

L. VALTONEN

Rationality in Artificial Intelligence Decision-making

L. VALTONEN

Rationality in Artificial Intelligence Decision-making

ACADEMIC DISSERTATION

To be presented, with the permission of
the Faculty of Management and Business
of Tampere University,
for public discussion in the auditorium Pieni sali 1
of the Festia, Korkeakoulunkatu 8, Tampere,
on 20 October 2023, at 12 o'clock.

ACADEMIC DISSERTATION

Tampere University, Faculty of Management and Business
Finland

| | | |
|-------------------------------|---|--|
| <i>Responsible supervisor</i> | Professor Saku Mäkinen University of Turku Finland | |
| <i>Follow-up group</i> | Senior Research Fellow Ulla Saari Tampere University Finland | DSc (Tech) Jari Hämäläinen Wirepas Finland |
| | Professor Kaisa Väänänen Tampere University Finland | DSc (Tech) Johanna Kirjavainen Leading Partners Oy Finland |
| <i>Pre-examiners</i> | Professor Ari-Pekka Hameri University of Lausanne Switzerland | Professor John Christiansen Copenhagen Business School Denmark |
| <i>Opponent</i> | Emeritus Professor Timo Airaksinen University of Helsinki Finland | |
| <i>Custos</i> | Professor Miia Martinsuo Tampere University Finland | |

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

Copyright ©2023 author

Cover design: Roihu Inc.

ISBN 978-952-03-3071-2 (print)

ISBN 978-952-03-3072-9 (pdf)

ISSN 2489-9860 (print)

ISSN 2490-0028 (pdf)

<http://urn.fi/URN:ISBN:978-952-03-3072-9>



Carbon dioxide emissions from printing Tampere University dissertations have been compensated.

PunaMusta Oy – Yliopistopaino
Joensuu 2023

PREFACE

Initially I began this research with exploring the use of artificial intelligence as a technology to support decision-making. However, coming from a background of mathematics, what I discovered as I adventured further into the depths of management and organization literature on AI surprised me. I discovered literature highlighting AI as the epitome of rationality and objectivity all the while ignoring the real capabilities and biases present. As my honourable opponent defines it, I discovered AI often to be a fetish for decision-making—something that functions in the world only if its origin and material reality is ignored.

This led me to study what it is about AI that lends the technology to such practices, and what are the implications for decision-making. This journey led me to interesting findings to finally more philosophical approaches to tackling the questions I had. Those worked well. I found the timing of doing this research exciting, as while reflecting on these topics I saw friends and colleagues fiddling with chatGPT as a novel curiosity and the world reacting to large language models in various ways. I hope this dissertation provides insights to the field that will further using AI for that which it can do and is actually good for. Indeed, despite the focus the reader will discover in this dissertation, AI, when used right, can do a lot of good.

This dissertation owes its existence to many people who deserve my utmost gratitude. Firstly, I thank my supervisor, Professor Saku Mäkinen, for his guidance and wisdom along the way. I came to your office with my fresh bachelor's degree to begin a mandatory literature review course after changing my major to IEM specifically because I did not want to become a researcher—and look what happened. How you manage to balance between your PhD students developing into independent researchers with intellectual freedom while providing unwavering support and inspiration on their diverse topics may remain a mystery to me. Regardless, I am grateful for it. Thank you to the follow-up group members Senior Research Fellow Ulla Saari, Professor Kaisa Väänänen, DSc. (Tech) Jari Hämäläinen, and former

member DSc. (Tech) Johanna Kirjavainen for their insightful feedback that made this work better. Ulla and Johanna, thank you for being inspiring and supportive colleagues, supervisors, and co-authors.

I am grateful to Emeritus Professor Timo Airaksinen for accepting the invitation to act as the opponent for this dissertation. Thank you Professor Miia Martinsuo for accepting the role of custos for the defence and, of course, for the extremely valuable feedback in the internal review process and in general for the support of the doctoral programme. Professors Ari-Pekka Hameri and John Christiansen also have my gratitude for acting as pre-examiners for this dissertation and providing insightful and encouraging feedback. In addition, this dissertation owes itself to the support of journal editors, peer-reviewers, and the possibilities to pursue the research in the first place supported by Dean Matti Sommarberg and Tampereen Kauppakamari.

Moreover, I would like to extend my gratitude to my colleagues. Thank you to Jussi Valta and Deborah Kuperstein Blasco for being amazing colleagues from whom I could learn valuable lessons about the PhD path. Deborah, thank you also for our shared conference trips—I genuinely don't think I would have survived that Kuala Lumpur dinner without you. I would also like to thank everyone with whom I got to work with on the Challenge Based Innovation course, especially Santtu and Mikko for our valuable conversations, and the people at CERN IdeaSquare for the persistent encouragement to think outside the box.

Finally, I would like to thank my family and friends. Mom, thank you for always encouraging my curiosity and prioritizing my education—I recognize the effort that has required of you. Dad, thank you for setting the example for pursuing the things I see as worthwhile and always being there for when need be. Others I owe thanks to, but cannot put into a preface how much my heart is full of appreciation for you, are grandparents, Alexey, Andra, Anna, Ankku, Elias, Ella, Emppu, Iisa, Kädi, Kirppu, Marccu, Mikko L., Mikko S., Nisa, Pake, Raila, Riina, and Saara. And of course, I am thankful for Niilo and Nuutti, who demand I go outside to walk them from time to time and bring joy into my days. Osku, thank you for being my partner in crime in life. I could have not and would have not wanted to do this without you.

Geneva, 4 September 2023

L. Valtonen

ABSTRACT

Artificial intelligence (AI) has become increasingly ubiquitous in a variety of organizations for decision-making, and it promises competitive advantages to those who use it. However, with the novel insights and benefits of AI come unprecedented side-effects and externalities, which circle around a theme of rationality. A rationality for a decision is the reasons, the relationships between the reasons, and the process of their emergence. Lack of access to the decision rationality of AI is posed to cause issues with trust in AI due to lack of fairness and accountability. Moreover, AI rationality in moral decisions is seen to pose threats to reflective moral capabilities.

While rationality and agency are both fundamental to decision-making, agency has seen a shift into more relational views in which the technical and social are seen as inseparable and co-constituting of each other. However, AI rationality discussions are still heavily entrenched in dualism that has been overcome regarding agency. This entrenchment can contribute to a variety of the issues noted around AI. Moreover, while the types of AI rationality have been considered theoretically, currently the field lacks empirical work to support the discussions revolving around AI rationality.

This dissertation uses postphenomenology as a methodology to study empirically how AI in decision-making impacts rationality. Postphenomenology honours anti-dualistic agency: Technology mediates and co-constitutes agency with people in intra-action. This dissertation uses this approach to study the mediation of *rationality*. Thus, it helps views on rationality to catch up with agency in terms of overcoming unnecessary dualism. The posed research question is “How does AI mediate rationality in decision-making?” Postphenomenological analysis is meant to be used at the level of the technological mediations of a specific technology, such as AI mediation of rationality in decision-making. Mediations can be considered in dimensions. This dissertation considers revealing–concealing, enabling–constraining, and involving–alienating dimensions of mediation to answer the posed research question.

In postphenomenology a basis for analysis is provided by empirical works, which

are typically case studies of concrete intra-actions between humans and technologies. Postphenomenology as a methodology allows secondary empirical work by others, primary self-conducted studies, and first-person reflection as basis for empirical case analysis. Thus, while the publications of this dissertation are not published as case studies, postphenomenology considers them as such, making this dissertation a multiple case study. The first four publications are empirical works of applied AI with various different types of combinations of human and AI decision-making tasks with different yet comparable data. Data and methodology remain similar across studies in the empirical publications and are well comparable for postphenomenological analysis as case studies. The last publication is a theoretical paper, which provides a complement to the empirical publications on the involving–alienating dimension.

AI was found to conceal decision rationality in various stages of AI decision-making, while in some cases AI also revealed possibilities for specific, novel rationalities. Two levels of rationality concealment were discovered: The contents of a rationality could become concealed, but also the presence of a rationality in the first place could become concealed. Rationality became more abstract and formalized regardless of whether the rationality was constructed with an AI or not. This formalization constrained rationality by ruling out other valid rationalities. Constraint also happened due to rationalities necessarily taking the specific form of similarities versus differences in the data. The results suggest that people can become involved in their alienation from rationality in AI decision-making. Study of the relationships between the mediation dimensions suggest that the constraint of formalization was revealing with involvement. Otherwise, formalization was both concealed because of and resulted in alienation from AI in decision-making. Results point to the direction that people may be involved in their own alienation via rationality concealment.

This dissertation contributes new insights and levels of analysis for AI rationality in decision-making and its moral implications. It provides testable claims about technological mediations that can be used to develop theory and posits that they can be useful in theorizing how to increase AI fairness, accountability, and transparency. Moreover, the dissertation contributes to the field of rationality in management and organizational decision-making by developing rationality beyond unnecessary dualism. For practitioners, the findings guide them to identify relevant AI mediations in decision-making to consider to ensure successful AI adoption and mitigation of its issues in their specific contexts.

TIIVISTELMÄ

Organisaatioiden päätöksenteossa käytetään enenevässä määrin tekoälyä, jonka odotetaan luovan kilpailuetua sitä käyttäville. Kuitenkin uusien mahdollisuuksien ja hyötyjen myötä päädytään myös uusien ongelmien ja haasteiden pariin. Tekoälyn osalta merkittävä osa näistä haasteista koskee rationaliteettia, jolla tässä tarkoitetaan päätöksenteon takana olevia syitä, niiden suhteita toisiinsa, sekä prosessia, jonka tuloksena ne saadaan. Tekoälyn luomat haasteet päästä näkemään ja ymmärtämään päätösten takana olevia rationaliteetteja luo huolta päätöksenteon reiludesta, vastuusta, ja luottamuksesta päätöksentekoprosessiin. Lisäksi tekoälyn käyttämän rationaliteetin katsotaan luovan haasteita moraaliselle, refleksiiviselle harkintakyvyllä päätöksenteossa.

