-making skill. Often overlooked is that imagining involves looking backwards (what has been) to inform the future (what might be) based on the present trajectory (what is happening now) (Emirbayer and Mische, 1998). It requires the introspection to examine what has worked and not worked in the past. This is important, not only for the purpose of capturing history, but also to inform the future, capturing the essence of what has come to be known as "mental time travel" (Michaelian et al., 2016; Suddendorf and Corballis, 2007) or "temporal work" by strategy scholars (Cattani et al., 2018; Kaplan and Orlikowski, 2013). Of course, the future does not yet exist; except as a possibility which has not yet been formed. AI's computational capacity to scan vast data landscapes across industries and geographies, and to connect accumulated episodic past incidents, is often hailed as a way to expand the horizons of strategy-making through projecting possibilities into the future.

Futuring techniques that modulate timing and pacing are characterized by managers using AI to: (1) consistently update issue trajectories; and (2) consistently calculate issue magnitudes. As a result, timing and pacing are most effective when the crafting time horizons technique is in place, since it allows managers to learn more about future risks, especially the type of complex events that are unlikely to occur but would have a major impact (that is, "black swans" such as a sudden crash of the stock market) (Taleb, 2007). Constructing strategic plans and detecting issues are often months-long or even year-long processes that are difficult for organizations to repeat frequently. AI, however, is capable of simulating trajectories of issues frequently, establishing probabilities, and calculating magnitudes for multiple possible chains of events that could have occurred. Timing and pacing restore AI's promise to explore possible and preferable future scenarios without taxing human attention

in a significant way. In this way, AI as an imagining tool would focus on preparing for the unknown ahead of time. Additionally, the method of exploring alternative histories can be employed to anticipate different types of risks (Krotov, 2019).

Minimizing Artificial Selection Effects through Futuring Techniques

With improvements in issue detection arguably generating more issues, discussing how AI benefits artificial selection becomes important. More issues in the strategy-making process creates a knock-on problem of clogging up the communication channels, which only reinforces the selection problem for determining what issues are prioritized, and how top managers develop strategic agendas for their organizations (Dutton and Jackson, 1987; Ocasio et al., 2018).

For a moment, let us assume that the number of issues detected from the external environment will increase with the help of AI, and that new perspectives may emerge as well. Under these circumstances, two possibilities emerge for AI to impact the strategic agenda of the organization: (1) the selection of issues by middle managers is improved, such that the issues presented to top managers are superior than would be the case without AI; or (2) the issues increase in number, so that the “AI-generated” issues are presented alongside the human-generated issues.¹ In either of the two scenarios, the top managers need to improve the likelihood of selecting AI-recommended issues for strategy-making, to meaningfully change toward focusing on longer-term issues.

To further advance in the direction of AI’s promise, we propose that mobilizing AI’s voice in selection processes is most likely when the recommendations provided by AI can be used to arm top managers with information, rather than standing alone as a machine-based recommendation. AI’s range of computational capacity can then augment top manager intuition and broaden the contingencies and value judgments considered in strategy-making. Next, we describe the decision-making structures that are best suited to minimizing the previously identified artificial selection effects.

AI promises to become the voice of otherwise overlooked stakeholders of organizations (for example, customers, underrepresented stakeholders) thus, bringing to the surface or amplifying issues that were not acknowledged otherwise. We can anticipate instances when AI identifies the unarticulated voice of the customer, including tacit and unmet needs for new products and services in the long term, emerging technological needs and product trends, economic and market trends, demographic and behavioral trends, and negative externalities including stakeholders most likely to be impacted by climate-driven changes. As these trends converge, top managers may indeed see an emerging and previously overlooked set of stakeholders that may or may not be customers. Opportunities may arise to develop products of value to would-be customers in anticipation of future (but presently unarticulated) customer demand, before the competition. Alternatively, catastrophic events may be avoided if AI projects a likelihood that decisions further destabilize planetary conditions that impact overlooked stakeholders. Using AI’s voice to elevate overlooked stakeholders supports firms to establish a long-term perspective in strategy-making.

Voice clearing techniques to minimize artificial selection processes

In our example, AI voice allows organizations to account for diversity of thought in strategy-making and imagines future strategic paths in which this diversity is core to the organization's strategic agenda. AI's voice, therefore, becomes akin to heterogeneity in top management teams that proposes improved decision-making from a greater number of perspectives (Carpenter et al., 2004). In a very concrete way, elevating AI's voice would allow organizations to embrace a different voice, the AI voice of overlooked stakeholders, as the voice of heterogeneity that did not yet exist in that specific organization.

Yet, the challenge in realizing the promise of AI in this case of AI voice emerges because voice rights differ from decision rights; a key observation made by Turco (2016) in the context of social media platforms' use within organizations. Counterintuitively, we propose that the AI voice is most likely to generate different organizational strategy outcomes if the artificial selection processes that minimize AI voice are suppressed. We argue that voice clearing involves structuring decision processes around decentralized expertise (a tacit form of autocracy), rather than structuring decisions around democratic processes.

We propose that voting-based democratic processes present few opportunities for AI to truly alter the strategic agendas of organizations, due to voice weight. When a number of issues are presented by top managers (for example, each functional or business unit presents the issues believed to be most important), issues generated by AI become a function of the voting process. If AI is given an independent voice alongside other opinions at the table, the democratic process automatically assigns a small weight to the issues raised by AI. Put differently, because the AI becomes one voice among many, with equal weight of voting, voice rights may be present but decision rights are low. The determining factor will be the degree to which other top managers depart from the issues they independently raise to agree with AI.

Conversely, autocracy by expertise presents an opportunity for AI to be more influential in the decision-making process by arming non-experts with alternatives. When experts are relied upon to determine issues of importance, they hold disproportionate influence over strategic decisions (for example, the chief technology officer determines the key technological issues). When AI presents issues to all top managers, an opportunity opens up for non-expert top managers to be armed with AI-generated issues and information. AI-generated issues gain greater voice through these non-experts to challenge or reinforce expert opinions. AI's tendency to quantify issues can also realign goals in favor of non-experts (Mazmanian and Beckman, 2018), which may be used either to override expert perspectives that may be overly subjective, or to reinforce expert opinions in the instances where AI and experts agree. For this technique to be effective, ambiguity in AI's selection process should be low, and non-threatening to top manager identities (Morewedge, 2022). Under these conditions, AI voice may indeed surface new issues from overlooked stakeholders' perspectives that may be incorporated into strategic agendas in a way that qualitatively improves strategy-making.

As the example of AI voice illustrates, the possibilities for leveraging AI to overcome known issues in strategy-making may indeed result in different outcomes than those experienced otherwise. But the path to different outcomes is narrow, and contingent upon the existing organizational dynamics. As our AI voice example shows, these processes may play out in counterintuitive ways. At a practical level, however, voice clearing allows organizations to overcome the known issue of path dependent skillsets and downplayed AI voices, which we highlighted earlier as a key problem that AI is likely to face in artificial selection environments.

CONCLUSION

The computational might of AI tools and platforms is not under debate, and the potential impact on organizations is increasingly acknowledged. We contribute an understanding specifically on how AI may change the core strategy-making processes of issue detection and artificial selection within organizations, which is an emergent topic in current conversations amongst organizational leaders, thinkers, and scholars. To this end, we propose a way of thinking through and overcoming AI's limitations in both issue detection and artificial selection that can serve as a starting point for practical uses of AI in strategy-making.

We argue that a more fruitful avenue for AI use in strategy-making will come through shifting frames, from AI as a validating tool, to AI as imagining tool. The shift towards the latter can, as we present through examples, generate strategy-making outcomes that are different to those anticipated without AI's use. This shift requires a shift in perspective whereby the focus is directed towards what we refer to as futuring. We propose three techniques—crafting time horizons; timing and pacing; voice clearing—that we collectively label as “futuring techniques.” We argue that such techniques, amongst others, are needed to support leaders interested in AI's promise to transition intentionally away from established issues and paths that are likely to limit AI's promise in strategy-making. Our argument for a futuring lens articulates how each futuring technique may be leveraged to overcome the challenges that we identified in reaching the AI promise in strategy-making.

Our analysis, however, leaves a number of broader organizational topics for future discussion that are beyond the scope of this chapter. First, we wonder whether AI can challenge the “performance-oriented” culture to decrease the emphasis on short-term results among top managers. Our argument centers on the notion that solutions should not be the target of AI's role in strategy-making: rather, issues should. The focus on solutions and results is pervasive among top managers (Bansal and DesJardine, 2014), which is understandable given that top manager employment depends on producing results. The consequences of short-term thinking are equally problematic, and reflected in a recent turn toward discussion of corporate purpose (George et al., 2021). We believe that AI's role in strategy-making should support identifying issues that reveal and reinforce purpose in organizations. Butting up against this belief is

a broader social embeddedness of performance culture, and questioning whether AI reinforces or counters performance culture is an important question.

Relatedly, our discussion on AI voice and AI's augmentation of top manager skillsets in strategy-making relies on the social structures within organizations, and AI's inclusion in such social structures. We wonder whether non-experts can develop enough trust to support recommendations from so-called "black box" technologies. Voice clearing only amplifies AI voice if non-experts leverage AI to challenge the authority of experts. This largely boils down to understudied relationships between human–AI collaborations and the need to upskill top managers before trust can be developed. Trust between AI and humans is already being deeply studied, and AI's technological development will undoubtedly change in the coming months and years. The impact of these developments on human trust will certainly have implications for AI's use in strategy-making.

Despite these scope conditions, we believe that the futuring techniques outlined offer a way to establish AI's promises in improving strategy-making, while mitigating the problems that AI is likely to face. While strategy-making may very well look the same with or without AI for most organizations, those organizations that walk the narrowly lit path guided by futuring techniques hold the potential to broaden strategy-making, and AI's role within strategy-making, to hopefully peer into the longer term a little bit more than managers do today.

NOTE

- 1 This does not preclude the possibility of fewer issues being presented to top managers. We believe this to be reflected in superior issues being identified.

REFERENCES

- Amabile, T. (2019). GUIDEPOST: Creativity, Artificial Intelligence, and a World of Surprises. Guidepost Letter for Academy of Management Discoveries. *Academy of Management Discoveries*, 6(3), amd.2019.0075. <https://doi.org/10.5465/amd.2019.0075>.
- Anthony, C. (2021). When Knowledge Work and Analytical Technologies Collide: The Practices and Consequences of Black Boxing Algorithmic Technologies. *Administrative Science Quarterly*, 000183922110167. <https://doi.org/10.1177/00018392211016755>.
- Audia, P.G., and Greve, H.R. (2021). Organizational Learning from Performance Feedback: A Behavioral Perspective on Multiple Goals. In *Organizational Learning from Performance Feedback: A Behavioral Perspective on Multiple Goals* (Vol. 3859). <https://doi.org/10.1017/9781108344289>.
- Bansal, P., and DesJardine, M.R. (2014). Business Sustainability: It is About Time. *Strategic Organization*, 12(1), 70–78. <https://doi.org/10.1177/1476127013520265>.
- Bansal, P., Kim, A., and Wood, M.O. (2018). Hidden in Plain Sight: The Importance of Scale in Organizations' Attention to Issues. *Academy of Management Review*, 43(2), 217–241. <https://doi.org/10.5465/amr.2014.0238>.

- Baumann, O., Schmidt, J., and Stieglitz, N. (2019). Effective Search in Rugged Performance Landscapes: A Review and Outlook. *Journal of Management*, 45(1), 285–318. <https://doi.org/10.1177/0149206318808594>.
- Baumann, O., and Siggelkow, N. (2013). Dealing with Complexity: Integrated vs. Chunky Search Processes. *Organization Science*, 24(1), 116–132. <https://doi.org/10.1287/orsc.1110.0729>.
- Bundy, J., Shropshire, C., and Buchholtz, A.K. (2013). Strategic Cognition and Issue Salience: Toward an Explanation of Firm Responsiveness to Stakeholder Concerns. *Academy of Management Review*, 38(3), 352–376. <https://doi.org/10.5465/amr.2011.0179>.
- Carpenter, M.A., Geletkanycz, M.A., and Sanders, W.G. (2004). Upper Echelons Research Revisited: Antecedents, Elements, and Consequences of Top Management Team Composition. *Journal of Management*, 30(6), 749–778. <https://doi.org/10.1016/j.jm.2004.06.001>.
- Casciaro, T., and Piskorski, M.J. (2005). Power Imbalance, Mutual Dependence, and Constraint Absorption: A Closer Look at Resource Dependence Theory. *Administrative Science Quarterly*, 50(2), 167–199. <https://doi.org/10.2189/asqu.2005.50.2.167>.
- Cattani, G., Sands, D., Porac, J., and Greenberg, J. (2018). Competitive Sensemaking in Value Creation and Capture. *Strategy Science*, 3(4), 632–657. <https://doi.org/10.1287/stsc.2018.0069>.
- Chen, J., and Nadkarni, S. (2017). It's about Time! CEOs' Temporal Dispositions, Temporal Leadership, and Corporate Entrepreneurship. *Administrative Science Quarterly*, 62(1), 31–66. <https://doi.org/10.1177/0001839216663504>.
- Chen, M., Kaul, A., and Wu, X. (Brian) (2019). Adaptation Across Multiple Landscapes: Relatedness, Complexity, and the Long Run Effects of Coordination in Diversified Firms. *Strategic Management Journal*, smj.3060. <https://doi.org/10.1002/smj.3060>.
- Cho, T.S., and Hambrick, D.C. (2006). Attention as the Mediator Between Top Management Team Characteristics and Strategic Change: The Case of Airline Deregulation. *Organization Science*, 17(4), 453–469. <https://doi.org/10.1287/orsc.1060.0192>.
- Christensen, C.M., and Bower, J.L. (1996). Customer Power, Strategic Investment, and the Failure of Leading Firms. *Strategic Management Journal*, 17(3), 197–218. [https://doi.org/10.1002/\(SICI\)1097-0266\(199603\)17:3<197::AID-SMJ804>3.0.CO;2-U](https://doi.org/10.1002/(SICI)1097-0266(199603)17:3<197::AID-SMJ804>3.0.CO;2-U).
- Cohen, M.D., March, J.G., and Olsen, J.P. (1972). A Garbage Can Model of Organizational Choice. *Administrative Science Quarterly*, 17(1), 1. <https://doi.org/10.2307/2392088>.
- Crilly, D., and Sloan, P. (2012). Enterprise Logic: Explaining Corporate Attention to Stakeholders from the 'Inside-Out.' *Strategic Management Journal*, 33(10), 1174–1193. <https://doi.org/10.1002/smj.1964>.
- Cyert, R.M., and March, J.G. (1963). *A Behavioral Theory of the Firm*. Prentice-Hall.
- DesJardine, M.R., and Shi, W. (2020). CEO Temporal Focus and Behavioral Agency Theory: Evidence from Mergers and Acquisitions. *Academy of Management Journal*, amj.2018.1470. <https://doi.org/10.5465/amj.2018.1470>.
- Dutt, N., and Joseph, J. (2019). Regulatory Uncertainty, Corporate Structure, and Strategic Agendas: Evidence from the U.S. Renewable Electricity Industry. *Academy of Management Journal*, 62(3), 800–827. <https://doi.org/10.5465/amj.2016.0682>.
- Dutton, J.E., and Jackson, S.E. (1987). Categorizing Strategic Issues: Links to Organizational Action. *Academy of Management Review*, 12(1), 76–90. <https://doi.org/10.5465/amr.1987.4306483>.
- Dutton, J.E., Ashford, S.J., O'Neill, R.M., and Lawrence, K.A. (2001). Moves that Matter: Issue Selling and Organizational Change. *Academy of Management Journal*, 44(4), 716–736. <https://doi.org/10.5465/3069412>.
- Elsbach, K.D., Sutton, R.I., and Principe, K.E. (1998). Averting Expected Challenges Through Anticipatory Impression Management: A Study of Hospital Billing. *Organization Science*, 9(1), 68–86. <https://doi.org/10.1287/orsc.9.1.68>.