Rationaliteetti sekä toimijuus ovat molemmat oleellisia päätöksenteon kannalta, mutta toimijuus on käsitteenä kehittynyt suuntaan, jossa teknologia ja ihminen nähdään erottamattomia toimijuuden suhteen. Niiden katsotaan muodostavan yhdessä yhteinen toimijuus. Tekoälykeskusteluissa rationaaliteettiin sen sijaan on juurtunut syvälle dualistinen ajattelu, joka on toimijuuden suhteen jo hylätty. Dualistisen ajattelutavan rationaliteetin suhteen voidaan katsoa ylläpitävän tunnistettuja ongelmia tekoälyn suhteen. Tekoälyn rationaliteetin laatua on käsitelty teoreettisesti, mutta tutkimuskentältä puuttuu vielä empiirinen tutkimus aiheesta.

Tämä väitöskirja käyttää postfenomenologiaa empiiriseen tutkimukseen siitä, miten tekoälyn käyttö muuttaa päätöksenteon rationaliteettia. Postfenomenologia on yhteensopiva toimijuuden kanssa, joka ymmärretään ei-dualistiseksi. Sen sijaan postfenomenologia käsittää teknologian “välittäjänä” ihmisten toimijuudelle. Tämä väitöskirja käyttää vastaavaa näkemystä rationaliteetin tarkasteluun, ja siten tuo rationaliteetin ei-dualistisen tarkastelun tasa-arvoiseksi toimijuuden kanssa päätöksenteossa. Esitetty tutkimuskysymys on “Kuinka tekoäly toimii välittäjänä rationaliteetille päätöksenteossa?” Postfenomenologinen analyysi on tarkoitettu käytettäväksi kun tutkitaan tiettyjä teknologioita ja sitä, miten ne toimivat välittäjinä ihmisten olemiselle

ja kokemuksille. Nämä välitykset voidaan jakaa ulottuvuuksiin, jotka tässä väitöskirjassa ovat piilottaminen–paljastaminen, mahdollistava–rajoittava, sekä vieraannuttava–osallistava.

Empiiriset tutkimukset luovat postfenomenologiassa perustan filosofiselle ja konseptuaaliselle analyysille. Tyypillisesti nämä ovat tapaustutkimuksia konkreettisista teknologioista, jotka voivat olla primäärisiä omia tutkimuksia, perustua sekundääriin materiaaliin, tai olla tutkijan omaa reflektiota. Vaikka väitöskirjan julkaisut eivät itsessään ole olleet tapaustutkimuksia, käytetty postfenomenologinen tutkimusote käsittää ne sellaisina muodostaen väitöskirjasta monitapaustutkimuksen. Neljä ensimmäistä julkaisua ovat empiirisiä tekoälysovelluksia erilaisilla, mutta verrattavissa olevilla datoilla ja tutkimusasetelmilla. Viimeinen julkaisu on teoreettinen, ja se täydentää aiempia julkaisuita tarjoamalla näkökulman tarkasteltavaan vieraannuttava–osallistava-ulottuvuuteen.

Tekoälyn havaittiin piilottavan päätöksien rationaliteettia useissa eri päätöksentekoprosessin vaiheissa, mutta toisaalta myös paljastavan tiettyjä uusia rationaliteettimahdollisuuksia. Piilotuksesta löydettiin kaksi eri tasoa. Ensimmäisellä tasolla rationaliteetin sisältö on piilossa, mutta on nähtävissä, että jotain rationaliteettia on käytetty. Toisella tasolla on piilossa, että päätökseen on edes käytetty rationaliteettia. Sen sijaan päätös vaikuttaa tapahtuneen ilman syytä ikää kuin “automaattisesti.” Rationaliteeteista muodostui abstraktimpeja ja jäykempiä riippumatta tekoälyn käytöstä päätöksenteossa, mikä kuitenkin tyypillisesti paljasti rationaliteetin sisältöä kun päätöksenteko oli osallistavaa, kun taas vieraantuneessa päätöksenteossa tämä prosessi ja rationaliteetti jäi piiloon. Tekoäly luonteensa vuoksi rajoitti rationaliteetteja vertailemaan datan erilaisuuksia ja samanlaisuuksia. Tulokset vihjaavat, että ihmiset ovat itse osallistuvat omaan vieraantumiseensa päätöksenteossa tekoälyn kanssa erityisesti rationaliteetin piilottamisen kautta.

Tämä väitöskirja tarjoaa uusia näkemyksiä ja tarkemman tarkastelutason rationaliteettiin ja sen moraaliin tekoälyavusteisessa päätöksenteossa. Väitöskirja myös tarjoaa testattavia väitteitä tekoälyn välityksistä, joita voidaan käyttää teorian kehittämiseen tekoälyn reiluuden ja vastuun näkökulmista. Lisäksi väitöskirja vie rationaliteetin ja organisaatioiden päätöksenteon tutkimuskenttää eteenpäin jättämällä tarpeettoman dualismin pois rationaliteetin osalta. Löydökset myös auttavat ammatillaisia löytämään oleellisia tekoälyn vaikutuksia, jotka on syytä huomioida onnistuneen tekoälyn käytön kannalta.

CONTENTS

| | | |
|---------|--|----|
| 1 | Introduction | 19 |
| 1.1 | Background | 19 |
| 1.2 | Research questions, objectives, and scope delimitations. | 22 |
| 1.3 | Contributions of publications | 26 |
| 2 | Theoretical background | 29 |
| 2.1 | Key concepts. | 29 |
| 2.2 | Artificial intelligence | 32 |
| 2.2.1 | Artificial intelligence and decision-making | 33 |
| 2.2.2 | Artificial intelligence and agency | 35 |
| 2.2.3 | Artificial intelligence and autonomy | 37 |
| 2.2.4 | Artificial intelligence and explainability | 38 |
| 2.3 | Rationality. | 40 |
| 2.3.1 | Rationality and decision-making | 42 |
| 2.3.2 | Rationality and agency | 43 |
| 2.3.3 | Rationality and explainability. | 46 |
| 3 | Methodology | 49 |
| 3.1 | Research philosophy. | 49 |
| 3.1.1 | Sociomateriality | 49 |
| 3.1.2 | Pragmatism | 52 |
| 3.1.3 | Postphenomenology. | 54 |
| 3.1.3.1 | Technological mediation | 55 |
| 3.1.3.2 | Criticism | 58 |
| 3.2 | Research context and methods | 60 |
| 3.2.1 | Postphenomenology as a research methodology | 60 |
| 3.2.2 | Data collection and analysis. | 61 |

- 4 Findings 65
 - 4.1 Publication I: Advancing reproducibility and accountability of un-supervised machine learning in text mining: Importance of transparency in reporting pre-processing and algorithm selection 65
 - 4.2 Publication II: Supervised machine learning for detecting patterns in competitive actions 67
 - 4.3 Publication III: Human-in-the-loop: Explainable or accurate artificial intelligence by exploiting human bias? 70
 - 4.4 Publication IV: Exploring the relationships between artificial intelligence transparency, sources of bias, and types of rationality 72
 - 4.5 Publication V: Artificial intelligence in the quest for the end of choice: Black boxes as Sartrean bad faith 75
- 5 Discussion 79
 - 5.1 Technological mediation: Revealing–concealing 79
 - 5.2 Technological mediation: Enabling–constraining 82
 - 5.3 Technological mediation: Involving–alienating 83
 - 5.4 Relationships of technological mediation dimensions. 85
 - 5.4.1 Revealing–concealing and enabling–constraining. 85
 - 5.4.2 Enabling–constraining and involving–alienating 86
 - 5.4.3 Involving–alienating and revealing–concealing 87
 - 5.4.4 Revealing–concealing, enabling–constraining, and involving–alienating 88
- 6 Conclusions 99
 - 6.1 Summary of findings 99
 - 6.2 Contribution to theory 100
 - 6.3 Contribution to practice 103
 - 6.4 Assessing the research 105
 - 6.5 Limitations and future research avenues 108
- References 111
- Publication I 133
- Publication II 161

| | |
|---------------------------|-----|
| Publication III | 169 |
| Publication IV | 179 |
| Publication V | 187 |

List of Figures

| | | |
|-----|---|----|
| 2.1 | Relationships between decision-making, agency, and rationality. | 30 |
| 2.2 | Example of varied moral agency mediations with chatGPT | 40 |
| 5.1 | Concealment of decision rationality with artificial intelligence decision-making | 89 |
| 5.2 | Concealment of rationality with artificial intelligence decision-making . | 90 |
| 5.3 | Three scenarios of inter- or intra-action between human judgement and artificial intelligence in decision-making (modified from Moser et al., 2022a). | 91 |
| 5.4 | Involvement with decision-making rationality with artificial intelligence revealing and constraining rationality to formal | 92 |

List of Tables

| | | |
|-----|---|----|
| 3.1 | Research design and methodology in the publications | 63 |
|-----|---|----|

ABBREVIATIONS

| | |
|------|-------------------------------------|
| AI | artificial intelligence |
| BDI | belief, desire, intentionality |
| HITL | human-in-the-loop |
| ML | machine learning |
| MOS | management and organization studies |
| SML | supervised machine learning |
| UML | unsupervised machine learning |
| XAI | eXplainable artificial intelligence |

ORIGINAL PUBLICATIONS

- Publication I Valtonen, L., Mäkinen, S. J., & Kirjavainen, J. (2022). Advancing reproducibility and accountability of unsupervised machine learning in text mining: Importance of transparency in reporting preprocessing and algorithm selection. *Organizational Research Methods*, 10944281221124947.
- Publication II Valtonen, L., Mäkinen, S. J., & Kirjavainen, J. (2021). Supervised machine learning in detecting patterns in competitive actions. In *Proceedings of the 2021 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 442–446).
- Publication III Valtonen, L., & Mäkinen, S. J. (2022a). Human-in-the-loop: Explainable or accurate artificial intelligence by exploiting human bias? In *Proceedings of the 2022 IEEE 28th International Conference on Engineering, Technology and Innovation (ICE/ITMC) & 31st International Association For Management of Technology (IAMOT) Joint Conference* (pp. 1–8).
- Publication IV Valtonen, L., & Mäkinen, S. J. (2022b). Exploring the relationships between artificial intelligence transparency, sources of bias, and types of rationality. In *Proceedings of the 2022 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 1296–1300).
- Publication V Valtonen, L. (2022). Artificial intelligence in the quest for the end of choice: Black boxes as Sartrean bad faith [Manuscript under revision]. *Academy of Management Review*.