- Emirbayer, M., and Mische, A. (1998). What Is Agency? *American Journal of Sociology*, 103(4), 962–1023. <https://doi.org/10.1086/231294>.
- Esterman, M., and Rothlein, D. (2019). Models of Sustained Attention. *Current Opinion in Psychology*, 29, 174–180. <https://doi.org/10.1016/j.copsyc.2019.03.005>.
- Fu, R., Tang, Y., and Chen, G. (2019). Chief Sustainability Officers and Corporate Social (Ir) responsibility. *Strategic Management Journal*, 17(3), smj.3113. <https://doi.org/10.1002/smj.3113>.
- Gaba, V., Lee, S., Meyer-Doyle, P., and Zhao-Ding, A. (2022). Prior Experience of Managers and Maladaptive Responses to Performance Feedback: Evidence from Mutual Funds. *Organization Science*. <https://doi.org/10.1287/orsc.2022.1605>.
- Gavetti, G. (2012). Perspective: Toward a Behavioral Theory of Strategy. *Organization Science*, 23(1), 267–285. <https://doi.org/10.1287/orsc.1110.0644>.
- Gavetti, G., and Levinthal, D. (2000). Looking Forward and Looking Backward: Cognitive and Experiential Search. *Administrative Science Quarterly*, 45(1), 113. <https://doi.org/10.2307/2666981>.
- Gavetti, G., and Menon, A. (2016). Evolution Cum Agency: Toward a Model of Strategic Foresight. *Strategy Science*, 1(3), 207–233. <https://doi.org/10.1287/stsc.2016.0018>.
- Gavetti, G., Greve, H.R., Levinthal, D., and Ocasio, W. (2012). The Behavioral Theory of the Firm: Assessment and Prospects. *Academy of Management Annals*, 6(1), 1–40. <https://doi.org/10.1080/19416520.2012.656841>.
- George, G., Haas, M.R., McGahan, A.M., Schillebeeckx, S.J.D., and Tracey, P. (2021). Purpose in the For-Profit Firm: A Review and Framework for Management Research. *Journal of Management*, 49(6), 014920632110064. <https://doi.org/10.1177/01492063211006450>.
- Goyal, M., Knackstedt, T., Yan, S., and Hassanpour, S. (2020). Artificial Intelligence-based Image Classification Methods for Diagnosis of Skin Cancer: Challenges and Opportunities. *Computers in Biology and Medicine*, 127, 104065. <https://doi.org/10.1016/j.combiomed.2020.104065>.
- Greenwood, B.N., Agarwal, R., Agarwal, R., and Gopal, A. (2019). The Role of Individual and Organizational Expertise in the Adoption of New Practices. *Organization Science*, 30(1), 191–213. <https://doi.org/10.1287/orsc.2018.1246>.
- Henderson, R.M., and Clark, K.B. (1990). Architectural Innovation: The Reconfiguration of Existing Product Technologies and the Failure of Established Firms. *Administrative Science Quarterly*, 35(1), 9. <https://doi.org/10.2307/2393549>.
- Huising, R. (2014). The Erosion of Expert Control Through Censure Episodes. *Organization Science*, 25(6), 1633–1661. <https://doi.org/10.1287/orsc.2014.0902>.
- Jarrahi, M.H. (2018). Artificial Intelligence and the Future of Work: Human–AI Symbiosis in Organizational Decision-Making. *Business Horizons*, 61(4), 577–586. <https://doi.org/10.1016/j.bushor.2018.03.007>.
- Joseph, J., and Gaba, V. (2019). Organizational Structure, Information Processing, and Decision Making: A Retrospective and Roadmap for Research. *Academy of Management Annals*. <https://doi.org/10.5465/annals.2017.0103>.
- Joseph, J., and Ocasio, W. (2012). Architecture, Attention, and Adaptation in the Multibusiness Firm: General Electric from 1951 to 2001. *Strategic Management Journal*, 33(6), 633–660. <https://doi.org/10.1002/smj.1971>.
- Kahneman, D. (1973). *Attention and Effort*. Prentice-Hall.
- Kaplan, S., and Orlikowski, W.J. (2013). Temporal Work in Strategy Making. *Organization Science*, 24(4), 965–995. <https://doi.org/10.1287/orsc.1120.0792>.
- Krotov, V. (2019). Predicting the Future of Disruptive Technologies: The Method of Alternative Histories. *Business Horizons*, 62(6), 695–705. <https://doi.org/10.1016/j.bushor.2019.07.003>.

- Laverty, K.J. (1996). Economic “Short-Termism”: The Debate, the Unresolved Issues, and the Implications for Management Practice and Research. *Academy of Management Review*, 21(3), 825. <https://doi.org/10.2307/259003>.
- Leiblein, M.J., Reuer, J.J., and Zenger, T. (2018). Special Issue Introduction: Assessing Key Dimensions of Strategic Decisions. *Strategy Science*, 3(4), 555–557. <https://doi.org/10.1287/stsc.2018.0073>.
- Levine, S.S., Bernard, M., and Nagel, R. (2017). Strategic Intelligence: The Cognitive Capability to Anticipate Competitor Behavior. *Strategic Management Journal*, 38(12), 2390–2423. <https://doi.org/10.1002/smj.2660>.
- Levinthal, D. (1997). Adaptation on Rugged Landscapes. *Management Science*, 43(7), 934–950. <https://doi.org/10.1287/mnsc.43.7.934>.
- Levinthal, D. (2021). *Evolutionary Processes and Organizational Adaptation: A Mendelian Perspective on Strategic Management*. Oxford University Press.
- Levinthal, D., and March, J. (1993). The Myopia of Learning. *Strategic Management Journal*, 14(S2), 95–112. <https://doi.org/10.1002/smj.4250141009>.
- Levinthal, D., and Posen, H.E. (2007). Myopia of Selection: Does Organizational Adaptation Limit the Efficacy of Population Selection? *Administrative Science Quarterly*, 52(4), 586–620. <https://doi.org/10.2189/asqu.52.4.586>.
- Li, Q., Maggitti, P.G., Smith, K.G., Tesluk, P.E., and Katila, R. (2013). Top Management Attention to Innovation: The Role of Search Selection and Intensity in New Product Introductions. *Academy of Management Journal*, 56(3), 893–916. <https://doi.org/10.5465/amj.2010.0844>.
- March, J.G. (1991). Exploration and Exploitation in Organizational Learning. *Organization Science*, 2(1), 71–87. <https://doi.org/10.1287/orsc.2.1.71>.
- March, J.G. (2010). *The Ambiguities of Experience*. Cornell University Press.
- March, J.G., and Simon, H.A. (1958). *Organizations*. Wiley.
- Mazmanian, M., and Beckman, C.M. (2018). “Making” Your Numbers: Engendering Organizational Control Through a Ritual of Quantification. *Organization Science*, 29(3), 357–379. <https://doi.org/10.1287/orsc.2017.1185>.
- Melone, N.P. (1994). Reasoning in the Executive Suite: The Influence of Role/Experience-Based Expertise on Decision Processes of Corporate Executives. *Organization Science*, 5(3), 438–455. <https://doi.org/10.1287/orsc.5.3.438>.
- Michaelian, K., Klein, S.B., and Szpunar, K.K. (2016). *Seeing the Future: Theoretical Perspectives on Future-Oriented Mental Time Travel*. Oxford University Press.
- Morewedge, C.K. (2022). Preference for Human, Not Algorithm Aversion. *Trends in Cognitive Sciences*, 26(10), 824–826. <https://doi.org/10.1016/j.tics.2022.07.007>.
- Neumann, N., Tucker, C.E., Kaplan, L., Mislove, A., and Sapiezynski, P. (2022). *Data Deserts and Black Box Bias: The Impact of Socio-Economic Status on Consumer Profiling*. Working Paper.
- Ocasio, W. (1997). Towards an Attention Based View of the Firm. *Strategic Management Journal*, 18(S1), 187–206. [https://doi.org/10.1002/\(SICI\)1097-0266\(199707\)18:1+<187::AID-SMJ936>3.3.CO;2-B](https://doi.org/10.1002/(SICI)1097-0266(199707)18:1+<187::AID-SMJ936>3.3.CO;2-B).
- Ocasio, W., and Joseph, J. (2005). An Attention-Based Theory of Strategy Formulation: Linking Micro- and Macroperspectives in Strategy Processes. In Szulanski, G., Porac, J. and Doz, Y. (eds), *Strategy Process*, Advances in Strategic Management series (Vol. 22, pp. 39–61). Emerald Group Publishing. <http://www.emeraldinsight.com/doi/10.1108/MRR-09-2015-0216>.
- Ocasio, W., Laamanen, T., and Vaara, E. (2018). Communication and Attention Dynamics: An Attention-Based View of Strategic Change. *Strategic Management Journal*, 39(1), 155–167. <https://doi.org/10.1002/smj.2702>.
- Ocasio, W., Yakis-Douglas, B., Boynton, D., Laamanen, T., Rerup, C., Vaara, E., and Whittington, R. (2022). It’s a Different World: A Dialog on the Attention-Based View in

- a Post-Chandlerian World. *Journal of Management Inquiry*, 105649262211034. <https://doi.org/10.1177/10564926221103484>.
- Pietronudo, M.C., Croidieu, G., and Schiavone, F. (2022). A Solution Looking for Problems? A Systematic Literature Review of the Rationalizing Influence of Artificial Intelligence on Decision-Making in Innovation Management. *Technological Forecasting and Social Change*, 182, 121828. <https://doi.org/10.1016/j.techfore.2022.121828>.
- Raisch, S., and Krakowski, S. (2020). Artificial Intelligence and Management: The Automation-Augmentation Paradox. *Academy of Management Review*, 2018.0072. <https://doi.org/10.5465/2018.0072>.
- Simon, H.A. (1962). The Architecture of Complexity. *Proceedings of the American Philosophical Society*, 106(6), 467–482.
- Simon, H.A. (1973). Applying Information Technology to Organization Design. *Public Administration Review*, 33(3), 268. <https://doi.org/10.2307/974804>.
- Simon, H.A. (1978). Rationality as Process and as Product of Thought. *American Economic Review*, 68(2), 1–16.
- Starbuck, W.H., and Milliken, F.J. (1988). Executive Perceptual Filters: What They Notice and How They Make Sense. In Donald Hambrick (ed.), *The Executive Effect: Concepts and Methods for Studying Top Managers*. JAI Press, pp. 35–65.
- Suddendorf, T., and Corballis, M.C. (2007). The Evolution of Foresight: What is Mental Time Travel, and is it Unique to Humans? *Behavioral and Brain Sciences*, 30(3), 299–313. <https://doi.org/10.1017/S0140525X07001975>.
- Sund, K.J., Galavan, R.J., and Sigismund Huff, A. (2016). Introducing New Horizons in Managerial and Organizational Cognition. In K.J. Sund, R.J. Galavan, and A. Sigismund Huff (eds), *Uncertainty and Strategic Decision Making* (pp. xiii–xvii). Emerald Group Publishing. <https://doi.org/10.1108/S2397-52102016026>.
- Sydow, J., Schreyögg, G., and Koch, J. (2020). On the Theory of Organizational Path Dependence: Clarifications, Replies to Objections, and Extensions. *Academy of Management Review*, 45(4), 717–734. <https://doi.org/10.5465/amr.2020.0163>.
- Taleb, N.N. (2007). *The Black Swan: The Impact of the Highly Improbable* (1st edn). Random House.
- Turco, C. (2016). *The Conversational Firm: Rethinking Bureaucracy in the Age of Social Media*. Columbia University Press.
- Tversky, A., and Kahneman, D. (1974). Judgment Under Uncertainty: Heuristics and Biases. *Science*, 185(4157), 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>.
- Verhoeven, D., Bakker, J., and Veugelers, R. (2016). Measuring Technological Novelty with Patent-based Indicators. *Research Policy*, 45(3), 707–723. <https://doi.org/10.1016/j.respol.2015.11.010>.
- von Krogh, G. (2018). Artificial Intelligence in Organizations: New Opportunities for Phenomenon-Based Theorizing. *Academy of Management Discoveries*, 4(4), 404–409. <https://doi.org/10.5465/amd.2018.0084>.

19. Artificial intelligence as a mechanism of algorithmic isomorphism

Camille G. Endacott and Paul M. Leonardi

Technology companies are using advances in artificial intelligence (AI) to provide businesses and individuals with products that promise to increase workplace productivity and effectiveness (Trapp, 2019). AI, as a set of computational processes designed to mimic human intelligence to make complex decisions (Berente et al., 2021), leverages patterns found in data to learn and improve over time. Across a variety of professional domains, organizations are implementing technologies equipped with AI to accomplish work.

Emerging technologies equipped with AI are improving in their capabilities to autonomously make decisions on behalf of people and organizations (Leonardi and Neeley, 2022). AI technologies make decisions by drawing on patterns in existing data to make predictions about which actions are most likely to lead to given desirable outcomes across work settings. Though people vary in how they go about their work, AI technologies must appeal to a critical mass of users by making decisions based on patterns. The data that provide these patterns inform decisions made by AI such that actions are selected based on the probability that they will produce the desired outcome, as defined by developers of machine learning algorithms that make these choices, across use cases.

For AI companies that sell software as a service (SaaS), the business case for their tools lies in their ability to improve the computing abilities of their technologies over time. To make the case for the viability of their product, companies selling software powered by AI must demonstrate that their products will produce a return on investment for users that outweighs the cost of switching and learning services (Davenport et al., 2018). Many technology companies champion the power of big data and predictive analytics to distinguish the value of their AI solution. This approach allows them to appeal to clients at scale, and to increase the gross margins of their business (Casado and Bornstein, 2020). It is in companies' best interest to help the machine learning algorithms that power their AI tools improve, and to leverage vast amounts of aggregated data gathered across all users. Thus, the sophistication of these tools depends on their ability to account for as many use cases as possible.

Yet, that is not typically how AI is sold to many organizations. The popular rhetoric surrounding most AI tools—and indeed a key reason why they have captured attention so rapaciously—is that they promise to learn from specific user behaviors and become more customized to their users over time. Although such customization through machine learning on user-specific data is certainly possible if an individual or organization has a massive corpus of proprietary data on which to train the AI (at

the time of writing in April 2023, BloombergGPT represents one such example), most users of consumer-facing AI applications and most organizations simply do not have such massive quantities of data to assure that AI-based technologies are fine-tuned toward their particular needs. Instead, to help AI tools continue to learn, developers need to feed their machine learning algorithms with data from as many different users as possible; thus potentially undermining the very argument of customization that so many companies purport to offer.

Due to the fact that the many emerging consumer-facing AI tools learn by ingesting data not just from a single user but from a broad array of users, we argue that organizations' implementation of AI technologies can perpetuate a new mechanism of isomorphism, through which the work practices of organizations might become increasingly similar over time. We call this algorithmic isomorphism. Unlike mechanisms of isomorphism that depend on knowledgeable actors' responses to the externalities of the institutional field (Koçak and Özcan, 2013; Munir and Phillips, 2005), algorithmic isomorphism occurs as AI technologies implement patterns gleaned from aggregated data across time and space. As with other technologies, AI technologies are designed by developers to carry particular versions of reality and patterns of structuring action (Pollock et al., 2016). But unlike other digital technologies, AI technologies change in how they reach the outcomes specified by developers, based on what they learn by identifying patterns that emerge from across a wide userbase. AI technologies make an emergent form of isomorphism possible through aggregation of human action, as optimized for a determined set of constraints. AI technologies learn from aggregated data gathered across organizations, and make decisions based on what is likely to be statistically significant for the entire userbase. In so doing, they foster behaviors that may intentionally or unintentionally lead to increased homogeneity in work (Hancock et al., 2020; Pachidi et al., 2020).