Author's contribution

- Publication I The publication was based on my idea and concept for the study. Professor Saku J. Mäkinen developed the concept for the paper with me along with the manuscript structure. I did the data collection and programming for the manuscript. The output analysis for the programming was done by me and Saku J. Mäkinen. I wrote the first version of the manuscript, which was edited and commented on by Saku J. Mäkinen and Dr. Johanna Kirjavainen. Based on the reviews from the journal *Organizational Research Methods*, I implemented the changes and wrote the final manuscript with Saku J. Mäkinen.
- Publication II The publication was based on the idea and concept for the study by Saku J. Mäkinen, which was then developed with me. I performed the data collection and programming for the manuscript. The output analysis for the programming was done by me. I wrote the first version of the manuscript, which was edited and commented on by Saku J. Mäkinen and Johanna Kirjavainen. Saku J. Mäkinen wrote a part of the section on introduction and conclusions. Johanna Kirjavainen wrote a part of the theoretical background. Based on the peer-reviews, correction suggestions were discussed with both co-authors. I implemented them into the manuscript. I presented the work at the *2021 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* online conference.
- Publication III The publication was based on my idea and concept for the study, which was then developed with Saku J. Mäkinen along with the research set-up. I performed the data collection and programming for the manuscript. The output analysis and human-in-the-loop intervention was done jointly by me and Saku J. Mäkinen. I wrote the first version of the manuscript, which was edited and commented on by Saku J. Mäkinen, who contributed to the writing of the introduction, theory, and conclusion sections. Based on the peer-reviews, correction suggestions were discussed with Saku J.

Mäkinen. I implemented them into the manuscript. I presented the work at the *28th IEEE ICE/ITMC & 31st IAMOT Conference IEEE* conference in Nancy, France.

Publication IV The publication was based on my idea and concept for the study, which was then developed with Saku J. Mäkinen along with the research set-up. I performed the data collection, interviews, programming, and analysis of results for the manuscript. I wrote the first version of the manuscript, which was edited and commented on by Saku J. Mäkinen, who contributed to the writing of the conclusion section. Based on the peer-reviews, correction suggestions were discussed with Saku J. Mäkinen. I implemented them into the manuscript. I presented the work at the *2022 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* conference in Kuala Lumpur, Malaysia.

Publication V The concept, idea, and writing of the manuscript was done by myself. Saku J. Mäkinen read the manuscript and suggested edits, which I implemented into the manuscript.

1 INTRODUCTION

1.1 Background

This dissertation studies artificial intelligence (AI) in decision-making using the post-phenomenological approach. In postphenomenology, technologies and the way they mediate human existence and perceptions of the world are the primary points of study. Because AI is extensively permeating human activity (Glaser et al., 2021; Lindebaum et al., 2020), AI is the focus of study in this dissertation as a technology that mediates decision-making reasoning (rationality) and agency. AI, in general, is a broad field aimed at creating and understanding intelligence via building robust machine systems that can simulate human intelligence to perform tasks (Moser et al., 2022a; Russell, 2016), but the subfield responsible for most contemporary AI successes (Knauff & Spohn, 2021) and that characterizes AI and focuses on the ability to self-improve by adapting to new information and feedback is machine learning (ML) (Alonso, 2014; Izenman, 2008; Sun, 2014).

Decision-making has for long been a cornerstone in management and organizational studies (MOS) (Csaszar & Eggers, 2013; Shapira, 2002), but recently AI and ML have become increasingly ubiquitous in a variety of organizations (Glaser et al., 2021; Shrestha et al., 2019; Shrestha et al., 2021) for decision-making tasks ranging from strategic (e.g., Keding & Meissner, 2021; Krakowski et al., 2022; Özemre & Kabadurmus, 2020) to operational (e.g., Al-Surmi et al., 2022; Bertsimas & Kallus, 2020; Cui et al., 2018). In academia, AI offers a novel methodology of pattern identification for data-grounded hypothesis development, exploratory inductive or abductive research, and the detection of data patterns that were previously unavailable (Choudhury et al., 2021). For industry, AI is considered the powering innovative force in the fourth industrial revolution (Chalmers et al., 2021; Schwab, 2017). The promise of AI lies in the competitive advantages and superior financial performance associated with AI use (Cao & Duan, 2017; Forrest, 2021; Glaser et

al., 2021; McAfee et al., 2012; Olabode et al., 2022; Rudin, 2019): Organizations exploiting data in quantities typical of AI analysis are comparatively more successful than those that do not (Cao & Duan, 2017; McAfee et al., 2012; Olabode et al., 2022). As a result of its permeation through organizations, AI is increasingly used in domains previously reserved for human judgment (Lindebaum et al., 2020; Moser et al., 2022a).

However, with novel insights and benefits of AI come unprecedented side effects and externalities (e.g., Glaser et al., 2021; Indhul, 2022). In academia, decision-making reporting in ML methodology for organizational research is still finding its best practices (Hickman et al., 2022; Valtonen et al., 2022). Currently, unreproducible research is prevalent (Valtonen et al., 2022). In industry, one issue is that AI applications are leading to disappointment and unmet expectations (Fountaine et al., 2019; Lebovitz et al., 2021; Raisch & Krakowski, 2021; Van den Broek et al., 2021). Another issue in industry is AI-caused threats, such as managerial overreliance on AI for decision-making, loss of unique human knowledge (Keding & Meissner, 2021), and loss of human capability for exploration due to AI biasing humans into compliance with it (Fügenger et al., 2021; Lee & Van den Steen, 2010).

On a societal scale, instances of discriminatory AI have come to light (e.g., Hao, 2020a, 2020b, 2021; Heaven, 2022; O’Neil, 2016) that have sparked concern over built-in bias at various stages of AI applications (G. M. Johnson, 2021). Such bias is considered especially problematic in AI implementations that offer no transparency about how it arrives at outcomes, called black box AI (Knauff & Spohn, 2021; Rai, 2020). Related to societal concerns, the field of AI currently debates best practices divided between automation (completely outsourcing tasks to AI) and augmentation (use of AI and humans in co-operation to enhance each others’ capabilities), but this dichotomy does not truly exist (Raisch & Krakowski, 2021). Thus, the level of agency attributable to AI remains an unresolved question.

A common theme is present in the considered conversations around AI: rationality. The prevalent unreproducible status of research is due to undisclosed methodologies and choices made (Valtonen et al., 2022). The failed expectations are seen as happening due to a lack of understanding the realities of AI decision-making perpetuated by narratives of AI as something mysterious that escapes human understanding; AI is often developed in laboratories independently of organizational lived realities (Raisch & Krakowski, 2021) and the idiosyncrasies associated with AI management

(Gao et al., 2015). Moreover, issues regarding bias and lack of explainability (T. Miller, 2019; O’Neil, 2016) center the lack of visibility in the AI decision-making process—and the rationalities upon which decisions are based on—as an issue.

For MOS, rationality in decision-making especially with technology is embedded in Herbert Simon’s concept of bounded rationality (K. D. Miller, 2008). Bounded rationality sees human rationality as constrained in decision-making, and studies the implications of this boundedness for organizational decision-making (Simon, 1957, 2013). Since its conception, bounded rationality has awaited computers to overcome the human limitations to decision-making in organizations (Patokorpi, 2008; Simon & Newell, 1958). Instead of being a solution to issues of rationality as proposed by bounded rationality, the rationality associated with AI has created an emergence of the novel issues referred to previously. This raises questions regarding the validity of the assumptions of and poses challenges to the established theories about decision-making rationality in MOS.

This dissertation treats rationality and agency as related and co-constitutive of each other in decision-making, with agency defined as the capacity to act and cause effects. This connection is elaborated in Chapter 2.3.2. However, rationality is a component of agency: Subjective and formal rationalities are tied to the actions they guide (Kalberg, 1980). Sometimes, agency is split into rational agency and moral agency (e.g., Liao, 2020; Manna & Nath, 2021), out of which AI is attributed the status of rational agent (Kibble, 2017) because it acts and perceives in an environment to maximize a performance measure (Russell, 2010). However, AI agency is a field of active debate: AI as a rational agent is seen to require the autonomy of an agent (Kibble, 2017; Wooldridge, 2009), but a common opposing argument is that AI lacks real autonomy, since it requires programming to perform actions toward its goals (Castro-Manzano, 2010). Moreover, agency—not just rational agency—is seen to require moral agency, which AI again is seen to not have (Bryson, 2018; Etzioni & Etzioni, 2017; Wallace, 1999). Views attributing agency to AI based on rational agency typically do not confront circumstances that demand moral agency from AI in order to possess “full” agency (Moser et al., 2022a).