In this chapter, we discuss how AI technologies can prompt algorithmic isomorphism in organizations. First, we discuss the nature of institutional isomorphism (DiMaggio and Powell, 1983) and how its mechanisms have been used to study organizational adoption of technologies. Second, we discuss why existing mechanisms of isomorphism are insufficient for studying how AI technologies shape organizations, and how AI technologies' capabilities to learn from aggregated data and act autonomously lead to an algorithmic mechanism of isomorphism when these technologies are embedded within organizations. Third, we offer possible outcomes of the algorithmic isomorphism that AI technologies make possible. Finally, we discuss the implications of our arguments, and directions for future research.

MECHANISMS OF INSTITUTIONAL ISOMORPHISM

Institutional isomorphism concerns the tendency for organizations' forms and practices to become more similar to one another over time. In their seminal work, DiMaggio and Powell (1983) define and describe institutional isomorphism as an explanation for how organizational homogeneity emerges. Drawing on Giddens

(1979), DiMaggio and Powell argue that organizations resemble one another over time because of the structuration of organizational fields. In Giddens's (1979) terms, organizations are structured by institutional-level structures such as professional standards, legislation, and norms. Organizations are enabled and constrained by these structures and draw on them to produce action. Because organizations exist in shared institutional fields, their structuring efforts in response to institutional structures tend to resemble one another, such that organizational change does not lead to increased differentiation but instead increased similarity over time.

DiMaggio and Powell (1983) identified three mechanisms through which isomorphic change occurs in organizations. The first mechanism is coercive isomorphism, in which organizations become more similar to one another because of political influence and pressure to conform to cultural expectations. The second mechanism is mimetic isomorphism, in which organizations imitate one another as a means to manage uncertainty. The third mechanism, normative isomorphism, occurs when organizations' actions are shaped by growing professionalization, especially professional norms that are legitimized in communities of practice. These mechanisms have generated empirical work on organizations in areas such as organizational entry into new markets (Koçak and Özcan, 2013), organizational strategy (Washington and Ventresca, 2004), corporate social responsibility (Lammers, 2003), and innovation (Tschang, 2007). Each of the three mechanisms identified by DiMaggio and Powell (1983) assumes that institutional isomorphism occurs as organizations interact with and respond to actors in their surrounding social system, such as government agencies, other organizations, or professional communities. While the actions of these entities may shape the practices of a focal organization (that is, creating sanctions, implementing new technologies, or adopting norms), they do so from outside the boundaries of the organization. We refer to these existing mechanisms of isomorphism as exogenous isomorphism because they are driven by organizations' active and reflexive responses to external entities.

Studies of technology and organizational change have drawn on institutional isomorphism to explain how organizations choose and implement the use of new technologies (Faik et al., 2020). Studies have shown how institutional isomorphism leads organizations to adopt sustainable technologies (Xu et al., 2022), and how organizations' dependence on information technologies mediates their likelihood of adopting them (Pal and Ojha, 2017). Other work has shown how companies strategically draw on discourse to shift institutional fields (Munir and Phillips, 2005), how organizational analysts attend to other organizations' decisions about technology implementation (Benner, 2010), and how professionals resist organizations' technological initiatives in the face of institutional isomorphism (Currie, 2012).

Existing studies of institutional isomorphism as it relates to technological change have primarily studied exogenous mechanisms of isomorphism. This work explains technological change largely as stemming from organizations' responses to their institutional field (Barley and Tolbert, 1997). Though studies have shown how the organization-specific changes in work practices that surround new technologies are enacted and negotiated among individual organizational members (Barley, 1986;

Leonardi, 2009a), when institutional isomorphism occurs it is assumed to be exogenous in nature.

To study organizational decision making about AI, exogenous mechanisms of isomorphism could be examined to understand how AI technologies are adopted and used within organizations. Such an approach could certainly make incremental contributions to our understanding of exogenous institutional isomorphism as it pertains to new technologies. Current research has shown, for example, that organizations are more likely to invest in AI when they feel pressured to satisfy customers and to become more competitive within their institutional field (Iwuanyanwu, 2021). Caplan and Boyd (2018) discuss the ways that organizations change their practices to adapt to other organizations' algorithms, in their analysis of the Facebook newsfeed algorithms. And in one of the most extensive treatments of an institutional perspective on AI technologies, Larsen (2021) distinguishes institutional and digital realms to argue that both institutions and digital infrastructure shape how organizations manage the uncertainty of adopting AI technologies.

Each of the studies described above contributes to our understanding of how mechanisms of exogenous isomorphism shape organizations' decisions about implementing AI technologies. In these studies, however, AI technologies are treated similarly to other technologies. AI technologies are treated as the outcome of isomorphism: that is, whether and how AI technologies are adopted and the practices through which people bring them into use (Orlikowski, 2000) are the outcomes of organizations' response to external forces.

However, such an approach is limited because AI technologies are not like all technologies. Unlike other iterations of digital technologies, AI technologies are capable of learning, and of making decisions without explicit human instruction in heterogeneous ways. In the next section, we describe the capabilities of AI technologies to act autonomously and learn. In combination, these capabilities allow AI technologies to shape organizational actions such that they are more similar to other organizations based on emergent patterns in data, a mechanism that we call algorithmic isomorphism. Relatedly, AI technologies are then not only the outcome of isomorphism but also actors that make isomorphism possible. We discuss why these capabilities require theorizing the use of AI technologies as the engine of organizational change towards homogeneity, not only as the outcome of organizations' isomorphic decision making.

ARTIFICIAL INTELLIGENCE AND ALGORITHMIC ISOMORPHISM

AI, most broadly, refers to computational processes designed to mimic human intelligence (Nilson, 2010). Generally, AI refers to complex predictive models that can outperform human decision making as opposed to rule-based computations (Berente et al., 2021). Technologies equipped with AI have two capabilities that are relevant to their capacity to shape organizational practice. The first capability is that AI technol-

ogies can learn, improving their decision making through computational processes without explicit human instruction. The second capability is that AI technologies can make decisions autonomously on behalf of people and organizations, as opposed to only facilitating human decision making. Below, we describe how these capabilities work together to make AI technologies' decision making a mechanism of algorithmic isomorphism.

Like human intelligence, AI relies on processes of learning to improve over time. Machine learning, or the processes through which AI technologies improve, occurs as AI technologies encounter data. AI technologies begin their learning by analyzing training data, a labeled set of data from which AI technologies can identify patterns between actions and particular outcomes that are either inductively identified or specified by a programmer (Nilson, 2010). AI technologies may not work effectively at first because they are still learning to account for a range of possible use cases (Shestakofsky and Kelkar, 2020), but over time AI technologies are designed to learn from the patterns of activity that users generate in deploying them. As actors naturalistically use AI technologies in a range of social settings, they generate a wider, more robust set of data that facilitates machine learning. While digital technologies without AI are improved through human action (that is, changing written code, fixing glitches; Neff and Stark, 2004), AI technologies can learn and refine their predictions from data themselves.

In addition to being able to learn from data, AI technologies are also capable of making decisions without explicit human instruction. While many AI technologies are designed to respond to human prompts (that is, natural language requests such as prompts for AI-generated imagery, or voice commands to virtual assistants), they can respond to these prompts without explicit instruction about how to do so. AI technologies vary in the extent to which they execute their decisions (for example, some AI technologies offer suggestions to users, whereas others take action without any human input; see Endacott and Leonardi, 2022), but share an ability to make decisions without being given exact parameters. Increasingly, AI technologies make decisions about specific tasks or actions on behalf of actors; for example, making decisions to direct customers toward different services on behalf of organizations, or making decisions about work practices on behalf of individuals in organizations. AI technologies' capability to make decisions on behalf of others allows them to meaningfully shape organizational practice through the actions that they generate and implement.

Taken together, AI technologies' capabilities to learn and autonomously make decisions form two logics that shape their actions: aggregation and optimization. By logics, we mean the "organizing principles" through which AI technologies operate (Thornton and Ocasio, 1999, p. 804) and the guiding values through which the work of these technologies is conducted. Aggregation refers to sophisticated AI technologies' reliance on wide swaths of data to identify patterns that are robust across use cases. Optimization refers to AI technologies' aim to predict which decisions have the highest probability of securing desired outcomes. As AI technologies' work is guided by the logics of aggregation and optimization, these technologies make deci-

sions based on what is most useful to a wide userbase, inclusive of a variety of social contexts. In other words, AI technologies make decisions that are suitable for a given set of constraints for the average user.

The challenge that the logics of aggregation and optimization pose to organizational practice is that most organizational decision making does not assume that decisions need to be appropriate across social settings. Most organizational decision making occurs based on what is appropriate for the specific social context of the organization, given its surrounding institutional field. Because different fields have different institutional logics, organizations vary in the “assumptions and values, usually implicit” through which organizational reality should be interpreted and behavioral decisions made (Thornton and Ocasio, 1999, p. 804). While the premise of institutional isomorphism assumes that organizations are affected by other organizations which share their normative expectations, it does not assume that organizations are affected by all organizations. Instead, organizations make decisions that are mechanisms of isomorphism based on the actions of other relevant organizations. For example, DiMaggio and Powell propose that organizations are more likely to model organizations on which they are dependent. Such a view assumes that organizational practices are selected based on the likelihood that they are the right courses of action for a particular time, place, and social group.

If AI technologies make decisions based on what is best for the entire userbase, and organizations make decisions based on what is best for their specific organization, then organizations’ decisions to implement AI technologies contrive a situation in which organizational practices may shift, if the actions that AI technologies generate are accepted (Endacott and Leonardi, 2022). As AI technologies make decisions based on aggregation and optimization, they implicitly select practices that combine a set of potentially conflicting institutional logics. The situated nature of organizational decision making is transformed into a process shaped by what is best for the average actor (that is, individual user or organization).

The selection of practices by AI technologies that are learning from aggregated data presents the possibility of practices becoming stretched over time and space into realms in which they did not originate, a phenomenon that Giddens (1984) called “time-space distanciation.” As AI technologies implement work practices that are probabilistically best for the average user, they replace practices that are specific to the organizational field, replacing domain-specific actions with homogenized ones. For example, Hancock et al. (2020) describe how written text can become homogenized when it is drafted using AI suggestions, using the example of Google’s predictive text function. As users draft emails, the function automatically suggests text based on data gathered from all users’ emails (if a user begins to write, “I hope,” suggested text of “this finds you well” will be displayed). It could be, however, that the user had intended to write “I hope you’re staying safe.” But the machine learning algorithms that power this tool must predict text based on what is most likely to be optimal for the greatest number of users. If more and more users accept the suggestions of the tool, a likely outcome is that writing in general will become more

homogenized, as people implement the patterned work practices gleaned from big data analysis that are suggested or implemented by AI technologies.

To the extent that these AI technologies can autonomously choose organizational practices, organizations may be made more similar to one another over time. AI technologies are capable of making decisions about organizational practices such as hiring (van den Broek et al., 2021), meeting with co-workers (Endacott and Leonardi, 2022), communicating with customers (Pachidi et al., 2020), and allocating organizational resources, that is, deploying staff (Waardenburg et al., 2022). In outsourcing any one of these organizational practices to AI, organizations may find that practices are transformed to be more like the average use case in the tool's training data (Glaser et al., 2021). Some technologies may draw on more field-specific data to make decisions (for example, predictive policing would learn from data gathered across cities related to policing), but other technologies draw on data from a variety of institutional fields (for example, scheduling tools learn from a variety of organizations involved in knowledge work). In both cases, however, AI technologies can enact mechanisms of isomorphism because the logics through which they make decisions require identifying practices likely to hold across institutional fields, or across organizations within an institutional field.

Because of the opacity of machine learning processes, many organizations may not even realize the criteria through which AI technologies are making decisions, nor the scope of the data on which they are trained. Burrell (2016) describes this type of opacity as emerging from the scale of data on which machine learning algorithms are trained. Because so many data points with "heterogenous properties" are analyzed in machine learning, the criteria on which decisions are made become increasingly complex (Burrell, 2016, p. 5). No single organization will be able to sufficiently understand the scope and nature of the data from which a particular AI technology is learning, nor would it be able to offer the same complexity of predictions. This opacity allows AI technologies to implement organizational practices gleaned from aggregated data without organizations' explicit awareness that they are doing so. This dynamic, in which AI technologies can choose organizational practices that regress toward the mean, based on emergent patterns rather than criteria pre-determined by developers, led us to call this a mechanism of algorithmic isomorphism.

To this point, we have discussed how AI technologies can bring about algorithmic isomorphism largely in the abstract. Next, we offer one illustrative example of how algorithmic isomorphism occurs, from our ongoing study of AI scheduling technologies.

AN ILLUSTRATIVE EXAMPLE OF ALGORITHMIC ISOMORPHISM: ARTIFICIALLY INTELLIGENT SCHEDULING TECHNOLOGIES

We researched how AI shaped organizational members' decisions in the domain of scheduling and by studying the development and use of an AI technology designed

to help users manage their calendar. Though our original research interests were in understanding how using AI technologies shaped how users represented themselves to others, we noticed that technologies also made users' work more similar to that of one another. We present a description of the case, how developers said that their AI technology enacted logics of aggregation and optimization, and users' descriptions of how their use of the AI technology minimized variation in their work.

Case: Time Wizards

We studied a company, which we call Time Wizards, which was working to develop a conversational agent that could schedule meetings on users' behalf. The Time Wizards tool was a conversational agent, which could be given the feminine name "Liz" or the masculine name "Leo." The tool was designed to recognize and generate scheduling requests in natural language. Much like an executive could copy an executive assistant into an email thread to arrange a meeting with an important investor, Liz and Leo were designed to autonomously schedule users after being prompted within email threads. Liz and Leo could then autonomously message meeting guests to inquire about their availability, identify times of shared availability, schedule the meeting, and place it on all parties' calendars. To accomplish this, Liz and Leo had to learn to recognize and generate a wide range of scheduling utterances. After several years of working with human trainers to teach Liz and Leo relevant expressions (that is, temporal expressions, meeting requests, and prompts) Time Wizards released its product to the public, which was available on a subscription basis. The tool was deployed by individual users with others within their organization or with people outside of it to schedule meetings in natural language. In adopting this AI technology, users ceded control over their calendar and how decisions about their schedule were communicated to others by Liz or Leo.

Time Wizards Developers on Aggregation and Optimization

We spoke to five developers at the company about the processes through which Liz and Leo were designed to learn. The developers affirmed that users wanted Liz and Leo to learn from within-case data, or to learn from their patterns alone. For example, one developer, Diego, the lead data scientist for the company, explained that if the tool could "pre-populate preferences for the user based on their calendar, like based on their habits, that would delight them."

Despite user interest in a tool that would learn their unique patterns, it was clear from interviews with developers that the tool was designed to learn from all aggregated data at the level of the userbase. The tool did not learn from unique users; rather, its decision making was continuously being updated based on the aggregated data that users generated in their deployment of the technology. Developers described how their tool was designed to learn via a logic of aggregation. For example, co-founder Mikkel explained how the value of the tool lay in the vast degree of data from which it would learn. He explained, "On an individual basis, you can't really do

any optimization in your own inbox. It's too sparse of a dataset, you don't have the time, you don't really think about it. But we [as an AI company] can start to really think about this." While at first human coders supplied the input of data from the machines to learn, after releasing the software into the market, the company relied on users' data to continuously update its models. Michelle, who oversaw human training at Time Wizards, explained that all users' data will help Liz and Leo to improve, and increase the probability that they will make more appropriate decisions: "Liz and Leo are constantly improving. Any input will feed the model." Diego explained that the tool needs to learn from many, many data points so that it can accurately guess at what a user wants and what it should do next. He said that users helped Liz and Leo to learn what actions were most likely to lead to a desirable outcome: "What I can say is after looking at a million meetings, I can say at this juncture of the meeting, people typically want to say one of these twenty things." These quotations indicate that the complexity of decisions made by the AI agent required learning from aggregated data collected across the userbase, suggesting that our proposed logic of aggregation shapes the work of this AI technology.

The aggregated data from which Liz and Leo learned allowed the tool to make decisions using the logic of optimization. While Liz and Leo had initially learned from paid human coders, their underlying algorithms were constantly being refined through data generated by users. As Diego explained, "Liz is trained in aggregate and the dialogue is optimized in aggregate." Diego's comments suggest that Liz and Leo could generate appropriate responses in natural language because they had learned from aggregated data and continued to be optimized based on aggregated data. Mikkel explained that Liz and Leo then learn how to make decisions about work practices and how optimal actions begin to emerge from the data. He explained how Liz and Leo learn how to optimally negotiate the time and duration of meetings:

You [the developer] can certainly chart it in such a way that you can start to see which particular dialogue paths, are more likely to yield success because if I can see that this certain path that is more likely to yield success then I can start to direct the dialogue, like any good negotiator, down a path where I am now more confident that you and me will come to a positive outcome.

Mikkel and Diego's comments suggest that the logic of optimization did shape how their AI technology was designed to make decisions about the practices that should be implemented on users' behalf.

Time Wizards Users on Isomorphism in Their Work

To learn how AI technologies shaped organizational practices, we also interviewed users of Time Wizards technology. We spoke with 15 users about their use of the Time Wizards AI agent, and associated consequences. We contacted users again six to nine months after their first interview to ask them to participate in a second interview so that we could better understand how their use of the technology shaped

work; 13 of the original 15 users participated in a second interview. Some users worked in larger organizations (that is, as technology officers, researchers, or sales professionals), while most users worked in small businesses (that is, small consulting firms, start-ups, financial advising office). Some users worked with organizations as contractors, that is, business consultants.

Users described how once they started using Liz and Leo, they noticed that the tool scheduled them differently than they would do themselves. Often, users described how the tool scheduled them in ways that would be appropriate for someone in sales or recruiting, who had a high volume of meetings that were relatively similar in time and duration. For example, users Joe and Bob both described the perfect user for the tool as “a recruiter or kind of a salesperson” (Joe). Joe, who ran a small consulting company for the legal field, explained that the tool is designed to help someone with a high “volume and uniformity” of meetings. Some users noticed that their meeting needs were quite varied. As Bob, a business consultant, explained:

I’m very rarely giving somebody an audience. Not because I’m a jerk, just that my role doesn’t call for it right now. Recruiters and salespeople, meetings where someone’s doing me a favor—I don’t do a lot of that anymore. It’s more often, I’m putting together strategic partnerships where it’s me and a couple of other people at my level trying to figure out how to make something happen.

Another user, Richard, described that based on how the tool made decisions, the perfect user would be “a salesperson or a dev [development] type person whose primary thing is they want to set up meetings.” These quotations show that users understood the tool as designed to optimize the number of meetings in which they participated in a given week.

As users outsourced their scheduling practices to the AI agent, they noticed that their schedules became more similar to that of a recruiter or salesperson. For example, Bradley, who worked for a start-up, explained that his work had typically involved some meetings, but also more independent creative work. He explained that he perceived one effect of using Leo for scheduling is that he has “more external meetings, more meetings with new people.” Many users shared Bradley’s observation that they had many more meetings since using Liz and Leo. For example, Joe explained that the tool’s “default” way of organizing meetings is to schedule “ASAP.”

Users had the option to reschedule or delete meetings scheduled by Liz or Leo (and doing so would, according to developers, help to train the model). However, many users choose not to do so, since the request had already been sent to their meeting partner. Without active intervention, Liz and Leo arranged users’ work based on what the tool was trained to do: negotiate meetings as efficiently as possible, as optimized for the greatest amount of successfully scheduled meetings. For users whose work deviated from the norms of a salesperson or recruiter, they experienced an influx of meetings onto their calendar. Their work began to resemble that of a salesperson, who made up a significant portion of the Time Wizards userbase. One user, Benjamin, who oversaw company partnerships for the government, pointed out that

this transformation of work had taken its toll on him, even though meetings were vital to his work. He said, “The robot doesn’t know that I’m hungry and need a break ... I get tired of the sound of my own voice.” He said that he had meetings scheduled with people who, if he had to schedule the meetings himself, he would not have put in the requisite effort to meet. But since Liz and Leo scheduled meetings with little necessary input from him, he found himself in more meetings. Using Liz and Leo has changed his “understanding of getting work done and how much there is to do and how if I don’t [schedule my work blocks], my schedule will be full, and then I won’t get any of that time for my work.” As Benjamin’s comment shows, users perceived that the practices through which their work was organized and accomplished were changed as they entrusted an AI agent to manage their calendars. The practices were arranged according to the logics by which the AI agent was designed to act, as enabled by the aggregated data from users across many different work contexts. This case suggests that AI technologies can spread practices across work contexts, removing sources of variation in how people choose to organize their work by masking differences in their work practices.

Our illustrative case occurred at the individual level, with individuals’ work practices being altered as an AI technology made choices on their behalf. We argue that such changes can scale up toward isomorphic changes in organizations, just as traditional mechanisms of institutional isomorphism occur as organizational leaders make strategic decisions. Even if people anthropomorphize organizations as agentic entities (Ashforth et al., 2020), organizational decision making is enacted by people. Thus, we argue that isomorphism is possible when organizational members allow AI technologies that implement patterns from across the userbase to make decisions about organizational practices.

Possible Outcomes of Algorithmic Isomorphism

AI technologies’ algorithmic shaping of organizational practices to be more similar to the average case yields several potential significant outcomes for organizations. Certainly, as recent studies have shown, the implementation of patterns located in training data can perpetuate bias, for example in client services in policing (Brayne and Christin, 2020) and medicine (Lebovitz et al., 2022). But a more subtle effect of the homogenization of work may also be a reduction in the requisite variety of inputs, including knowledge and practices, needed to facilitate organizational creativity and innovation (Weick, 1979). Organizations should consider whether, when, and how they should retain variation as work processes are increasingly mediated and organized by AI. Below, we describe how the mechanism of algorithmic isomorphism could lead to a spread and reinforcement of work-related bias and lower requisite variety for innovation, but also a greater awareness of organizational practices.

Spread and reinforcement of work-related bias

One potential outcome of algorithmic isomorphism is the spread and reinforcement of work-related bias. Ongoing conversations about the ethics of AI have focused on

the ways in which bias related to gender, race, and class can become perpetuated by AI technologies that learn from biased data (Brayne and Christin, 2020; Lebovitz et al., 2022). But AI technologies can also be biased toward particular logics of work. This bias can arise through aggregation. If enough of the userbase works or makes decisions in a similar way, the AI technology will learn to imitate those patterns (that is, the “recruiter” or “salesperson” in our example above). Consequently, AI technologies can have a homogenizing effect through an aggregation of individual-level changes in practices. The bias can also stem from optimization, as the outcome criteria for which the tool is trained to pursue can reflect bias about the ideal mode of working or organizing (that is, Time Wizards developers optimizing their agent to make as many meetings happen as possible). As AI technologies stretch practices across time and space, they may reinforce the dominant logics through which work is organized. In doing so, AI technologies can reconfigure how people enact their work roles (Pachidi et al., 2020).

Lower requisite variety for innovation

Across theoretical perspectives, including evolutionary models of organizational routines (Weick, 1979), behavioral theories of the firm (March, 1991; Simon, 1997), and network theories of innovation (Leonardi and Bailey, 2017; Uzzi and Spiro, 2005), a shared assumption is that organizations need to experience a certain degree of variety in their organizational practices in order to learn and innovate. If organizations increasingly use AI technologies to make decisions about their work, it is likely that there will be lower variety of practices both within and across organizations. Within organizations, AI technologies may implement decisions that are less reactive to specific situational contexts, because the technologies are optimized for fixed outcomes. Across organizations, increasing use of AI technologies may reduce the variety of organizational practices, because these practices serve the needs of the crowd rather than the preferences of the individual user. Such homogenization may ensure that the most data-supported best practices (as determined by the userbase) are implemented in organizations, but it may also reduce the requisite variety of actions that organizations can select and retain (Weick, 1979).

More awareness of organizational practices

One potential unintended but useful consequence of organizations deploying AI technologies is that doing so may help organizations to interrogate their practices. As our illustrative example of AI scheduling shows, users became more aware of how their own work was organized when they implemented an AI agent to schedule on their behalf. They noticed ways in which the tool shaped their work to pursue different logics than they themselves used when organizing their workday. As AI technologies introduce new organizational practices into an organization’s landscape, this may serve as an occasion for organizations to better understand taken-for-granted structures. For example, as people in organizations observe differences between decisions generated by machine learning and unassisted decision making, they can interrogate existing knowledge boundaries (van den Broek et al., 2021; Waardenburg et al.,

2022). When organizations are better positioned to understand and articulate their practices, they may be more apt to interrogate them, and change them if necessary.

IMPLICATIONS

In this chapter, we have theorized the implementation of AI technologies that can make decisions about organizational practices on behalf of organizations and individuals as a mechanism of algorithmic isomorphism. Rather than exerting organizational change towards homogeneity by eliciting a conscious organizational response to outside pressures, we argue that AI technologies push organizations toward homogeneity by implementing organizational practices that are selected via aggregation and optimization. Because the sources of data and the optimization criteria for decisions are usually opaque, AI technologies can implement practices drawn from other organizational contexts without organizational members' awareness. Below, we discuss the implications of this argument for theories of technology and institutional isomorphism, the methodological implications for studying intelligent technologies, and practical implications for organizational leaders.

Theoretically, our argument offers a new way for technologies to be implicated in isomorphic change: as implementors of new organizational practices, rather than as the content of that change. Our argument moves beyond technological adoption as an outcome of isomorphism to theorize how intelligent technologies that can make decisions on behalf of organizations introduce new practices into an organizational setting. Such an approach highlights how existing theoretical perspectives such as institutional isomorphism can be reconfigured to account for the unique characteristics of AI (the ability to learn from aggregated data, and to probabilistically make decisions based on that learning). Theorizing AI technologies as actors within institutional fields has the potential to yield a much more significant intellectual contribution than continuing to examine how organizations make decisions about whether to adopt AI, or by not sufficiently accounting for the dynamic nature of homogeneity that AI technologies can produce.

Practically, understanding the implementation and use of AI technologies as a mechanism of algorithmic isomorphism may help organizations be more attuned to intended and unintended consequences of novel technologies. Without greater awareness, organizational leadership may assume that AI technologies are designed to learn from their specific organizations, which may make them less likely to question decisions made by AI. By highlighting the logics through which AI technologies make decisions, our argument may help organizations to become more aware of ways in which AI technologies make decisions that depart from their desired institutional logics. In such instances, human input might be especially useful in overriding or amending decisions made by AI (Schestakofsky and Kelkar, 2020). When organizations are mindful of AI technologies' capacity to transform their practices, AI technologies may contrive occasions for organizations to notice areas in which

practices diverge from existing routines, allowing them to retain and enact the most useful changes.

Our conceptual argument also has methodological implications. Future research on AI technologies requires deep understanding of processes on both sides of the implementation line: that is, of both development, including machine learning, the outcomes for which algorithms are optimized; and the nature of training data, and use, including changes that AI technologies engender in situated organizational contexts (Bailey and Barley, 2020; Leonardi, 2009b). One analytical strategy could be to design studies to elicit the underlying institutional logics of both the AI technologies and the organizational contexts where they are embedded. A researcher could select several cases to see whether, how, and when an organization's existing institutional logics are reconfigured by implementation of AI technologies. That research design might help to illuminate use cases in which organizations would be especially susceptible to algorithmic isomorphism and what the implications of reconfigurations toward homogeneity could be.

Our argument has several limitations. We note that our argument is primarily premised on people's acceptance of the actions that AI technologies generate, and that people may respond to AI technologies in many ways. However, most AI technologies can learn even from users' rejection of their actions, which means that AI technologies may still eventually lead to isomorphism as they adapt and improve over time. A second possible critique of our discussion here is that our algorithmic mechanism of isomorphism may be capturing the diffusion of forms and practices (that is, "mere spread"; Greenwood and Meyer, 2008, p. 262), rather than the institutionalized practices that DiMaggio and Powell described. Future research should assess the extent to which AI technologies change organizational practices to adhere to different institutional logics, by studying AI technologies as they are implemented and used in organizations.

CONCLUSION

In this chapter, we have theorized a new mechanism of isomorphism that is brought about by AI technologies that operate through logics of aggregation and optimization: algorithmic isomorphism. Understanding how AI technologies transform organizational practices toward greater homogeneity is important to understanding how organizations may adapt and change within an algorithmically mediated institutional field, as well as exploring interventions required for maintaining the variety of actions needed for organizational learning.

REFERENCES

- Ashforth, B.E., Schinoff, B.S., and Brickson, S.L. (2020). "My company is friendly", "Mine's a rebel": Anthropomorphism and shifting organizational identity from "What"

- to “Who.” *Academy of Management Review*, 45, 29–57. <https://doi.org/10.5465/amr.2016.0496>.
- Bailey, D.E., and Barley, S.R. (2020). Beyond design and use: How scholars should study intelligent technologies. *Information and Organization*, 30. <https://doi.org/10.1016/j.infoandorg.2019.100286>.
- Barley, S.R. (1986). Technology as an occasion for structuring: Evidence from observations of CT scanners and the social order of radiology departments. *Administrative Science Quarterly*, 78–108. <https://www.jstor.com/stable/2392767>.
- Barley, S.R., and Tolbert, P.S. (1997). Institutionalization and structuration: Studying the links between action and institution. *Organization Studies*, 18, 93–117. <https://doi.org/10.1177/017084069701800106>.
- Benner, M.J. (2010). Securities analysts and incumbent response to radical technological change: Evidence from digital photography and internet telephony. *Organization Science*, 21, 42–62. <https://doi.org/10.1287/orsc.1080.0395>.
- Berente, N., Gu, B., Recker, J., and Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, 45, 1433–1450. <https://doi.org/10.25300/MISQ/2021/16274>.
- Brayne, S., and Christin, A. (2020). Technologies of crime prediction: The reception of algorithms in policing and criminal courts. *Social Problems*, 1–17. <https://doi.org/10.1093/socpro/spaa004>.
- Burrell, J. (2016). How the machine “thinks”: Understanding opacity in machine learning algorithms. *Big Data and Society*, 3. <https://doi.org/10.1177/2053951715622512>.
- Caplan, R., and Boyd, D. (2018). Isomorphism through algorithms: Institutional dependences in the case of Facebook. *Big Data and Society*, 5. <https://doi.org/10.1177/205391718757253>.
- Casado, M., and Bornstein, M. (2020, February 22). The new business of AI (and how it’s different than traditional software). *The Machine (Venture Beat)*. venturebeat.com.
- Currie, W.L. (2012). Institutional isomorphism and change: The national programme for IT—10 years on. *Journal of Information Technology*, 27, 236–248. <https://doi.org/10.1057/jit.2012.18>.
- Davenport, T.H., Libert, B., and Buck, M. (2018). How B2B software vendors can help their customers benchmark. *Harvard Business Review*. <https://hbr.org>.
- DiMaggio, P.J., and Powell, W.W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*, 48, 147–160. <https://www.jstor.org/stable/2095101>.
- Endacott, C.G., and Leonardi, P.M. (2022). Artificial intelligence and impression management: Consequences of autonomous conversational agents communicating on one’s behalf. *Human Communication Research*, 48, 462–490. <https://doi.org/10.1093/hcr/hqac009>.
- Faik, I., Barrett, M., and Oborn, E. (2020). How information technology matters in societal change: An affordance-based institutional logics perspective. *MIS Quarterly*, 44, 1359–1390. <https://doi.org/10.25300/MISQ/2020/14193>.
- Giddens A. (1979). *Central problems in social theory: Action, structure, and contradiction in social analysis*. Palgrave.
- Giddens, A. (1984). *The constitution of society*. UC Press.
- Glaser, V.L., Pollock, N., and D’Adderio, L. (2021). The biography of an algorithm: Performing algorithmic technologies in organizations. *Organization Theory*, 2, 1–27. <https://doi.org/10.1177/26317877211004609>.
- Greenwood, R., and Meyer, R.E. (2008). Influencing ideas: A celebration of DiMaggio and Powell (1983). *Journal of Management Inquiry*, 17, 258–264. <https://doi.org/10.1177/1056492060836693>.
- Hancock, J.T., Naaman, M., and Levy, K. (2020). AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, 25, 89–100. <https://doi.org/10.1093/jcmc/zmz022>.