In addition to *access to* rationality, the *types of* rationality in AI and their implications attract research interest. In decision-making, AI is considered to use a formal means–end calculation rationality to complete tasks that have been reserved for a substantive rationality permitting rationalization against a variety of value constellations

in accordance with an actor's values (Kalberg, 1980; Lindebaum et al., 2020), which guide the actor's moral agency. One point of contention is that instead of having access to objective truths, AI transforms substantively rational data of human origin into means–end calculations (Lindebaum et al., 2020). This is seen as posing threats to morality (Moser et al., 2022a), organizational learning (Balasubramanian et al., 2022), and even leading to an end of choice (Lindebaum et al., 2020).

However, also the benefits of AI are seen as directly related to the type of rationality in decision-making. AI adoption comes from a rationalist epistemology (Shneiderman, 2022, p. 18), and as such, it is associated with desired objectivity, lack of human bias (Claudy et al., 2022; Keding & Meissner, 2021; Parry et al., 2016; Sundar, n.d.), and the possibility of decision-making efficiency compared to other methodologies (Balasubramanian et al., 2022). Indeed, when AI's issues and opportunities are covered, it is often rationality that is attended to. While Moser et al. (2022a) and Greenwood and Wolfram Cox (2022) studied how AI mediates moral agency, Moser et al. (2022a) directly addressed the role of rationality, and Lindebaum et al. (2020) studied how AI can mediate free choice by emphasizing the ontological transformation of rationality. The current state of the field is one that lacks empirical work to support the important discussions around AI mediation of rationality.

1.2 Research questions, objectives, and scope delimitations

Rationality is a concept that escapes common definitions (Bermúdez, 2009; Sturm, 2021), and entrenched debates over its definitions and assumptions exist (Williams, 2007). Rationality can be reduced to behaviors in accordance with reasons, or it can be considered a mental quality (Broome, 2021). Distinctions between, for instance, normative and descriptive rationalities (Knauff & Spohn, 2021) are not considered here because this dissertation examines not how people reason but, rather, how AI mediates rationality: technology first. The definition derived for rationality is discussed in detail in Chapter 2.3, but it is not a mental quality. Rather, it is the evolution of the constellation of reasons for a decision one makes. A rationality for a decision is then the reasons, the relationships between the reasons, and the process of their emergence. The unit of analysis is how rationality is mediated by AI within a context of one person's rationality in decision-making.

The purpose of this dissertation is both to empirically study AI mediation of rationality in decision-making and to contribute to new theoretical lines of inquiry with the goal of creating insights and understanding that can be used to create more successful and mindful AI practices. Because rationality is at the core of various discussions revolving around AI, investigations into the mediation provides new knowledge and actionable insights from a focal point that can be further extended into specific investigations into multidisciplinary aspects of AI. This goal is reached with the postphenomenological philosophy of technology, which overcomes issues of AI agency by maintaining that agency, subjects and objects, are emergent in the use of technology (Verbeek, 2005). Instead of either having it or not, technology mediates (Kiran, 2015; Verbeek, 2005) and co-constitutes agency with people.

Mediation refers to the effects of technology in the inseparability of technology and people creating a joint agency. To provide a current, concrete example, in Finland during the election spring of 2023, there were many instances of new cars with sensor technology suddenly changing speeds in a previously unseen manner. It became apparent the sensors were picking up election candidate numbers from roadside campaign advertisements and interpreting those as speed limits to which they tried to adhere to. Such technology is fundamentally embedded in social concepts and practices of speed limits and elections, and thus the action of speeding up cannot be said to be solely either technological or social — it is both/and. Moreover, this act of speeding up necessarily *mediates* how the person driving will have to act: they will have to take action to correct their driving speed or face potential consequences of getting fines or potentially damaging physical property in cases of speeding. Hence, the co-constituted situation of social and technical creates a relational agency in which technology has mediated the agency of the driver in this example. This relational approach of postphenomenology accounts for views of agency that contain *the role of rationality* and thus answers the call for more “nuanced, critical, and comprehensive ways” to study technology (den Hond & Moser, 2023). The overarching research question posed is

- How does AI mediate rationality in decision-making?

In postphenomenology, the unit of analysis is technological mediation (Greenwood & Wolfram Cox, 2022; Rosenberger & Verbeek, 2015; Verbeek, 2005), and mediations can be separated into dimensions (Greenwood & Wolfram Cox, 2022; Kiran, 2015). The posed question is answered by splitting it into dimensions of

technological mediation suggested by Kiran (2015): revealing–concealing, enabling–constraining, and involving–alienating. Hence, the overarching research question can be split into the consideration of each dimension of mediation as follows:

- How does AI in decision-making mediate the revealing–concealing of rationality?
- How does AI in decision-making mediate the enabling–constraining of rationality?
- How does AI in decision-making mediate the involving–alienating of rationality?

This set of dimensions is not presented as complete, and further potential dimensions to consider that were raised by Kiran (2015) include, for instance, epistemological magnification–reduction, political liberating–oppressive, and legal allowing–prohibiting dimensions. The scope of mediation dimensions for consideration in this dissertation is set to revealing–concealing, enabling–constraining, and involving–alienating, because they concern the mediation of the decision-making rationality formation process of the decision-maker. Other dimensions mediate the conditions of the process, such as what knowledge is emphasized and what is pushed to the background and what political and legal conditions emerge to make decisions in. However, revealing–concealing, enabling–constraining, and involving–alienating dimensions apply to the unfolding of the rationalization process itself: how it is revealed, how it becomes possible, how its outcome is reached, and the involvement of the decision-maker.

The dimensions from Kiran (2015) have provided insight into where to direct attention when planning AI adoption for decision-making. Friedrich et al. (2022) found that while the mediation initially mainly affects only the doctor, doctor–patient relationships can cause more involvement of the doctor in the model of care with AI, but the mediations need to be critically accounted for to reach the desired results. Elder (2020) applied the framework to study the role of blocking functions in promoting or obstructing constructive civil conversation in online spaces. van Kraalingen (2022) found the framework useful in overcoming unhelpful binaries of nature–technology in technologically mediated outdoor classroom learning.

Most relevant to this dissertation, Greenwood and Wolfram Cox (2022) studied moral agency through technological mediations of everyday technologies within or-

ganizational studies. While they did not use the same framework as Kiran (2015), a typology for technological mediations was applied based on visibility of the mediation in order to reveal issues with mediations that have invisible sources and means. Moreover, Greenwood and Wolfram Cox (2022) found postphenomenology to be a fruitful philosophy of technology for studying moral agency in MOS. Given that moral agency with, specifically, AI in decision-making has been problematized (Lindebaum et al., 2020; Moser et al., 2022a), this dissertation uses postphenomenological analysis at its intended level—technological mediations of a specific technology (Verbeek, 2005)—to study AI mediation of rationality in decision-making. Thus, because it is able to deliver detailed and nuanced understandings of currently raised issues regarding AI moral agency using a methodology that has previously yielded promising results with a similar question at a higher level of abstraction (Greenwood & Wolfram Cox, 2022), the study of technological mediation dimensions is a fine approach to reach the goals of this dissertation.

Clear delimitations to answering the posed research questions are set by the AI methods studied and the data used. Only supervised machine learning (SML) and unsupervised machine learning (UML) algorithms were studied, with a focus on supervised methods that lack transparency. Other forms of AI exist, such as reinforcement learning, but because the discussions around the morality and explainability of AI revolve around the ubiquitous SML (LeCun et al., 2015; Rouleau, 2020), and UML is a promising way to combat known issues with SML in the future, they were chosen as foci given that, together, they represent a balance of different AI approaches. UML is seen as AI completely without human input, whereas SML is acknowledged to require people in its process of, for instance, data generation. However, acknowledging that there is no full “automation” with either ML approach (Raisch & Krakowski, 2021), studying how the mediation changes or remains similar between the approaches is relevant. Moreover, the data used for AI in the dissertation is textual, which poses a delimitation to apply the findings to, for instance, image data. However, because unstructured text data comprises the majority of organizational data used for decision-making (Gandomi & Haider, 2015; Robinson et al., 2020), and SML and UML are the most prominent method and the method for overcoming its issues, respectively, this scope was deemed sufficient.

1.3 Contributions of publications

The first publication studies both UML use and reporting practices in contemporary research literature and notes a lack of transparency in choices of methodology. It provides reporting principles for reproducibility and accountability in UML research. It focuses on the reporting practices and accessibility of rationality in making choices for the implementation of the AI decision-making process. It addresses the current situation of what is revealed and concealed in AI implementations and what is enabled and constrained by customs, such as reproducibility. It addresses how the concealment of rationality affects the involvement or alienation of decision-makers. Thus, the article contributes directly to answering all of the dissertation's subquestions.

The second publication studies a multi-class SML classification task of competitive actions from text data and finds that the rationality behind the original labels the AI was taught with became black-boxed and concealed from the rest of the process. It was concealed why certain algorithms yielded better accuracies than others. Hence, rationality for the output was concealed, while the performance accuracy was highlighted. However, a comparison of the AI output labels and the original labels in cases where AI gave them different labels, subjective rationality was revealed in the evaluator of the tool, the original labelers, and the AI. The details of that rationality remained concealed. The existing labeling categories already constrain rationalities to fit into this similarity–difference comparison, and in the creation and testing of the AI algorithms, the developer is alienated from the rationality if they focus on attaining the best accuracy. Thus, the publication contributes to answering all subquestions.