- Iwuanyanwu, C.C. (2021). Determinants and impact of artificial intelligence on organizational competitiveness: A study of listed American companies. *Journal of Service Science and Management*, 14, 502–529. <https://www.scirp.org/journal/jssm>.
- Koçak, O., and Özcan, S. (2013). How does rivals' presence affect firms' decision to enter new markets? Economic and sociological explanations. *Management Science*, 59, 2586–2603. <https://doi.org/10.1287/mnsc.2013/1723>.
- Lammers, J.C. (2003). An institutional perspective on communicating corporate responsibility. *Management Communication Quarterly*, 16, 618–624. <https://doi.org/10.1177/089331890225064>.
- Larsen, B.C. (2021). A framework for understanding AI-induced field change: How AI technologies are legitimized and institutionalized. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. <https://doi.org/10.1145/3461702.3462591>.
- Lebovitz, S., Lifshitz-Assah, H., and Levina, N. (2022). To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization Science*, 33, 126–148. <https://doi.org/10.1287/orsc.2021.1549>.
- Leonardi, P.M. (2009a). Why do people reject new technologies and stymie organizational changes of which they are in favor? Exploring misalignments between social interactions and materiality. *Human Communication Research*, 35, 407–441. <https://doi.org/10.1111/j.1468-2958.2009.01357.x>.
- Leonardi, P.M. (2009b). Crossing the implementation line: The mutual constitution of technology and organizing across development and use activities. *Communication Theory*, 19, 278–310. <https://doi.org/10.1111/j.1468-2885.2009.01344.x>.
- Leonardi, P.M., and Bailey, D.E. (2017). Recognizing and selling good ideas: Network articulation and the making of an offshore innovation hub. *Academy of Management Discoveries*, 3, 116–144. <https://doi.org/10.5465/amd.2015/0515>.
- Leonardi, P., and Neeley, T. (2022). *The digital mindset: What it really takes to thrive in the age of data, algorithms, and AI*. Harvard Business Review Press.
- March, J.G. (1991). Exploration and exploitation in organizational learning. *Organization Science*, 2, 71–87. <https://doi.org/10.1287/orsc.2.1.71>.
- Munir, K.A., and Phillips, N. (2005). The birth of the “Kodak moment”: Institutional entrepreneurship and the adoption of new technologies. *Organization Studies*, 26, 665–1687. <https://doi.org/10.1177/0170840605056395>.
- Neff, G., and Stark, D. (2004). Permanently beta: Responsive organization in the Internet era. In P.N. Howard and S. Jones (eds), *Society online: The internet in context*. SAGE.
- Nilson, N.J. (2010). *The quest for artificial intelligence: A history of ideas and achievements*. Cambridge University Press.
- Orlikowski, W.J. (2000). Using technology and constituting structures: A practice lens for studying technology in organizations. *Organization Science*, 11, 404–428. <https://www.jstor.org/stable/2640412>.
- Pachidi, S., Berends, H., Faraj, S., and Huysman, M. (2020). Make way for the algorithms: Symbolic actions and change in a regime of knowing. *Organization Science*, 32, 18–41. <https://doi.org/10.1287/orsc.2020.1377>.
- Pal, A., and Ojha, A.K. (2017). Institutional isomorphism due to the influence of information systems and its strategic position. In *Proceedings of the 2017 ACM SIGMIS Conference on Computers and People Research* (pp. 147–154).
- Pollock, N., Williams, R.A., and D'Adderio, L. (2016). Generification as a strategy: How software producers configure products, manage user communities and segment markets. In S. Hyysalo, T.E. Jensen, and N. Oudshoorn (eds), *The new production of users: Changing innovation collectives and involvement strategies* (pp. 160–189). Routledge.
- Shestakofsky, B., and Kelkar, S. (2020). Making platforms work: Relationship labor and the management of publics. *Theory and Society*, 49, 863–896. <https://doi.org/10.1007/s11186-020-09407-z>.

- Simon, H.A. (1997). *Administrative behavior: A study of decision-making processes in administrative organizations* (4th edn). Free Press.
- Thornton, P.H., and Ocasio, W. (1999). Institutional logics and the historical contingency of power in organizations: Executive succession in the higher education publishing industry, 1958–1990. *American Journal of Sociology*, 105, 801–843.
- Trapp, R. (2019, May 30). AI could be better for the workplace than we think, but we still need to be careful. *Forbes*. <https://forbes.com>.
- Tschang, F.T. (2007). Balancing the tensions between rationalization and creativity in the video games industry. *Organization Science*, 18, 989–1005. <https://doi.org/10.1287/orsc.1070.0299>.
- Uzzi, B., and Spiro, J. (2005). Collaboration and creativity: The small world problem. *American Journal of Sociology*, 111, 447–504.
- van den Broek, E., Sergeeva, A., and Huysman, M. (2021). When the machine meets the expert: An ethnography of developing AI for hiring. *MIS Quarterly*, 45, 1557–1580. <https://doi.org/10.25300/MISQ/2021/16559>.
- Waardenburg, L., Huysman, M., and Sergeeva, A.V. (2022). In the land of the blind, the one-eyed man is king: Knowledge brokerage in the age of learning algorithms. *Organization Science*, 33, 59–82. <https://doi.org/10.1287/orsc.2021.1544>.
- Washington, M., and Ventresca, M.J. (2004). How organizations change: The role of institutional support mechanisms in the incorporation of higher education visibility strategies. *Organization Science*, 15, 1369–1683. <https://doi.org/10.1287/orsc.1110.0652>.
- Weick, K.E. (1979). *The social psychology of organizing*. McGraw-Hill.
- Xu, N., Fan, X., and Hu, R. (2022). Adoption of green industrial internet of things to improve organizational performance: The role of institutional isomorphism and green innovation practices. *Frontiers in Psychology*, 13, 917533.

20. Ethical implications of AI use in practice for decision-making

Jingyao (Lydia) Li, Yulia Litvinova, Marco Marabelli, and Sue Newell

INTRODUCTION

In recent times, artificial intelligence (AI) has become the new hype. Despite disinvestments in emerging technologies such as blockchain, and massive layoffs due to the post-pandemic “back to normal” (Kessler, 2023), high-tech companies are intensively investing in AI (Rotman, 2023). One of the main reasons for this AI hype lies in the remarkable progress that generative AI has made since 2021 (Dwivedi et al., 2023). Yet, with technology progress, ethical issues surface (Bender et al., 2021). In this chapter, we focus specifically on the ethical implications of AI systems, based on algorithms embedded in everyday devices. We suggest that, to build cumulative knowledge on AI, we need to recognize that most ethical issues of AI emerge with use, and it is nearly impossible to prevent these issues, systematically, when the AI is first designed and implemented.

We see the AI lifecycle as a nonlinear, messy unfolding of practices where AI is constantly tweaked (design, implementation) as societal consequences (use) surface (Cook and Brown, 1999; Newell et al., 2009). We argue that AI design, implementation, and use in practice are mutually influenced and constantly mangled. To study AI with this approach, we draw on Marabelli et al.’s (2021a) framework concerning the lifecycle of algorithmic decision-making systems (ADMS) (design, implementation, and use in practice). This approach offers an alternative way to consider the ethical issues associated with AI; aside from concerning the ethical design of AI (d’Aquin et al., 2018; Martin, 2019). We thereby focus on AI use in practice, but we do so by taking into consideration how AI is designed and implemented.

Here we question assumptions suggesting that AI design and implementation can prevent all negative consequences of AI in practice, as technology affordances are unpredictable (Costa, 2018; Leonardi, 2011). “Good” design is certainly important (Finocchiaro et al., 2021) and can help to prevent negative outcomes, such as the perpetuation of human biases, discrimination, and so on. Research, for example, has largely demonstrated that the lack of team diversity during development generates AI largely biased against those not involved in design (Noble, 2018). For instance, in the 1970s, car seatbelts were designed without thinking of how they could protect pregnant women, in part because no pregnant women were involved with the design teams (Perez, 2019).

Implementation too deserves attention. In fact, AI systems generally have a very narrow focus (based on the data they were trained with), and work as expected (most of the time) only in the contexts they were designed for (Luca et al., 2016). In our opinion, it is only by observing and assessing their use in practice that we can identify and seek solutions for the most relevant and problematic issues. We claim that both design and implementation issues are often visible and can therefore be addressed by good practices, ethical codes, and laws/regulation. The use in practice of AI remains more nebulous, for two main reasons.

First, not all ethical issues are clearly evident from the onset of AI use. They may be invisible to the designers (in good faith) and implementers/companies (Markus et al., 2019). For example, it took over a year for community officials to realize that Zillow's AI system, aimed at maximizing profits for sellers, was actually penalizing mostly Black and Latino homeowners. The algorithm was fed with, among other variables, zip code data concerning local crimes in areas where homes were on the market. But zip codes ended up acting as a proxy to discriminate (Ponsford, 2022). In this case the Zillow AI system was not designed to be racist, but operated otherwise.

Second, and related, when AI systems embed ethical issues such as biases, these issues generally affect minorities, underrepresented populations, and more generally powerless people who would need to be protected by laws and regulations because they can hardly speak up for themselves (Crawford, 2021; O'Neil, 2016). The reason behind this is that companies tend to benefit from "efficient" AI, that generates revenues even if this is at the expense of some. This dystopian view of consequentialism (an ethical construct suggesting that we should assess the ethicality of acts based on their outcomes) needs to be prevented, addressed, and regulated by laws. Yet, laws and regulations are notoriously late in addressing ethical issues associated with technology (mis)use: first issues emerge, and only after (some) people are penalized by AI, lawmakers attempt to fill gaps, loopholes, and so on. Companies may try to benefit from this institutional inertia. Just to give an example, former Google chief executive officer Eric Schmidt recently suggested that we should wait until something bad happens and then we figure out how to regulate it, otherwise, "you're going to slow everybody down" (Kaye, 2022).

In summary, the difficulties of spotting potential AI issues at the design and implementation stages, along with lawmakers' inability to produce regulations protecting minorities, call for considering specific characteristics of AI which typically emerge as concerns during their use in practice. We argue that it is only by focusing on these characteristics—and how they can generate ethical issues—that we can be more proactive in mitigating negative outcomes of AI use.

AI CHARACTERISTICS

In this section, we lay out six AI characteristics that emerge with AI long-term use. For each characteristic, we provide one meaningful example (vignette) which is illustrative of an ongoing issue. While these AI characteristics are key for making

our point on the relevance of scrutinizing AI use, we do not aim at providing an exhaustive and comprehensive taxonomy of the phenomenon. We instead use them to illustrate the importance of ongoing monitoring of AI use in practice. These characteristics concern training data, power asymmetries, automation, transparency, accountability, and the environment.

AI and Training Data

Issue

In theory, AI can (or should) be less biased than humans, because algorithms do not have feelings—and so would not favor job candidates, loan applicants, and so on—and process data they are trained with in an objective way. However, the fact that AI is fed with historical data has been shown to perpetuate longstanding discriminations (Browning and Arrigo, 2021). These issues can be partially dealt with in design and implementation. Designers, for example, have the important task of generating algorithms that mitigate or even “correct” past biases; while implementers need to make sure that systems using historical data (most AI systems) are rolled out in appropriate settings. For instance, an AI system relying on historical data on clinical trials conducted with White people—something very common in the past (Obermeyer et al., 2019)—should be revised (looped back to the design/data collection point) before being unleashed in hospitals, to avoid perpetuating health inequalities (Tong et al., 2022). This, however, was not the case at the outset of the COVID-19 pandemic, where research on White people was applied to AI systems managing X-rays to determine whether a patient had COVID-19 (Feng et al., 2020). The result was that in 2021, in the United States (US), several Black people were triaged incorrectly (Marabelli et al., 2021b). But it was during the long-term use of AI, in practice, that these data issues surfaced.

A similar example of an important issue with training data concerns facial recognition systems. These systems need to be trained with millions of faces (and several photos of the same person) to be able to recognize individuals; for example, in airports, at borders, and even by police with street cameras. It recently emerged that Facebook, for years, had used user photos (and associated tags) to train its own facial recognition system. Here, the problem concerning the reliability of facial recognition systems—evidence suggests that they are not reliable and penalize people with dark skin—is combined with privacy issues. In fact, Facebook users never consented to having their photos used to train facial recognition software. Next, we unpack a specific training data issue that concerns generative AI (ChatGPT), that can write impressively coherent essays and computer code, generate pictures from text, and more. The problem is that when such systems are unleashed on the general public, it can be very difficult to spot problems if they are not sufficiently scrutinized.

Vignette 1: Training data and ChatGPT

Generative pre-trained transformer (GPT) models are based on natural language processing (NLP) systems, and are capable of producing human-like texts and pic-

tures based on being fed with a huge amount of data, generally obtained by scraping the web, something contested from an ethical standpoint (Marabelli and Markus, 2017). But training NLP systems with data from the web has implications that go beyond typical issues associated with privacy and appropriation of intellectual property. Bender et al. (2021) noted that feeding NLP models using web data involves important issues leading to potential discrimination. Internet access is not evenly distributed worldwide, even within countries in the Global North that are believed to have broad internet access (compare rural areas of some US states). Also, scraping websites such as Reddit leads to data collected prevalently from males and young people (Bender and Friedman, 2018). Negative consequences of these biases in training data were observed only when ChatGPT models were unleashed in society (that is, designers did not capture the issues, and making GPT models available to the general public seemed to be a good thing). Incidents included chatbots being racist, sexist, and overall replying to user prompts in ways that are often far from objective. What is even more important is that these systems are now being used by organizations to aid (or sometimes replace) decision-making processes. For instance, a judge in Colombia employed ChatGPT to make a court ruling.¹ While for now this is an isolated case, it is illustrative of the potential worrying consequences of large-scale adoption of NLP systems for decision-making in organizations.

Epilogue

Data are key for AI, because these systems do not learn as humans do: unlike a toddler who takes along 2–3 photos of an elephant to recognize a live animal at the zoo, AI requires tens of thousands of photos to identify patterns and make accurate predictions, illustrating how AI is data-hungry. However, what AI can do with training data leads to unpredictable outcomes. While designers program AI to be responsive to customers (chatbots), once the issue is taken forward, for instance with NLP models, consequences cannot be accurately predicted, as we have seen in the case of ChatGPT (who would imagine that these systems would be used to rule in juridical settings?). The issue with implementation is different here, because several NLP models are freely available online, which makes it difficult to target context-specific rollouts. It is then paramount to focus on AI use in practice, to capture concerns and address them, looping back to the AI lifecycle.

AI and Power Asymmetries

Issue

As we illustrated in the previous vignette, AI systems are often used to aid decision-making processes or even to make (unsupervised) decisions such as hiring/firing systems (Acquisti and Fong, 2019), grade student work automatically (Mowat, 2020), and calculate employee performance, such as systems employed by Amazon in its warehouses (Delfanti and Frey, 2020). AI systems (and the companies which use them) hold power over customers (compare: nudging), patients (Tong et al., 2022), and citizens more generally. For instance, at the onset of the COVID-19 pan-

demic, algorithms were used to prioritize access to vaccines. Supply chains too are now dominated by AI, and allow suppliers to move goods towards more profitable (thus, wealthy) geographical areas.

Here, the “power problem,” as clearly outlined by Kalluri (2020), is that AI is developed and implemented in a top-down fashion, with little involvement of “end users.” It is hard to think of a collective development and implementation of AI, because the technical competences and financial resources to do so are in the hands of large companies. However, scrutinizing the unfolding of power dynamics when AI is used in practice may help to mitigate the power asymmetries that are currently taking place.

One notable example of AI generating power asymmetries concerns the gig economy, where delivery companies (for example, Deliveroo, Instacart), hospitality services (for example, Airbnb), and ridesharing companies (for example, Uber, Lyft) rely heavily on AI to manage their workforce; so-called algorithmic management (Mohlmann et al., 2023; Tarafdar et al., 2022). Designing and implementing systems aimed at managing the workforce is not a negative idea in principle; for instance, the literature on people analytics explains how automated systems can help in identifying opportunities for process improvement and assessing performance in an accurate fashion (Leonardi and Contractor, 2018). However, the long-term use of these systems to manage the workforce has proven detrimental to job satisfaction (Giermindl et al., 2021). A case in point is AI managing Uber drivers, as we illustrate in the following vignette.

Vignette 2: Uber drivers and the algorithm

Uber drivers are managed by an app that runs on most smartphones (IOS, Android), which connects (and matches) drivers with customers (riders). Here, the app rules. It tells drivers the route they need to follow to pick up and drop off customers, generates a reward system based on customer reviews, and in this way is able to punish drivers who do not comply—for instance, drivers who use an alternative Global Positioning System (GPS) that suggests more convenient and faster routes. In the beginning the Uber app was considered innocuous; or actually convenient, because it replaced humans managing drivers, allowing more profits (in theory, for both Uber and the drivers). However, after its implementation some key issues started to emerge.

Examples include the drivers feeling disempowered: they cannot make their own decisions on how to reach a destination when they carry passengers/customers (Mohlmann et al., 2023). The Uber drivers also feel powerless when they realize that they are not able to communicate with the app as they would with a manager. In fact, the Uber app is prescriptive when telling drivers what to do, and is not able to record and implement feedback if something does not work as it should; something that recent research called “broken loop learning” (Tarafdar et al., 2022). The idea to save money by assigning AI managers to Uber drivers was a good one. The algorithm did its job (design) and was implemented in the appropriate settings (the gig economy, where efficiency and automation are key to ensure workers’ good earnings). Yet,

their use in practice unveiled shortcomings that were hard to predict in the previous stages of the AI lifecycle.

Epilogue

AI generates power asymmetries. These are mainly due to the fact that actual users are not involved in the first two phases of the lifecycle (design and implementation). It is not realistic, for example, to think of involving the workforce in collectively designing AI. However, even when AI is designed and implemented in ways that seem to generate a win–win situation (companies make money, workers are fairly rewarded), its long-term use in practice reveals shortcomings. In the Uber driver vignette, feelings of frustration for not being able to provide feedback to virtual managers (AI) needs to be addressed by going back to the design point and focusing on (re)designing systems that, for example, accept and learn from user feedback. A more careful implementation, for instance explaining to users how some of the mechanisms of the underlying AI work, might also contribute to reducing frustration. However, some argue that, for instance, telling Uber drivers too much about how the app’s algorithm calculates driver performance might lead to drivers trying to game the system (Christin, 2017; Kellogg et al., 2020). This example of power imbalances at Uber is illustrative of emerging issues being captured only during long-term use of these technologies. Even whether Uber drivers would try to game the system is something that can be realized only once Uber drivers actually use the driving/managing app, which makes decisions for them as if it were their boss (Tarafdar et al., 2022).

AI and Automation

Issue

While whether AI is actually replacing humans and creating unemployment is a contentious subject, the fact that automation of basic tasks is taking place in most industries is a matter of fact. In accounting and audit, for instance, AI systems are able to spot irregularities much faster than humans, and are often extremely accurate in doing so (Richins et al., 2017). In the delivery industry, for instance in Amazon warehouses, AI monitors employees’ microtasks and takes notice even every time they take a short break to go to the restroom (Faraj et al., 2018). Local law enforcement agencies are currently employing robot dogs which are supposed to be used to defuse bombs, yet they are also armed with guns that can kill people (Vincent, 2021). Supermarkets, restaurants, and hospitals, in the aftermath of COVID-19, are increasingly using robots to provide basic services (ordering, checkouts, assistance, and so on) (Marabelli et al., 2021b)

Here the issue is twofold: one concerns replacing workers, thereby potentially creating unemployment; the other is about surveillance, where automation is in place and humans are managed by AI. This latter issue—the focus of the next vignette—is related to, yet different from, the previous power example. Automation is a term that, at least given the state of the art of AI systems, does not fully reflect all the overwhelming control that “automated systems” require. Here, paradoxically, AI hands

over decision-making processes to humans, often poorly qualified and exploited. The next vignette is about the dark side of automation with respect to its monitoring, and refers to Amazon.

Vignette 3: Surveillants and surveilled at Amazon

Amazon warehouses have become popular for making use of robots able to move items across shelves, picking and packing goods, and finalizing the shipment with labeling, quality checks, and so on. Warehouses, however, are far from being fully automated. In fact, a number of task workers are employed in Amazon warehouses who are required to supervise all this automation and make sure that every time a robot has issues (something rather frequent; for instance, when items fall from shelves) they “fix” the situation. All this can be seen as a necessary process towards full automation. However, the long-term use of these AI systems at Amazon has revealed an important aspect of surveillance. “Cameras are working on your station at all times,” a worker told a journalist of *The Verge* (McIntyre and Bradbury, 2022). “It’s kind of demeaning to have someone watching over your shoulder at every second.”

These cameras are managed by AI systems. They capture videos of warehouse workers and send them to a data center in India, where offshore Amazon workers review the videos and provide inputs aimed at improving Amazon’s AI systems, along with feedback on how well warehouse workers were able to help robots doing their job. Offshore reviewers receive up to 8000 videos per day and are paid as little as \$500/month. A report to the Bureau of Investigative Journalism indicated that these workers experience physical issues such as headaches, eye pain, and even deterioration in their eyesight. Other reviewers reported that they are also watched by cameras, while reviewing videos, which adds to this global chain of surveillance. The point here is that automation has a dark side, which goes behind the initial goal of the creation (design) and use (implementation) of these systems. Exploited individuals are required to provide feedback and make decisions in regimes of surveillance, all this with the goal of improving the accuracy of AI systems with “unprofessionally vetted” insights. To improve the systems, training here does not involve the previously discussed training data, but also a whole set of human activities that implicate hidden monitoring and poor work conditions.

Epilogue

Automation is a process that was introduced in organizations over a century ago, with the first assembly line systems at Ford Motor Company. But in the last decade, the increased use of AI systems to automate, along with technology’s ability to monitor the minutiae of our lives, has led to scenarios that can be defined as uncomfortable, unethical, or otherwise not in line with the spirit of progress once associated with automation. And given our example of Amazon warehouse automation, we wonder: how could one capture these issues without observing long-term use of automated systems at companies? Here, however, it is difficult to think of whether it is possible to go back to the design point and tweak AI for automation, or whether it is just the

implementation that needs to be substantially revised; for instance by generating training systems based solely on data, and that do not exploit human surveillance capabilities (and offshore outsourcing) to do so.

AI and Transparency

Issue

AI systems are made up of algorithms that often become increasingly complex as they are used in practical settings. This is the case of machine learning (ML), for instance, which changes as algorithms learn when unleashed in practical (that is, “real”) settings and interact with humans. A very common example relates to auto-correct systems on smartphones, that over time learn the writing style of the owner (use of unique ways to formulate sentences and terms that are not in the default dictionary). These systems make decisions (or suggestions) on the basis of our past behaviors (typing certain words or sentences). More advanced systems do not just customize outputs (by learning), but also modify their own way to process data (that is, the algorithm) to deliver such outputs. It thus becomes clear that these systems evolve over time. It is however often unclear (or nontransparent) how they provide insights, even if the insights are most of the time correct. Here, by transparency we refer to people’s ability (including the designers/programmers) to map the AI’s underlying algorithm and trace back an outcome to its inputs.

It is worth noting that less transparency often means more effective systems. In fact, traceable algorithms come at a price, which is associated with the limited number of variables being used. This is the reason why so-called “blackboxed” algorithms (algorithms that produce poorly traceable outputs) are widely used; these systems are continuously trained with “real-world” data, can process more variables than traditional AI systems, and do not need to be reconfigured manually every time a scenario changes (Asatiani et al., 2020). However, the underlying explanatory power when AI is used in this way is limited, decreases over time, and is difficult to understand even for the creators of the algorithms (Hosanagar and Jair, 2018). Lack of traceability comes with higher performance, increased ambiguity, and poor transparency. A different but related problem surfaces when companies do not disclose how AI systems work, to retain the algorithms’ intellectual property and to prevent others (employees or customers) from gaming the system (Christin, 2017; Kellogg et al., 2020). In this chapter, we focus on the former aspect of transparency: that of systems which need to be blackboxed to become more powerful, and make decisions that are not explainable. The following vignette is illustrative of AI operating in a way that confused the engineers who designed it.

Vignette 4: Occupy Wall Street

Occupy Wall Street (OWS) was a collective protest taking place in September 2011 in New York City, targeting big financial organizations and advocating for more income equality in the US, and denouncing the prominent role of money in politics. The protest attracted attention in the US and worldwide, and became a movement

(and a slogan) aimed at reducing the influence of corporations in politics, achieving a more balanced distribution of income, and promoting a reform of the banking/financial system to assist lower-income citizens. Social media were used prominently to promote the initial protest in September, with activists using Twitter intensively to coordinate the protest and make the public aware of what was happening in South Manhattan (the Wall Street district) where the protest originated.

Surprisingly enough, the various hashtags used by the activists were not trending sufficiently; meaning that, despite what was happening, Twitter (that is, the algorithm; note that Twitter uses advanced ML systems to help surface relevant content) did not seem to reflect the breadth and magnitude of the event. According to a report by journalist Tarleton Gillespie,² some suggested that Twitter was deliberately dropping the hashtag #occupywallstreet from its list of trending hashtags, and in doing so, preventing OWS from reaching a wider audience.

The initial reaction from the public was anger against Twitter, seen then as a mere political tool to control public opinion. Everyone was under the impression that the social media company was simply aligned with the financial and political status quo that OWS was trying to challenge, operating *de facto* censorship. However, Twitter's engineers denied any censorship, while not being able to explain why OWS had not become a trending topic; nor could they forensically trace back how all the data streams received by Twitter were processed by the platform's algorithms. In other words, the algorithm was so complex and blackboxed that its outputs could not be explained by the very programmers who designed and implemented it.

Epilogue

Transparency is a very fluid concept in the AI realm. When referring to the impossibility of tracing back an outcome to specific data, variables, and operations (that is, algorithms processing big data), it might pose ethical issues; especially because AI systems are supposed to aid or lead decision-making processes. What happens when an AI system makes a decision whose rationale is obscure, even to the people who programmed it? In the example outlined in the vignettes, protests were automatically censored. According to US laws, the OWS incident would not be considered a violation of the First Amendment (freedom of speech), which applies only to the government and affiliated institutions.

Yet, the fact that the algorithm's creators were not able to figure out why OWS's hashtag #occupywallstreet was not included in the trending ones is illustrative of the danger of AI systems currently used by several online platforms (social media, retailers, and so on). While this poses an issue of whether, and the extent to which, blackboxed algorithms should be implemented in "public" settings (or even designed as such: blackboxed), from our vignette it clearly emerges that it is only by scrutinizing these systems in practice that we can realize how dangerous they can be. Open source algorithms would to some extent jeopardize the competitive advantage of the companies developing such algorithms, yet this would provide a certain degree of transparency with respect to output being made available to the public (in fact, these systems can be scrutinized by online communities of programmers). Related to

transparency is accountability (who is responsible?), another peculiar characteristic of AI, which we discuss next.

AI and Accountability

Issue

Not knowing how and why an AI system delivers an output becomes problematic when these systems are, for example, discriminatory, because it is hard to hold someone accountable for something “done” by a machine. For instance, if a social media algorithm becomes discriminatory with respect to gender (as was a notorious Facebook algorithm that was showing certain job adverts only to men and others only to women) it is hard to isolate the proxy leading to this (Hao, 2021). The OWS example follows suit. To this end, it is problematic to point the finger at a specific error in programming the underlying algorithm, and thereby at a company which developed a specific AI system. Similar to transparency, accountability is problematic because decision-making processes that “go wrong” cannot be traced back to a specific actor.

Accountability cannot be addressed by designers, besides asking them to provide the specifics of their software. But intellectual property and the associated competitive advantage of building unique algorithms may prevent companies from disclosing code sources. Also, ML changes algorithms over time as they are fed with “real-world” data, and how these systems change based on new training data cannot be foreseen. Implementing a blackboxed algorithm makes it nearly impossible to hold a specific actor accountable. The following vignette illustrates precisely this, using another example from AI used on social platforms.

Vignette 5: AI promoting self-harm

In 2017, 14-year-old British citizen Molly Russel committed suicide, initially ruled as associated with depression, then investigated as somewhat “assisted” by social media posts about suicide (Meaker, 2022). In September 2021 the *Wall Street Journal* reported on a Meta employee (Frances Haugen) disclosing internal research demonstrating that Instagram algorithms took teenagers with eating disorders to online resources encouraging even more weight loss (Wells et al., 2021). In both cases, representatives of social media companies testified that (as was obvious) their algorithms were not designed or implemented to create harms. Thus, while social media AI manages exposure to news feeds and trending topics in a way that promotes “engagement” (that is, keeps the user on the platforms for long periods of time), the underlying algorithms were not designed to harm.

Epilogue

The vignette above exposes an evident tension between social media business models, which are aimed at keeping users on their platforms for as long as they can—so they can mine their data and serve targeted ads—and the potential side effects of this invasive strategy. The problem here is that the available research

linking mental health problems and social media is based on correlation and not causation, meaning that we do not know yet whether social media algorithms cause depression, or whether people affected by depression rely more on social media. One thing, however, is clear: AI accountability is still a grey area and needs to be investigated by ongoing observations of how AI systems affect our everyday lives. Automated decision-making can only magnify this issue, because mistakes made by these systems (that is, a hiring system that does not hire a qualified candidate because they have an accent, thereby discriminating) need to be traced back to someone (not something) who will be held accountable.

AI and the Environment

Issue

AI systems require immense computing capabilities to work effectively. For instance, they need to process vast amounts of (big) data to provide insights and support decision-making. Most recently, they were trained to translate text automatically, in different languages (Google Translator), talk with people (chatbot assistants, home assistants such as Alexa and Echo), produce “original” writing, and even create videos using texts/human instructions, as we previously explained. While AI (as a concept) has been around since the 1950s, it was only a decade ago that these things were possible, mainly because of the advent of big data and the increased computing capabilities (McAfee and Brynjolfsson, 2012).

All this, however, comes at a price. For instance, the emissions originated by training neural network models (advanced ML learning systems) can be compared to those of several jets carrying thousands of passengers overseas. Strubell et al. (2019) performed a lifecycle assessment of training AI models. They found that the process of training a single neural network can emit more than 626 000 pounds of carbon dioxide equivalent: nearly five times the lifetime emissions of an average automobile. Here, interestingly, these systems were designed with the awareness that they would need relevant resources to work and improve (that is, continuous learning of ML). But the implementers underestimated the long-term effect of the broad diffusion of AI. To highlight, the “resource” issue is something that emerged with AI use in practice, in the long term, as the systems build on each other and become more sophisticated and capable.