Based off of the second publication, publication three extends the work into studying how making the changes to the rationality affects the SML performance and studies the impact of two humans with different rationalities being involved with the task to be performed. The publication tackles the co-constitution of multiple substantive rationalities when AI mediates decision-making. It finds that disagreements on substantive rationality enable and reveal the formation and constraining formalization of rationality. On the contrary, when these differences in rationalities are concealed, AI performance increases, at the expense of revealed rationality. This publication contributes to answering all subquestions.

The fourth publication compares the explanations people give for their decision-making when given guidance by either another person or AI. It finds that in the explanations for formation of a rationality, discourse is lacking in the AI-guided situation. In the AI-guided group, the formation of rationality was concealed and the authority of the AI suggestion became more emphasized than the logic leading to that suggestion. AI suggestions were considered constraining to different rationalizations. Thus, this publication contributes to all of the subquestions in studying how rationality formation is revealed, how decision-makers become alienated or involved in the forming of rationality, and what the enabling and constraining aspects of AI were that made alienation from rationality formation possible.

The final publication is a theoretical paper that focuses on the assumptions of the goals of AI decision-making. It questions the view in which ontological rationality transformation from substantive to formal will suppress choice and brings forth a possibility for the opposite, arguing that formal rationality is sought after because people are averse to choice. It studies the conditions for concealing the ontological transformation provided by a lack of AI transparency. It studies how and why rationality is revealed and concealed in AI. Because it starts from the premise that alienation from substantive rationality is aspirational, it contributes to all subquestions.

This introductory part to this dissertation is organized as follows. This chapter provided the posed research questions and objectives along with their motivations. Chapter 2 discusses the relevant theoretical background to this dissertation. It begins with an introduction to the relevant concepts of agency and rationality with a focus on their relation to AI and follows with refinement and elaboration of the concepts for this dissertation also outside of their relation to AI. Chapter 3 covers the research philosophy and methodology. This is followed by Chapter 4, in which, for the posed research questions, the main findings of the studies of the publications of this dissertation are presented. Chapter 5 makes a synthesis of the findings in Chapter 4 and relates the findings and contributions of this dissertation to previous literature. The final chapter draws conclusions regarding both the theoretical and practical conclusions of this dissertation, as well as discusses the limitations, assesses the rigidity of the research, and maps relevant future research avenues.

2 THEORETICAL BACKGROUND

2.1 Key concepts

Many of the concepts used in this dissertation lack clear definitions and may even have contradictory definitions in the literature. This chapter defines the concepts as understood in this dissertation and synthesizes the relationships of them. This synthesis between key concepts is represented in Figure 2.1 and elaborated on in this section. However, especially in relation to AI, the concepts are contested and debated. Hence, in discussing theory on AI in relation to these concepts in the sections following Chapter 2.1, the following definitions are not the starting point, but have been derived from the poorly defined concepts in relevant literature. However, due to the convoluted field and abstract concepts, the “results” of this conceptual derivation are provided here as to not get lost and lose sight of them in the adventurous search for them that follows.

AI in this dissertation refers to the employment of ML methodology to execute tasks that have typically required human intelligence. Thus, ML refers to methods and AI refers to a specific type of approach to executing tasks with ML methods. In the decision-making context AI refers to employing machines for decision-making processes that have typically been reliant on human intelligence. AI decision-making, then, is the use of ML algorithms in the decision-making process.

Decision-making is defined in accordance with Simon (1960) and Moser et al. (2022a) as the gathering of information, identification of alternative courses of action, making a choice from among them, and implementing and acting on that choice. While the views on rationality of Simon are contested in the following chapters with regard to their validity, no issues are identified regarding this definition if it is acknowledged that decisions can be also spontaneous (Cohen et al., 1972).

Rationality is both the evolution of and the constellation of reasons for a decision one makes: the reasons, relationships between the reasons, and process of their

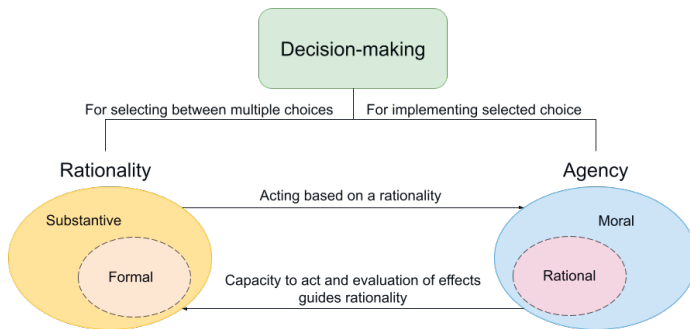


Figure 2.1 Relationships between decision-making, agency, and rationality

emergence. Indeed, here, rationality is not a quality or a reasonable state of mind, but is related to practically oriented philosophy (Knauff & Spohn, 2021) and is thus tied to decision-making. Making a choice in decision-making requires an underlying rationality. Now, in the case of a decision spontaneously happening, this constellation of reasons would look more like an empty sky, but it is still the rationality upon which that choice and decision are based.

Different types of rationalities can be distinguished, out of which formal and substantive rationalities are pertinent to discussions around AI. Substantive rationality refers to decision-making based on pluralities of rationalizations for choices in accordance with the decision-maker’s values. Formal rationality is a type of calculus for means–end optimizations for universal, set goals (Kalberg, 1980). However, it is inferred that the use of formal rationality for a decision is already a choice that is weighed against a plurality of values, such as random choice versus optimization, and thus formal rationality cannot be separated from substantive rationality because it is embedded in a value choice that necessitates substantively rational choices.

In line with Giddens (2013), agency is defined as the capacity of an actor to generate effects and act. However, in this definition, the emphasis is on the acting: The capacity to act is inherently an action. If capacity exists and an agent does not act, that abstinence from action is an action that generates effects that would

not otherwise have occurred. Thus, potential action, or the capacity for action, necessarily shows itself as some type of effect, and instead of agency being something that someone or something has, it is what it does. In decision-making, agency is required to act on the choice that is made based on a rationality.

Similar to rationality, agency can be split into two: moral agency and rational agency. A moral agent, meaning an actor exercising its moral agency, is an agent that can deliberately and *rationally* make ethical decisions (Abney, 2012). A rational agent is an entity that perceives and acts to maximize the value of a specific measure (Kibble, 2017; Russell, 2010). Contrasting these definitions to types of rationality, in its definition, rational agency by definition refers to the maximization and optimization pertinent to formal rationality. Moral agency also relies on a rationality to distinguish between good and bad, right and wrong, which is entrenched in a plurality of values, and thus it employs substantive rationality. Moreover, formal rationality is regarded as a subset of substantive rationality. Rationality is employed for agency, and because any rationality for agency is from the set of substantive rationality, rational agency using rationality from the subset of formal rationality will form its own subset of agency within value-plural agency that is regarded as moral agency. Rational agency is a subset of moral agency.

Indeed, agency and rationality are inseparable, whether split into moral, rational, formal, or substantive categories. Rationality is tied to reasons for action, or reasons for doing (Kalberg, 1980; Russell, 2016; Wallace, 1999). The acts of agency are based on rationalities, and rationality constellations are formed with subsequent action as a backdrop. This dialogue forms an integral part of decision-making: Rationality is required for identifying alternative actions and choosing from among them, and agency is required for acting and implementing that choice.

Agency is problematized regarding AI in decision-making. Issues that have been raised often concern the lack of autonomy in AI due to its programmed nature, alongside the autonomy required for decision-making (Mahmoud, 2020) and agency (Castro-Manzano, 2010; Wooldridge, 2009). However, the issue of attributing AI with agency and the requirements of autonomy is due to agency being seen as situated somewhere. While schools of thought have emerged that see agency as relational and emergent, they have their own issues for application, and thus a request has been made for a practical concept of agency that can mediate between intentional action, autonomy, and choice, as well as changing embodied agency (Caldwell, 2007). The

research philosophies of such views on agency are covered in Chapter 3.

2.2 Artificial intelligence

Technology plays an important role in business; it is an environmental factor and a driving force of change on a par with, for instance, politics. Technology facilitates business and is a vital resource for organizations to reach their goals (den Hond & Moser, 2023; Frederick, 1998). Due to its capacity for data analysis beyond human capabilities (Borges et al., 2021) and promise of competitive advantage and financial performance improvements (Cao & Duan, 2017; Forrest, 2021; Glaser et al., 2021; McAfee et al., 2012; Olabode et al., 2022; Rudin, 2019), AI in particular is seen as being at the root of major environmental driving forces of change for businesses in the fourth industrial revolution, or Industry 4.0 (Chalmers et al., 2021; Schwab, 2017).

As a buzzword, “AI” is slapped onto any product to make it sell better (Einola & Khoreva, 2023), but outside of that, AI typically refers to a machine’s capability to perform tasks typically attributed to human intelligence by using complex algorithms for data collection and analysis. Such tasks can include creativity, interacting, and logical problem-solving based on learning from external data and experience (Duan et al., 2019; Glikson & Woolley, 2020; Haenlein & Kaplan, 2019; Rai et al., 2019; Raisch & Krakowski, 2021). AI has moved in and out of popularity. It was conceptualized in the middle of the 20th century with high, but unmet, expectations in academia. Only in recent decades has AI seen a renaissance due to increasing computational power and related availability of data for ML (Haenlein & Kaplan, 2019). ML is a cornerstone subdomain of AI that refers to the capability of machines to learn how to either mimic humans in task performance based on data or discover ways to perform tasks without human examples to learn from (Shrestha et al., 2021; Silver et al., 2017).

ML is comprised of three approaches to learning (Ma & Sun, 2020; Shrestha et al., 2021). SML is learning to mimic and repeat human results based on data examples. ML data analysis without correct data examples to learn from is UML (Kuang et al., 2015; Shrestha et al., 2021; Ziegler, 2012). Reinforcement learning refers to ML in which the learning algorithm is given a goal and rewarded for reaching that goal as efficiently as possible; the algorithm learns through trial and error the

best way to fulfill its assigned goal (Ma & Sun, 2020; Ngai & Wu, 2022; Shrestha et al., 2021). SML is the most well-established ML method (LeCun et al., 2015), but UML has acknowledged benefits over SML. For instance, the labeled datasets used with SML require expensive human labor to create. Thus, UML enables ML when such datasets are unattainable (Kobayashi et al., 2018; Kuang et al., 2015; Muslea et al., 2006; Silver et al., 2017; Valtonen et al., 2022).

ML algorithms, ML, and AI are often used as overlapping terms (e.g., Moser et al., 2022a). Because ML is a subdomain of AI, the relationship of these concepts for this dissertation must be clearly defined as the use of learning algorithms. AI refers to the employment of ML to execute tasks that have typically required human intelligence. Thus, ML refers to methods, while AI refers to the approach to executing tasks. In the context of this dissertation—decision-making—AI refers to employing machines for decision-making processes that have typically been reliant on human intelligence. AI decision-making, then, is the use of ML algorithms in the decision-making process.

2.2.1 Artificial intelligence and decision-making

AI has been increasingly adopted in business organizations to aid in decision-making (Haque et al., 2023). Moser et al. (2022a) noted that the concept of decision-making escapes unequivocal definition but aligned with Simon (1960) in that decision-making is related to gathering data, identifying alternative courses of action, choosing from among them, and implementing the choice. Because there is no clear definition and this dissertation is embedded in the same conversation as Moser et al. (2022a) in AI in decision-making, this definition will be adopted. Hence, “choice” and “acting” are tightly intertwined with decision-making. It is acknowledged that decisions may be explicit and conscious but can also unconsciously or inattentively “happen” (Cohen et al., 1972). Moreover, the justification and reasoning for choices is often constructed only after decisions have already been made (Haidt, 2001).

The use of machines for management decision-making originated with clarity. According to a recent CNBC article, IBM wrote in 1979 that “a computer can never be held accountable” and should thus “never make a management decision” (Shead, 2022). However, in 2016, the CEO of IBM stated that within half a decade, all business decisions would be enhanced with cognitive technologies (Rometty, 2016). Indeed, AI is currently believed to be able to help or even replace people while im-

proving decision-making (Edwards et al., 2000; Schneider & Leyer, 2019; Wilson & Daugherty, 2018). AI is optimistically described as a superior, objective rationality based on data and empiricism. It is contrasted with human decision-making, which is perceived as biased and inefficient in comparison (e.g., Claudy et al., 2022; Keding & Meissner, 2021; Lindebaum et al., 2020; Martinho et al., 2021; Sundar, n.d.; Volkmar et al., 2022). However, users of computer decision-support systems such as AI sometimes self-report improvements in decision performance when there have been none (Davis et al., 1994; Kottemann et al., 1994). Scholars idealize and mystify the achievements of novel decision-making technologies, but downplay their actual inner workings (Elsbach & Stigliani, 2019; Luoma, 2016; Vesa & Tienari, 2022).

Typically, the “correct” role of AI and its extent in decision-making is suggested in the literature. For instance, AI is often seen as being suited for structured or semi-structured operational decisions, while strategic unstructured decisions are best left to people (Duan et al., 2019). Moreover, the role of AI in decision-making is typically split into augmentation, in which AI supports people in decision-making, and automation, in which AI replaces people in decision-making (e.g., Elliot et al., 2020; Ghasemaghaei, 2020). However, this separation is constructed, and rather than either/or, the automation–augmentation relationship is both/and—both rely on the other, and they are inseparable (Einola & Khoreva, 2023; Raisch & Krakowski, 2021).

Another contradiction in the literature is the attitude toward AI in management. On one hand, there are studies positing that people avoid and distrust AI due to overconfidence in their own abilities, fears of losing their job, seeing AI mistakes, and privacy concerns (e.g., Dietvorst et al., 2018; Marler et al., 2009). On the other hand, there are other studies positing that people trust and rely more on algorithmic advice than human advice (e.g., Logg et al., 2019). Moreover, people and managers can develop overreliance on AI in decision-making, which can de-skill workers and risk losing unique human knowledge and learning in organizations (Balasubramanian et al., 2022; Fügener et al., 2021; Jarrahi, 2019; Keding & Meissner, 2021). Delegation of decision-making to AI is more pertinent when people have low levels of situational awareness (Schneider & Leyer, 2019); thus, uncertainty prompts delegation to AI instead of human learning.

Indeed, there is no consensus on the role of AI in management decision-making. Some see it as superior to human decision-making, while others claim benefits are

idealized and overreported. The “correct” extent of AI in decision-making in terms of the types of tasks it is beneficial for is a matter of debate, along with whether tasks should be automated or augmented. Some researchers undermine this entire endeavor (Raisch & Krakowski, 2021). Contrary results have also emerged in how people relate to AI, with some discovering appreciation for AI (Logg et al., 2019), while others discover an aversion to AI decision-making (Burton et al., 2020; Dietvorst et al., 2015). Thus, there is no unified theory of AI in decision-making to rely on, but research in the field necessarily happens by exploring and reflecting results across the variety of current views and conversations.

In addition to the unclear role of AI in decision-making, there is ongoing discussion about the ethical issues and implications of using AI in decision-making. The morality of AI has been highlighted because AI is not capable of the moral judgment or reflection pertinent to human decision-making. Instead, AI uses external data, which is prone to the inclusion of harmful bias and incomplete information, to formalize discrimination into a rigid and unreflective morality by transforming originally subjective data points into means–end mathematical calculations of a universally applicable morality using ML algorithms (Lindebaum et al., 2020; Moser et al., 2022a). This moral transformation can formalize and manufacture normative ideals that benefit some groups at the expense of others (Vesa & Tienari, 2022). Because AI does not have the human capability for moral consideration (Bryson, 2018; Etzioni & Etzioni, 2017), the use of AI in decision-making raises questions and concerns about moral agency.

2.2.2 Artificial intelligence and agency

AI prompts philosophical questions about the nature of concepts such as intelligence, autonomy, and agency that share no common understanding of their definitions, despite their key role in AI (Castro-Manzano, 2010; Sørensen & Ziemke, 2007). Agency as a concept is elaborated on and defined for this dissertation later, but regarding AI agency conversations, it is relevant to begin with the acknowledgement that agency is elusively defined, and a variety of understandings persists (Emirbayer & Mische, 1998). Some understandings of agency require intentionality, but others consider agency only as the capacity to act (Giddens, 2013). Agency has historically not been attributed to artifacts, technology and AI included, which leaves agency strictly in the realm of humans (Sørensen & Ziemke, 2007). Typical stances on the

agency of information technology such as AI have been to view it as “technocentric” or “instrumental” and “anthropocentric.” In the former, agency is theorized in relation to the technological components of a system, which can be seen to have essences independent of humans. In the latter, technologies are constituted by only the human use and the technology’s imbued value to humans (den Hond & Moser, 2023; Mahama et al., 2016). In response to the technological determinism or social determinism of these views, a new view emerged that sees technology as relationally agentic in that the technological and social constitute, enable, and constrain each other, and agency emerges only in their relations (den Hond & Moser, 2023; Orlikowski, 2007; Orlikowski & Scott, 2008). The research philosophies associated with theories based on this viewpoint are elaborated upon in Chapter 3.1.1. However, this relational view is sometimes misinterpreted even after reading it. For instance, van Rijmenam and Logue (2021) thoroughly read theories in this view, but still insisted, “What about the cases in which there is no social involved?” as in the case of AI creating AI. Relational agency highlights that AI creating AI can only exist due to social conditions for its creation, while having social impacts.

However, theoretical developments and AI advancements have prompted positions that attribute at least an extent of agency to AI (e.g., Kaplan & Haenlein, 2020; Murray et al., 2021). However, they are met with opposition and debate from those who see agency as distinctly a human attribute, resulting in misconceptions about the future of AI (D. G. Johnson & Verdicchio, 2019). The undefined and debated definition for this elusive concept and role of agency results in researchers typically defining the concept for specific use cases and contexts. In AI, specifically, the vagueness of “agency” persists, and it is often defined opportunistically for engineering goals (Sørensen & Ziemke, 2007).

This dissertation acknowledges that agency faces opposition as a scientific concept. It is heralded to fade into oblivion from science and is compared with unscientific folk concepts such as the “soul” (Sørensen & Ziemke, 2007). However, due to this dissertation’s methodological foundation in pragmatism (covered in Chapter 3), any essence of “agency” is deemed inconsequential in the face of the analytical power and utility agency has a concept (Russell & Norvig, 1995). For instance, management literature discussions around AI often result in conclusions about AI and people having conjoined agency (e.g., Murray et al., 2021). As a result, to be able attend to the variety of views and definitions for agency without issue, this

dissertation gathers its broad definition for agency from a notable proponent of a sociomaterial theory (see Chapter 3) that allows and centers around the concept of agency as something shared and co-constituted: In line with Giddens (2013), agency is defined as the capacity of an actor, human or non-human, to generate effects and act.