Vignette 6: The unequal impact of climate change

AI systems in need of energy (in the form of electricity) and cooling systems are increasingly having a prominent impact on the environment and climate change. Besides the most obvious long-term threats to the planet (global warming), it is worth noting that impacts on climate are not shared evenly across countries. In fact, it is often the case that countries which contribute the least to climate change (countries with poor use of coal-related resources, and where AI is used in limited ways) pay the highest price of climate change. One such example is India, with nearly 18 percent of the world population, which generates 3 percent of the world pollution but pays

a high price because of global warming, with temperature records in June and July 2022 indicative of a long-term trend (Williams, 2022). In addition, just 12 percent of India's population can benefit from air conditioning at home. But in India, even the "privileged" experience issues, because of the frequent power outages (they cannot use air conditioning all day long); and power outages create water shortages, which in the warmest months of 2022 killed dozens of Indian citizens (Irfan, 2022).

Epilogue

AI is energy-demanding, and this affects global warming. It is worth noting that scientists are still weighing the benefits and pitfalls of AI with respect to the environment; some AI can actually help to lower carbon emissions, such as that managing smart agriculture, making transport on land more sustainable by predicting and optimizing traffic, and handling intermittency of renewable energy. However, in the short term, the main outcomes of AI with respect to the environment are negative, and most importantly they hit poor countries hardest (and within poor countries, poor populations, as in our vignette on India). This poses ethical issues concerning how AI should be designed in ways that limit the need to train systems with large amounts of data; something that people are already working on (Sucholutsky and Schonlau, 2021); and where it is worth implementing AI. For instance, AI is now used to mine bitcoins, another energy-demanding process that some might see as unethical, also given the most recent disturbing events associated with cryptocurrencies, such as the FTX bankruptcy (Reuters, 2022). But neither designers nor implementers could forecast the climate-related consequences of AI systems unleashed on such a broad scale. At what (environmental) cost are we willing to use automated decision-making systems?

DISCUSSION AND IMPLICATIONS

In this chapter, we have demonstrated that to capture most of the ethical implications of AI systems involved in decision-making processes, it is extremely important to focus on their use in practice, because it is often impossible to foresee these issues at the design and implementation stages of the AI lifecycle. Table 20.1 summarizes our insights.

Our focus on AI use in practice builds on prior research suggesting that technologies (physical and digital artifacts) hold affordances that were not embedded in the original design, but are discovered, over time, with use (Leonardi, 2011). A focus on design can be useful to prevent or limit unintended uses of AI (Noble, 2018). But we are skeptical that ethical design (cf. d'Aquin et al., 2018; Martin, 2019) will lead straightforwardly to good implementation and use. Implementation has its own issues, as we have illustrated with our vignettes. In particular, because AI is fed with only so many variables, context is extremely important, and AI portability becomes problematic (Luca et al., 2016). In this chapter, we demonstrate how critical, ethical issues can only be identified through observing the long-term use of AI. By doing so,

Table 20.1 *AI characteristics, ethical considerations and remedies*

AI characteristics	Key ethical considerations	Potential remedies
AI and training data	AI works with training data that are not inclusive	Carefully select more inclusive sources of training data(sets) Implement AI systems in contexts that are appropriate with respect to the dataset they were trained with
AI and power asymmetries	Algorithmic management generates feelings of oppression and does not allow for feedback from AI users	Address the “broken loop learning” by focusing on learning algorithms Conduct responsible implementation of these systems
AI and automation	Paradoxically, automation requires manual vetting of AI systems, often involving labor exploitation	Limit offshoring of the workforce vetting AI automations Ensure that supervision of AI is done by qualified individuals
AI and transparency	AI systems generate outputs that often cannot be explained; therefore unwanted outcomes can hardly be addressed	Consider releasing less powerful, yet more transparent systems Develop open source algorithms that can be scrutinized by online communities of programmers
AI and accountability	It is challenging to hold someone accountable when AI systems generate problematic outputs	Laws and regulations should push companies to release internal research findings on the potential harms of their technologies being released to the public
AI and environment	The burden of AI on climate changes is largely absorbed (unfairly) by Global South countries	Conduct cost–benefit analysis when releasing new technologies requiring high computational resources Consider ways to train systems with limited amounts of training data

we can better understand how to improve AI through redesign and reconsideration of its implementation.

We argue that it is of paramount importance to focus on affordances in practice and recognize that there will be a continuous need for adaptation in the post-implementation period (Costa, 2018). We recognize that this view of the AI lifecycle is largely optimistic and does not account for systems that are purposely designed and implemented for generating revenues at the expense of “end users” (for example, the case of Google’s algorithm that was optimized for the racially discriminatory patterns of past users; Benjamin, 2019; Noble, 2018). But regardless of whether design and implementation are purposely or unwittingly problematic, it is only by evaluating long-term societal effects that it is possible for society to spot and address issues. One prominent example is the increasingly pervasive use of ChatGPT by businesses for decision-making. While its adoption is massive, it is unclear what will be the long-term consequences of its use; large companies are considering slowing down rollouts, because looking at design and implementation appears to be insufficient to clearly understand how generative AI will affect organizational practices and people’s everyday lives.

Our argument that we need to focus on AI use in practice to spot and address ethical issues has two key organizational implications. First, viewing the AI life-cycle as a nonlinear, messy unfolding of practices demands that companies invest resources (and give up business opportunities) to go back to the drawing board, when needed, to review problematic algorithms or inappropriate implementations (that is, those having to do with AI portability). This seems highly unlikely to happen in the current hypercompetitive AI industry. In this regard, we see laws and regulations as the obvious remedy for poorly scrutinized AI applications (Marabelli et al., 2021a). For instance, in the US the AI market is still highly unregulated, with companies being able to use user data collected by home assistants (Alexa, Echo) to train voice recognition systems. The European Union (EU) has taken more substantial steps to regulate the use in practice of AI (European Union, 2021). However, in a global internet, more needs to be done to assess AI use.

Second, there is a lack of research (primary data, collected at AI companies) on how AI is designed and implemented. This is rather normal, given that companies would hardly give access to their research and development (R&D) department to people wanting to assess their ethical conduct in developing AI. For instance, Uber became famous for not allowing data-sharing with academics for most of its projects (Tarafdar et al., 2022). However, the fact that we as researchers cannot access data from AI companies has the negative effect of preventing us from undertaking studies that inform lawmakers on the potential ethical implications of AI use (Marabelli and Newell, 2022).

CONCLUSIONS

AI systems hold immense potential and opportunities for individuals, organizations, and society. However, ethical issues proliferate and, we argue, a focus on implementation and design is insufficient to address them. We have therefore suggested that looking at long-term use in practice of these systems is needed to readdress issues at the design and implementation points. We have relied upon Marabelli et al.'s (2021a) algorithmic decision-making systems framework to illustrate the importance of looping back to the initial phases of the AI lifecycle once an ethical concern is spotted as a system is unleashed on society.

NOTES

- 1 <https://www.vice.com/en/article/k7bdmv/judge-used-chatgpt-to-make-court-decision>.
- 2 <https://limn.it/articles/can-an-algorithm-be-wrong/>.

REFERENCES

- Acquisti, A., and Fong, C. (2019). An experiment in hiring discrimination via online social networks. *Management Science*, 66(3), 1005–1024.
- Asatiani, A., Malo, P., Nagbol, P.R., Penttinen, E., Rinta-Kahila, T., and Salovaara, A. (2020). Challenges of explaining the behavior of blackbox AI systems. *MIS Quarterly Executive*, 19(4), 259–278.
- Bender, E.M., and Friedman, B. (2018). Data statements for natural language processing: toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6, 587–604.
- Bender, E.M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.
- Benjamin, R. (2019). *Race after Technology: Abolitionist Tools for the New Jim Code*. Polity Press.
- Browning, M., and Arrigo, B. (2021). Stop and risk: policing, data, and the digital age of discrimination. *American Journal of Criminal Justice*, 46(2), 298–316.
- Christin, A. (2017). Algorithms in practice: comparing web journalism and criminal justice. *Big Data and Society*, 4(2), 1–14.
- Cook, S.D., and Brown, J.S. (1999). Bridging epistemologies: the generative dance between organizational knowledge and organizational knowing. *Organization Science*, 10(4), 381–400.
- Costa, E. (2018). Affordances-in-practice: an ethnographic critique of social media logic and context collapse. *New Media and Society*, 20(10), 3641–3656.
- Crawford, K. (2021). *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.
- d'Aquin, M., Troullinou, P., O'Connor, N.E., Cullen, A., Faller, G., and Holden, L. (2018). Towards an “ethics by design” methodology for AI research projects. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*.
- Delfanti, A., and Frey, B. (2020). Humanly extended automation or the future of work seen through Amazon patents. *Science, Technology, and Human Values*, 46(3), 655–682.
- Dwivedi, Y.K., Kshetri, N., Hughes, L., Slade, E.L., Jeyaraj, A., Kar, A.K., Baabdullah, A.M., Koohang, A., Raghavan, V., and Ahuja, M. (2023). “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, 1–63.
- European Union (2021). *Europe Fit for the Digital Age: Commission Proposes New Rules and Actions for Excellence and Trust in Artificial Intelligence*.
- Faraj, S., Pachidi, S., and Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. *Information and Organization*, 28(1), 62–70.
- Feng, C., Huang, Z., Wang, L., Chen, X., Zhai, Y., Zhu, F., Chen, H., et al. (2020). A novel triage tool of artificial intelligence assisted diagnosis aid system for suspected COVID-19 pneumonia in fever clinics. *The Lancet*. <http://dx.doi.org/10.2139/ssrn.3551355>.
- Finocchiaro, J., Maio, R., Monachou, F., Patro, G.K., Raghavan, M., Stoica, A.-A., and Tsirtsis, S. (2021). Bridging machine learning and mechanism design towards algorithmic fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.
- Gierminda, L.M., Strich, F., Christ, O., Leicht-Deobald, U., and Redzepi, A. (2021). The dark sides of people analytics: reviewing the perils for organisations and employees. *European Journal of Information Systems*, June 2, 1–26. <https://www.tandfonline.com/doi/epdf/10.1080/0960085X.2021.1927213?needAccess=true>.

- Hao, K. (2021). Facebook's ad algorithms are still excluding women from seeing jobs. *MIT Technology Review*. <https://www.technologyreview.com/2021/04/09/1022217/facebook-ad-algorithm-sex-discrimination/>.
- Hosanagar, K., and Jair, L. (2018). We need transparency in algorithms but too much can backfire. *Harvard Business Review*, July.
- Irfan, U. (2022). The air conditioning paradox. *Vox*. <https://www.vox.com/science-and-health/23067049/heat-wave-air-conditioning-cooling-india-climate-change>.
- Kalluri, P. (2020). Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature*, 583(7815), 169.
- Kaye, K. (2022). Why an "us vs. them" approach to China lets the US avoid hard AI questions. *Protocol*. <https://www.protocol.com/enterprise/us-china-ai-fear-military>.
- Kellogg, K., Valentine, M., and Christin, A. (2020). Algorithms at work: the new contested terrain of control. *Academy of Management Annals*, 14(1), 366–410.
- Kessler, S. (2023). Getting rid of remote work will take more than a downturn. *New York Times*.
- Leonardi, P.M. (2011). When flexible routines meet flexible technologies: affordance, constraint, and the imbrication of human and material agencies. *MIS Quarterly*, 35(1), 147–167.
- Leonardi, P., and Contractor, N. (2018). Better people analytics. *Harvard Business Review*, November–December. <https://hbr.org/2018/11/better-people-analytics>.
- Luca, M., Kleinberg, J., and Mullainathan, S. (2016). Algorithms need managers, too. *Harvard Business Review*, 94(1), 96–101.
- Marabelli, M., and Markus, M.L. (2017). Researching big data research: ethical implications for IS Scholars. *Americas Conference of Information Systems (AMCIS)*, Boston, MA.
- Marabelli, M., and Newell, S. (2022). Everything you always wanted to know about the metaverse* (*but were afraid to ask). *Academy of Management Annual Meeting*, Seattle, WA.
- Marabelli, M., Newell, S., and Handunge, V. (2021a). The lifecycle of algorithmic decision-making systems: organizational choices and ethical challenges. *Journal of Strategic Information Systems*, 30, 1–15.
- Marabelli, M., Vaast, E., and Li, L. (2021b). Preventing digital scars of COVID-19. *European Journal of Information Systems*, 30(2), 176–192.
- Markus, M.L., Marabelli, M., and Zhu, C.X. (2019). POETs and quants: ethics education for data scientists and managers. Presented at the third RICK Workshop, Cambridge, UK.
- Martin, K. (2019). Designing ethical algorithms. *MIS Quarterly Executive*, 18(2), 129–142.
- McAfee, A., and Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard Business Review*, 90(10), 61–67.
- McIntyre, N., and Bradbury, R. (2022). An offshore workforce is training Amazon's warehouse-monitoring algorithms. *Verge*. <https://www.theverge.com/2022/11/21/23466219/amazon-warehouse-surveillance-camera-offshore-workers-india-costa-rica>.
- Meaker, M. (2022). How A British Teen's Death Changed Social Media. *Wired*. <https://www.wired.com/story/how-a-british-teens-death-changed-social-media/>.
- Mohlmann, M., Alves De Lima Salge, C., and Marabelli, M. (2023). Algorithm sensemaking: how platform workers make sense of algorithmic management. *Journal of the Association for Information Systems*, 24(1), 35–64.
- Mowat, E. (2020). Marked down SQA results: Scotland's poorest kids TWICE as likely to have exam results downgraded to fails compared to rich pupils. *Scottish Sun*. <https://www.thescottishsun.co.uk/news/5888751/sqa-results-downgraded-marks-poor-rich-deprived/>.
- Newell, S., Robertson, M., Scarbrough, H., and Swan, J. (2009). *Managing Knowledge Work and Innovation*. Palgrave Macmillan.
- Noble, S.U. (2018). Algorithms of oppression. In *Algorithms of Oppression*. New York University Press.

- O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books.
- Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453.
- Perez, C.C. (2019). *Invisible Women: Data Bias in a World Designed for Men*. Abrams.
- Ponsford, M. (2022). House-flipping algorithms are coming to your neighborhood. *MIT Technology Review*. <https://www.technologyreview.com/2022/04/13/1049227/house-flipping-algorithms-are-coming-to-your-neighborhood/>.
- Reuters (2022). Factbox: Crypto companies crash into bankruptcy.
- Richins, G., Stapleton, A., Stratopoulos, T.C., and Wong, C. (2017). Big data analytics: opportunity or threat for the accounting profession? *Journal of Information Systems*, 31(3), 63–79.
- Rotman, D. (2023). ChatGPT is about to revolutionize the economy. We need to decide what that looks like. *MIT Technology Review*. <https://www.technologyreview.com/2023/03/25/1070275/chatgpt-revolutionize-economy-decide-what-looks-like/>.
- Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *arXiv preprint*. <https://arxiv.org/pdf/1906.02243.pdf%22%3EWachstum>.
- Sucholutsky, I., and Schonlau, M. (2021). Less than one-shot learning: learning N classes from M < N samples. *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Tarafdar, M., Page, X., and Marabelli, M. (2022). Algorithms as co-workers: human algorithm role interactions in algorithmic work. *Information Systems Journal*, 33(2), 232–267.
- Tong, Y., Tan, C.-H., Sia, C.L., Shi, Y., and Teo, H.-H. (2022). Rural–urban healthcare access inequality challenge: transformative roles of information technology. *Management Information Systems Quarterly*, 46(4), 1937–1982.
- Vincent, J. (2021). They're putting guns on robot dogs now. *The Verge*. <https://www.theverge.com/2021/10/14/22726111/robot-dogs-with-guns-sword-international-ghost-robotics>.
- Wells, G., Horwitz, J., and Seetharaman, D. (2021). Facebook knows Instagram is toxic for teen girls, company documents show. *Wall Street Journal*. <https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739>.
- Williams, R. (2022). India's deadly heatwaves, and the need for carbon removal. *MIT Technology Review*. <https://www.technologyreview.com/2022/07/05/1055436/download-india-deadly-heatwaves-climate-change-carbon-removal/>.

Index

- accountability 11, 84, 128, 133, 175, 260, 266–78, 368–9, 371
- actionable insights 195–209
- actor-network theory 99, 101, 106
- agentic IS artifacts 49–51
- AI experts 39–40, 42, 45, 47, 49–50, 52–3
- AI-in-the-loop 287, 289, 293, 295–6, 298–9
- Airbnb 363
- airplanes 83
- Alexa 369
- algorithmic isomorphism 342–55
- algorithmic management 291–2
- Amazon 136, 260, 362, 364–5
- Apple 260
- Aristotle 145, 147–8, 153
- artificial intelligence
 - black box issue *see* black box issue
 - chatbots 97–108, 123, 310, 313, 362, 369
 - circumspection 112–24
 - data analytics 20, 38, 44, 53, 115, 179–92, 197–9, 348
 - data sourcing for data-driven applications 17–32
 - data work as organizing principle in developing 38–54
 - decision-making in complex intelligent systems 160–76
 - definition of 99, 123, 145–6, 164–5, 345
 - ecology of explaining 214–23
 - ethical implications of 70–71, 112, 126, 130, 132–3, 135–7, 146–7, 359–72
 - see also* biases; fairness
 - experts *see* AI experts
 - explainable 84–5, 217–18, 221–2, 241
 - see also* interpretable AI
 - futuring techniques to enhance 322–37
 - human judgment and 144–56
 - human–AI workplace relationship configurations 282–99
 - hype/reality of in decision-making 1–3
 - integrative framework of in decision-making 9–12
 - interpretable 83–4, 240–61
 - see also* explainable AI
 - as mechanism of algorithmic isomorphism 342–55
 - missing link with data and decision-making 195–209
 - natural language processing 58–73, 97, 346, 349, 361–2
 - for naturalistic decision-making 80–94
 - and occupational identity 305–20
 - in public sector decision-making 266–78
 - responsible governance 126–37
 - synthetic stakeholders 226–37
- artificial selection environments 327–30, 334–6
- Ashby’s Law of Requisite Variety 84, 92
- Atlantic Lady* 80–82, 93
- attentional grain 325–7, 330, 332
- augmentation 2, 8, 50, 123, 151, 165, 171, 245, 249, 252, 256, 260, 282–5, 292–3, 299, 305–6, 308, 315–19, 333, 337
- autocracy by expertise 330, 335
- automated decision-making systems 144–56
- automation 8, 50, 69, 85, 123–4, 144–56, 199, 241, 283–5, 292–3, 298–9, 305–6, 308–9, 315–19, 323, 364–6, 371
- automotive industry 310–18
- autonomous cars 84
- autonomous weapons 290
- autonomy 84, 98, 112, 128, 146, 164–6, 176, 215, 297, 306–8, 313, 318
- AYLIEN 63, 66
- banking 308–16, 318
- Bard AI 323
- behavioral learning 18, 23–4, 27, 32
- biases 2, 42, 72, 84, 132–3, 136, 146, 196–8, 248, 252, 259–60, 315, 359–62
- confirmation 272
- experiential 324, 330
- gender 70, 353
- historical 70, 72
- racial 70, 353, 360–61
- work-related 352–3

- black box issue 84, 93, 100, 112, 137, 155–6, 181, 208, 217, 221, 240–41, 251, 253, 268, 277, 329, 337
- black swans 333
- blockchain 359
- Boeing 83
- Boston Consulting Group 130
- bounded rationality 3–4, 162–5, 171, 198, 229
- broken loop learning 363–4
- business intelligence and analytics 195–6, 198–200, 203–8
- business models 1, 60, 134–7, 160, 197, 368

- C# 63, 65
- Cambridge Analytica 133
- Captum 259
- certification 137, 160, 226, 228–9, 235
- chatbots 97–108, 123, 310, 313, 362, 369
- ChatGPT 296, 361–2, 371
- circumspection 112–24
- classical decision making 82
- climate change 226, 369–71
- coercive isomorphism 344
- COMPAS 272–5
- complex intelligent systems 160–76
- complex products and systems 160, 165–6, 170
- confirmation bias 272
- consequentialism 360
- conversational agents *see* chatbots
- convolutional neural networks 68–9
- corporate social responsibility 130, 226, 228, 344
- cosine similarity 62
- COVID-19 pandemic 98, 103, 151, 359, 361–4
- Cradle to Cradle 228
- criminal justice system 266, 272–5, 286, 362
see also law enforcement
- Cruelty Free 228
- Crunchbase 70
- cryptocurrencies 370
- customer lifecycle management 183–91
- customer relationship management 186–8
- cybernetics 84

- D3 65
- Danish General Practitioners' Database 30–31
- data analytics 20, 38, 44, 53, 115, 179–92, 197–9, 348
- data labeling 29–30
- data ownership 133
- data scientists 7, 10–11, 13, 23, 28–30, 45, 50, 86–90, 92, 116, 118–19, 121, 180–92, 196, 200–201, 214, 218–19, 247, 251, 349
- data sourcing 17–32, 87–8
- data-based effectuation 48–9
- data-centric models 246, 251–4
- data-driven applications 17–32
- data-driven decision-making 196–7, 200–201
- DDSM 250
- decision-making
 - automated decision-making systems 144–56
 - classical 82
 - in complex intelligent systems 160–76
 - data-driven 196–7, 200–201
 - engaging the environment in 226–37
 - ethical implications of AI in 359–72
 - futuring techniques to enhance AI in 322–37
 - human 10–12, 51, 82–3, 144–56, 162, 180, 197, 269, 286
 - hype/reality of AI in 1–3
 - integrative framework of AI in 9–12
 - literature on organizational 3–4
 - missing link with data and AI 195–209
 - natural environment in 228–9
 - naturalistic 80–94
 - public sector 266–78
 - regarding data sourcing 17–32
 - sequential 165, 170, 285
 - strategic 85, 161, 199–200, 230, 235, 322–37, 352
- deep learning 2, 39, 41–2, 123, 198–9, 319
- deep neural networks 240
- delegation 99, 101, 106, 108, 170, 283–4, 290–92, 295, 298
 - definition of 49–50, 99, 108
 - multi-faceted 49–52
- Deliveroo 363
- depression 368–9
- Dice similarity 62–3
- digital platforms 21, 133, 151
see also individual platforms
- distributed autonomous organizations 236–7
- distributed ledger technologies 226, 229–30, 236–7
- domain experts 39, 42, 45, 49–53, 82, 100

- EcdiSim 89–90, 93
- Echo 369
- electronic health record 167–8, 250, 307
- electronic support systems 80–85, 87–91
- emoticons 66
- emotions 150–52, 155
- entangled accountability 266–78
- Enterprise Miner 65
- environmental, social, and governance 226, 228–9
- epistemic differences 185–6
- epistemic knowledge 153
- epistemic uncertainty 39–44, 46, 52–3
- error reports 119–20
- ethical implications of AI 70–71, 112, 126, 130, 132–3, 135–7, 146–7, 359–72
 - see also* biases; fairness
- European Union 372
- Ever Given* 87
- Excel 28, 116, 186, 202–4
- experts
 - AI experts 39–40, 42, 45, 47, 49–50, 52–3
 - domain experts 39, 42, 45, 49–53, 82, 100
- explainable AI 84–5, 217–18, 221–2, 241
 - see also* interpretable AI
 - ecology of explaining 214–23
- Exxon Valdez* 228
- Facebook 105, 345, 361, 368
- facial expressions 68–9
- facial recognition systems 361
- Fair Trade 228
- fairness 39, 51–2, 69, 71–2, 98, 112, 122, 128, 228, 241, 255, 259–60, 311
- firefighting 82
- Ford Motor Company 365
- FTX bankruptcy 370
- futuring techniques 322–37
- gender biases 70, 353
- General Data Protection Regulation 97–8
- global positioning system 88, 363
- global satellite navigation systems 85–6, 88
- Google 70, 260, 347, 360, 369
- governance, responsible 126–37
- GPT-4 323, 326
- healthcare 305, 307, 325
 - electronic health record 167–8, 250, 307
 - medical data 30–31, 149, 151, 160–61, 164
 - medical imaging 17, 170–71, 240–61, 325, 328
 - personalized medicine 161, 166–76
- Hierarchical Document Process 68–9
- HireVue 69
- historical biases 70, 72
- historical data 2–3, 117, 179, 186, 271, 361
- horizon concept 123
- human decision-making 10–12, 51, 82–3, 144–56, 162, 180, 197, 269, 286
- human judgment 144–56, 168, 179, 197, 286
- human resources 47, 81, 99, 286, 298, 311–18
- human–AI workplace relationship
 - configurations 282–99
- human-in-the-loop 289–90, 293, 296–9
- human–machine interface 84, 91–3, 112, 155–6, 295
- hybrid intelligence 285
- IAIA-BL 252
- INbreast 250, 252
- innovation literature 71–2
 - see also* patents
- Instacart 363
- Instagram 368
- insurance 310–13, 317
- International Consortium for Personalized Medicine 167
- interpretable AI 83–4, 240–61
 - see also* explainable AI
- intuition 152–3, 155
- inverse document frequency 61–2, 71
- IS sourcing 6, 18–22, 26, 32
- isomorphism
 - algorithmic 342–55
 - mechanisms of institutional 343–5
- issue detection 322–7
- Jaccard similarity 62–3, 71
- Java 65–6
- key performance indicators 122, 192
- KNIME 63
- knowledge boundaries 41, 182, 190–91, 353
- knowledge interlace 46–8
- Kodak Corporation 18
- Kraslava* 80–82, 93
- large language models 209, 322–3, 326

- Last.fm 49
- Latent Dirichlet Allocation 68–9
- law enforcement 214–22, 267, 352, 364
- learning algorithms 29, 47, 63, 65, 112, 116, 123, 165, 179–83, 190–91, 201, 215, 217, 221, 342–3, 347–8, 371
 - see also* machine learning
- learning experiences 24–7, 30, 32
- lemmatization 59–60, 66
- LIME 259
- lived experience 152
- locus of design 230–36
- Loomis v Wisconsin* 266, 272–5
- Lyft 363
- machine learning 2, 39–42, 47, 49–51, 112, 115, 123, 152, 165, 172, 181, 195, 198–9, 209, 214–15, 221, 226, 240–42, 244–56, 258–61, 289–90, 342, 346–7, 353, 366, 368–9
 - see also* learning algorithms
- natural language processing 58–73, 97, 346, 349, 361–2
- tools 63–72
 - see also individual tools*
- mammograms 242–6, 250, 255–9
- maritime navigation 80–94
- maritime trade 27–9
- MATLAB 65
- medical data 30–31, 149, 151, 160–61, 164
 - electronic health record 167–8, 250, 307
 - personalized medicine 161, 166–76
- medical imaging 17, 170–71, 240–61, 325, 328
- mental health 368–9
- mental time travel 333
- metadata 20
- mimetic isomorphism 344
- model designers 240, 242, 245–52, 254–8
- model-centric models 246, 251–4
- multi-faceted delegation 49–52
- naïve-Bayes 64, 66–7
- National Marine Electronics Association 87
- natural environment 228–9
- natural language processing 58–73, 97, 346, 349, 361–2
- naturalistic decision-making 80–94
- Neptune consortium 85–94
- neural networks 1, 64–6, 68, 70, 72, 170, 179, 195, 218, 369
 - deep neural networks 240
- N-grams 66
- NLTK 63, 65–6, 68
- normative isomorphism 344
- occupational identity 305–20
- Occupy Wall Street 366–8
- Opinion Lexicon 61
- opinion mining 66
- organizational culture 24, 192
- organizational learning 6, 18, 23–7, 29, 32, 355
- Paris Agreement 133
- patents 60, 68–9, 71
- path dependence 135–6, 326–7, 329–30, 332, 336
- pattern recognition 144, 147, 152–3
- Pearson correlation 62
- personalized medicine 161, 166–76
- Personalized Medicine Initiative 167
- personification 100, 103, 231–4, 237
- phronesis 145, 147–55
- policing 214–22, 267, 352, 364
- Porter Stemmer 60
- power balances 23, 44, 362–4, 371
- precision medicine *see* personalized medicine
- predictive models 2, 116, 119–21, 123, 169, 186, 208, 307, 345
- predictive policing 214–22
- predictive text 347
- privacy 19, 128, 246, 362
 - General Data Protection Regulation 97–8
- profiling 222, 272–4, 276
- programming languages 63–72
 - see also individual languages*
- public sector decision-making 266–78
- public service delivery 97–108, 266–78
- Python 58, 63, 65–6, 68
- Qlik 65
- R 58, 63, 65
- racial biases 70, 353, 360–61
- radiologists 242–6, 255, 257–9
- RapidMiner 63–6
- recidivism profiling 272–4, 286
- Reddit 362
- regression models 71, 186, 218
- relational governance 19, 21–2

- reliability 20, 112, 172–4, 259, 290, 308, 361
- robo-advisors 291, 297–8
- robotics 123, 166, 168, 291, 297, 364

- SAS 65
- satisficing 162–4, 198
- security 19
- self-harm 368
- sensory perception 151–2, 155
- sentence embedding 70
- sentiment analysis 58–9, 69
- SENTIWORDNET 61
- sequential decision-making 165, 170, 285
- sexism 136
- SHAP 259
- Shapley Additive Explanations method 252
- SHL 69
- short-termism 328, 330
- social media 17, 64, 66, 105, 137, 181, 345, 361, 367–9
- software as a service 342
- speech recognition 59, 165
- sports analytics 286
- stakeholder theory 226–9, 235, 237
- stemming 59–60
- stereotypes 70, 118
- strategic decision-making 85, 161, 199–200, 230, 235, 322–37, 352
- suicide 368
- superforecasters 290
- synthetic stakeholders 226–37

- Tableau 65
- techne 153–4
- term frequency 58, 61–3, 71
- textual analysis 58–73
- textual corpora 58–9, 66–7
- textual similarity 58, 62–3
- time horizons 331–3, 336
- timing and pacing modulation 333–4, 336

- tokenization 59–61
- topic modeling 58, 67–9
- training data 360–62, 371
- transactional governance 19, 21–2
- transparency 22, 112, 128, 133, 218, 230, 241, 259–60, 319, 366–8, 371
- Treebanks 67
- trust 22, 30, 40, 43, 69, 90, 132, 186, 241
- Twitter 64, 66, 367

- Uber 292, 363–4, 372
- Ubisoft 287, 296, 299
- UKVI system 266, 271–5
- uncertainty 43–4, 162–3, 167
 - epistemic 39–44, 46, 52–3

- value judgments 150–51, 334
- vector space model 61–2, 67
- video games 287
- virtual assistants 98, 346
 - see also* chatbots
- virtual reality 94
- visas 266–7, 271–5
- voice clearing techniques 335–6

- Wayback Machine 70
- WEKA 63
- welfare services 97–108
- Whanganui River 226, 229
- whatness 148–50, 152
- WhatsApp 105
- wicked problems 152, 154–5
- Wikipedia 70
- word embeddings 69–71
- word2vec 70
- WordNet 60, 66

- X 64, 66, 367

- Zillow 1–2, 11–12