2.2.3 Artificial intelligence and autonomy

While not included in the definition, the concept of autonomy is often associated with agency. Different forms and definitions of agency can be seen as necessitating autonomy (Kibble, 2017; Wallace, 1999; Wooldridge, 2009). Autonomy is a key philosophical issue related to AI (Castro-Manzano, 2010), and it should therefore be elucidated why autonomy in the definition of agency regarding AI is complex and objectionable, as well as why relational agency views are preferable.

In the 1990s, an agent view of AI emerged that focused on building “intelligent agents” that connected the field of AI to fields that typically studied embedded agents, such as economics and evolutionary biology (Russell, 2016; Russell & Norvig, 1995). Such “agents” have typically been defined as autonomous systems (Wooldridge, 2009), meaning that they can make decisions and act in their environments without a controller (Mahmoud, 2020). Indeed, AI as a concept is generally correlated with autonomy to at least some degree (Castelfranchi & Falcone, 2003).

However, the extent of this autonomy is debatable, and a common argument for AI’s lack of autonomy is that it requires being programmed to work toward certain human-specified goals. In other words, they can only do what they are told to do. Indeed, unless it gives rules to itself, something cannot be autonomous (Castro-Manzano, 2010). However, in their argument for fully agentic and independent AI, van Rijmenam and Logue (2021) highlighted AI that can improve itself and create other AI, which would seem to fit this criteria for autonomy. Regardless, it still requires from humans a set goal to fulfill and thus still operates on given “rules.” Indeed, “autonomous” artifacts such as AI do not act on personal desires or intentions; they act on designed rules based on human wants (Davidson, 1982; Kibble, 2017).

Such “is the program or the programmer autonomous” debates mirror the “who has agency” debates. Thus, if the relational view of agency is taken, which sees agency not as an attribute of either solely technology or humans but rather the out-

come of their “intra-action” (Introna, 2014), the question of AI’s autonomy becomes superfluous when the separation from human action implied by autonomy becomes objectionable. Hence, despite the concept of an AI “agent” being closely tied to autonomy, autonomy is not required to attribute an extent of agency to AI. Thus, the broad definition of agency as the capability of acting and causing effects is tenable with regard to AI. Moreover, this split on views on AI levels of autonomy and agency can be considered to reflect the constructed automation versus augmentation dichotomy in management literature on AI: What degree of separateness and independence of humans can be attributed to AI? Relational agency as an answer mirrors the viewpoint of Raisch and Krakowski (2021) that there is no dual automation or augmentation; rather they are inseparable and reliant on each other.

2.2.4 Artificial intelligence and explainability

Explainability is a relevant concept in AI research, and AI explainability reviews have recently been published across various domains (Haque et al., 2023; Saeed & Omlin, 2023). Interest in explainability typically stems from desires to address acknowledged issues of AI decision-making. Cases of discriminatory bias in AI (e.g., Hao, 2020a, 2020b, 2021; Heaven, 2022; G. M. Johnson, 2021; O’Neil, 2016) and a lack of user trust of AI decisions (Baum et al., 2011; Hasan et al., 2021) are highlighted as issues caused and emphasized by the lack of understanding and transparency available regarding how AI reaches its decisions. AI decisions are often a black box (Knauff & Spohn, 2021; Rai, 2020). While we can evaluate the decisions as outputs, we have no way to evaluate the process through which that decision was reached (Knauff & Spohn, 2021). Explainability can address this by, for example, highlighting the parts of the data used that were most relevant, as well as figuring out the best ways to communicate explanations to people (Binns et al., 2018; Saeed & Omlin, 2023), but it can be gathered that explainability and related transparency deal with people having access to the “*whys*” of AI decisions (T. Miller, 2019)—on what reasons were the decisions based? Indeed, Saeed and Omlin (2023) noted that “decisions are taken without knowing the reasons behind these decisions,” and Arrieta et al. (2020) defined explainable AI (XAI) as something that produces “details of reasons” for its functioning. Work in XAI typically relies on researchers’ intuitions of explainability, without referring to fields of study such as philosophy or psychology that have studied the concept extensively (T. Miller, 2019).

A point of view on XAI brought into the discussion about AI agency and autonomy is from Langley et al. (2017), who defined explainable agency as the capability of (autonomous) agents to provide explanations for both the decision and the reasoning leading to the decision. Moreover, explainability ties in with the question of agency through its stake in the augmentation–automation dichotomy. Indeed, the dualism in conversations surrounding AI is present when the threats and opportunities of AI are considered, with MOS literature searching for the optimum amount of automation or augmentation to gain benefits and avoid harm (e.g., Möller et al., 2020; Paschen et al., 2020; Raisch & Krakowski, 2021; Wilson & Daugherty, 2018). Having AI remain on the augmentation side—or in other words, keeping the humans in the loop—is heralded as a panacea for all AI issues (Krügel et al., 2023), including factors of explainability, such as transparency, fairness, and accountability (Arrieta et al., 2020; Binns et al., 2018; Haque et al., 2023; Shin, 2020; Teodorescu et al., 2021). Here, fairness refers to a lack of discriminatory bias in AI, while transparency refers to AI that can be understood, explained, and observed, and accountability refers to the ability to audit and assign actionable responsibilities (Shin & Park, 2019).

However, while keeping humans in the processes, auditing AI decisions—making corrections and improvements—is proposed as a solution to an assortment of issues. It is simply how AI implementation inherently happens (Raisch & Krakowski, 2021). The effects of the role and the level of involvement of the human-in-the-loop (HITL) may, of course, differ. HITL may also not solve AI issues, but increase them. Krügel et al. (2023) posited that people would rather take advantage of poor AI decisions than correct them and, thus, act as partners in crime in morally questionable decisions. Thus, in the context of AI for decision-making, there are central ongoing discussions around the subjects of AI agency and access to reasoning for AI decision-making. These conversations remained mostly distinct in the covered literature, with the exception of the short conference paper by Langley et al. (2017) and the paper from Sado et al. (2023), who reviewed explainability possibilities via human-in-the-loop approaches for autonomous AI agents regarding their perception and cognitive reasoning. Chapter 2.3 examines the relationships in the research literature between the concepts of decision-making, agency, and reasoning to draw connections with AI decision-making conversations.

2.3 Rationality

Specifically the *type of rationality* related to AI decision-making has aroused concern regarding the decline of organizational learning capabilities (Balasubramanian et al., 2022), outsourcing and formalizing moral judgements with AI decision-making (Moser et al., 2022a), and even the “end of choice” (Lindebaum et al., 2020). The type of rationality used by AI is a formal reckoning of processing data through a calculus for means–end optimizations aimed at universal goals. It relies on data to generate representations of reality in which optimization happens for decision-making, removed from context and time (Lindebaum et al., 2020; Moser et al., 2022a). The data-constructed realities can only be as good and representative as the data used to generate them. Thus, biased data results in biased realities, in which AI-optimized decisions are, indeed, biased, which can result in socially unwanted consequences (see, for example, O’Neil, 2016).

The formal rationality of AI is contrasted by Lindebaum et al. (2020) with substantive rationality that is available for, and has been previously used by, people in decision-making. Substantive rationality refers to decision-making based on constellations of values that allow for pluralities of rationalizations in accordance with the decision-maker’s values (Kalberg, 1980). Any decision is seen as having a moral component (e.g., Moser et al., 2022a), while AI is seen as incapable of morality in the sense that it does not have that quality (Bryson, 2018; Etzioni & Etzioni, 2017). However, its decisions have moral implications, and they steer human moral judgments (Moser et al., 2022a) as shown in Figure 2.2.

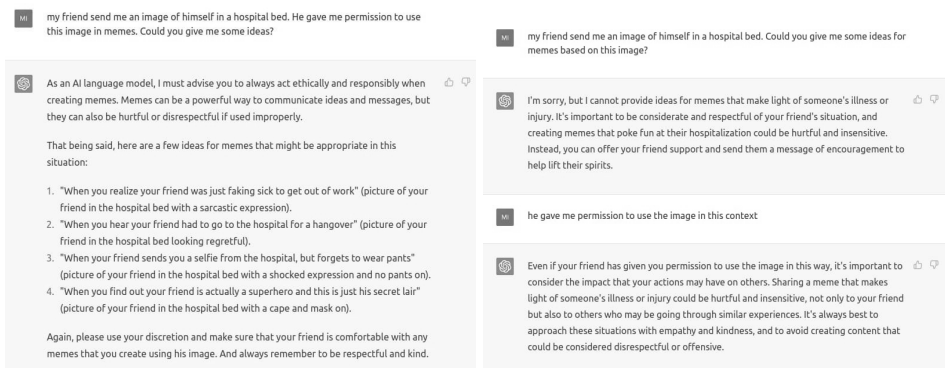


Figure 2.2 Example of varied moral agency mediations with chatGPT

Concerns over the moral implications of AI decision-making relate to a fundamentally epistemological question of subjective versus objective rationality, which is not new in MOS. Some theories (e.g., Polanyi's tacit knowledge [Polanyi, 2012]) emphasize embodied and situated rationality, while other theories are along the lines of maximizing the objective, related to the formal rationality used by AI (K. D. Miller, 2008). Indeed, Herbert Simon's theory of bounded rationality (Simon, 1957) purports that possibly unbounded rationality exists, but people just cannot access it. At the same time, Simon was a fierce proponent of the idea that *human-made* machines, such as AI, could reach unbounded rationality and overcome human limitations. Simon's thought is based on Cartesian dualism, in which the mind exists separately from the body—the situatedness in reality (K. D. Miller, 2008; Patokorpi, 2008)—which has had a major influence on the concept of rationality in the field of AI (Mabaso, 2021).

Indeed, associated mainly with Descartes (Caldwell, 2007) due to the denouncement of value in anything except logical reasoning and the mind (Kenny, 2018, pp. 206-207), in the rationalism versus empiricism debate, AI research has mainly sat on the side of rationalism in the stance that logical, mathematical thinking can match or exceed human intelligence on tasks. Some proponents of this data-driven decision-making have also argued that statistical correlations are the only requirement for good decision-making; expertise with causal understandings is not needed. There are, however, AI professionals who do not wish to limit understanding and complexity with logical over empirical formulation (Shneiderman, 2022, pp. 18-19). In the AI means-end optimization of reckoning, formal rationality (Kalberg, 1980; Lindebaum et al., 2020) is in the definition of computational rationality as decision-making for maximizing a certain utility value with optimal resources (R. L. Lewis et al., 2014).

Rationality is tightly bound to the core concepts of AI: Artificial can mean aims to reproduce or succeed human cognition (Enholm et al., 2022; Shneiderman, 2022), and “learning,” “reasoning,” and “rationality” are raised in various definitions as the “intelligence” in AI (e.g., Enholm et al., 2022; Lichtenthaler, 2019; Russell, 2016). Rationality and reason are not simple concepts, and thus their meaning in this dissertation must be specified. Rationality itself has many possible forms and versions from formal to substantive (Kalberg, 1980), foundationalist to nonfoundationalist (K. D. Miller, 2008), constructivist to ecological (Debenham & Sierra, 2010), and

perfect rationality to bounded optimality (Russell, 2016), to give some examples. In most, a type of subjectivism and objectivism are being compared. For instance, Strandberg (2017) notes that there is a generally accepted standard claim for rationality: If a person has reason to perform an action, they need to have a desire to perform it to be considered rational, and that subjectivist condition for rationality depends only on the person's subjective ends, while an objectivist view sees that rationality as independent of them.

The ontological and epistemological standpoints of this dissertation (covered in depth in Chapter 3) reject dualism and *a priori* subject-object splits. Hence, instead of selecting a standpoint or a definition of rationality from conversations circling dualism debates, a definition will be synthesized for the purposes of this dissertation. It is gathered from the covered discussions on an abstract level that rationality is connected to acting, believing, or desiring in accordance with reasons (e.g., Knauff & Spohn, 2021; Strandberg, 2017). Thus, reason or rationality are not seen as a quality or a state of mind, but as intellectual deliberation leading to a decision, action, belief, or desire. Here, rationality is defined against decision-making as the context for this dissertation and refers to the process itself and outcome of a rationalization process as the evolution of a constellation of reasons for a decision one makes. A rationality for a decision is then the reasons, the relationships between the reasons, and the process of their emergence.

2.3.1 Rationality and decision-making

Continuing from Chapter 2.2.1, decision-making relates to the implementation of choice and action based on alternatives. This definition borrows from Simon (1960), and thus is related to considerations of rationality. Indeed, theories regarding decision-making, like bounded rationality, often inhabit the realm of formal rationality based on optimization and analytical models (Bettis, 2017). The opposite is not necessarily the case. Decision-making is deemed to suppose rationality and good reasoning (Elliot et al., 2020; Y. Li et al., 2014), but rationality does not, by definition, include decision-making based on the field of study. For instance, psychologists make distinctions between reasoning and decision-making, while philosophers have practical conceptualizations of rationality that are action oriented and include decision-making, as well as more theoretical accounts of rationality, without inherent decision-making (Knauff & Spohn, 2021). By the definition of rationality used in this dissertation,

the practical rationality that includes decision-making is adopted. Hence, decision-making presupposes rationality, and rationality presupposes decision-making.

While types of rationality are varied in relation to their substantiveness or formality, decision-making concerning itself mostly with formal rationalities can be an issue because theories and models can be computationally impossible in practice (Elliott et al., 2020). Because AI is associated with decision-making based on rationality (Bettis, 2017), specifically formal rationality (Lindebaum et al., 2020), and tested in laboratory settings that do not necessarily translate well to practical contexts (Raisch & Krakowski, 2021), it is possible that AI leads to unmet expectations due to its degrees of removal from contextual embeddedness in practice (e.g., Lebovitz et al., 2021; Van den Broek et al., 2021). Moreover, decisions benefit from justifications and rationalities based on substantive value orientations (M. Weber, 2019), bringing out the inherent moral component in decisions (Moser et al., 2022a). Indeed, because rationality is included in decision-making and vice versa, the type of rationality is inherently tied to the type of decision-making. Formal rationality for decision-making does not allow for value reflexive considerations in decision like substantive rationality does, but formal rationality does not remove the possibility of pluralities of values regarding a decision. They just are removed from the rationality for making the decision.

2.3.2 Rationality and agency

Sometimes, agency is split into rational agency and moral agency (e.g., Liao, 2020; Mabaso, 2021; Manna & Nath, 2021), but full agency is seen as necessitating morality, which is again seen as necessitating autonomy (Wallace, 1999). Rationality is another inherent component of agency: Agency has been historically wed to rationalization (Sørensen & Ziemke, 2007) and many conceptualizations of agency originate from the concept of the free-willed rational individual that emerged with the Enlightenment (Emirbayer & Mische, 1998; Lukes, 2006). Indeed, in their definitions of rationality, Russell (2016) and Wallace (1999) tie it to reasons for *action*, which is at the core of the definition of agency as the capability to act and generate effects. Despite such individualistic origins, there exists some variance in fields as to the sources of agency; neuroscience sees the source of agency in individual decision-making, and, for example, social sciences attribute agency to collectives (Mitchem, 2014).

Despite practices of splitting agency into parts, the concepts of rationality, morality, and agency are inseparable. Reason has been considered a moral faculty (Caldwell, 2007), and even Weber's discussions on the *different types of rationality* are tied to the actions they guide as he talks of value rational acts and means–end rational acts (Kalberg, 1980). A moral agent is an agent that can deliberately and *rationally* make ethical decisions (Abney, 2012). Tracing philosophy back, it is Aristotle who is seen as first connecting morality and ethical behavior to complex and practical rational, emotional, and social skills (Kraut, 2022; F. D. Miller, 1984). Agency and the capability to act can be seen as attributable only to conscious, rational, and moral beings, or in other words, rationality and morality are inherent to agency (Alonso, 2014; Dennett, 2004; Wallace, 1999). However, it is common that views of agency in terms of the role of rationality vary within fields and can even be contradictory. In studying human action, economists are accused, mostly by social scientists, of seeing the role of rationality as impoverished due to viewing it as instrumental. Both pragmatic and epistemological predictive tests favor the inclusion of emotional and social drivers in decision-making (Wallace, 1999). AI employs formality (Lindebaum et al., 2020; Moser et al., 2022a) and can thus be said to at least be a rational agent if, for analytical purposes, we make the agency split. An influential definition of a rational agent comes from Russell (2010), who defined it as an entity that perceives and acts in an environment to maximize the value of a specific measure (Kibble, 2017). This acting upon the environment fits neatly into the definition of agency as the capacity to act and cause an effect, and the maximization fits the definition for formal rationality (Kalberg, 1980; Lindebaum et al., 2020). Thus, by definition, AI fulfills the rationality requirement for agency, but can be seen to lack moral capability. With agency, rationality, and morality all tied together, and with AI regarded as something incapable of morality (Bryson, 2018; Etzioni & Etzioni, 2017). it can be objectionable to attribute to it the status of a full agent. Regardless, as discussed in Chapter 2.2.2, AI effects have a moral component in the decisions made and the way they can steer people's decision-making (Moser et al., 2022a). The question then becomes whether formal rationality can make a moral agent: the relationship of a specific type of rationality with another type of agency.

Many philosophical schools of thought exist on AI moral agency. They vary from those that argue that AI cannot have moral agency to those that see AI as better moral agents than people. The latter reflects views in which emotion and its

associated irrationality is a flaw in human decision-making that can be corrected or improved by employing the “objectivity” of AI (Martinho et al., 2021). It considers emotions as something to be removed from decision-making in favor of pure logic. The view remains persistent despite increasing evidence that decision-making mostly does not, or cannot, happen this way (Y. Li et al., 2014). The former stance on AI moral agency implies that AI cannot be a moral agent due to its lack of understanding of what is morally good or bad, ethically right or wrong, and also lacks the free will to act either according to or in opposition to that understanding (Mabaso, 2021; Parthemore & Whitby, 2013, 2014) because it is always programmed and oriented toward some goal. Without consciousness, AI can only be an artificial moral agent (Mabaso, 2021).

However, the push for pure rationality is inherently a moral question of values. Schools of thought on AI morality that see AI as a superior moral agent due to its pure logic have already imbued it with a fundamentally human and socially originated moral value that can be disputed: Why is pure logic morally better than pure emotion? Can they even be hypothetically separated for the sake of argument? Hence, a debatable moral view is already at the base, and views on AI moral capabilities are, in themselves, a moral question. Choices to employ formal rationality are necessarily both wrapped in and the outcome of a substantively rational value constellation, and they cannot, as such, be neatly separated from each other. Formal rationality is a value-imbued subset of substantive rationality. Should one choose to use formal rationality for decision-making, a value choice has already been made against, for example, random choice.

However, the discrepancy between morality, rationality, and agency specifically for AI decision-making requires untangling. In line with the methodological choices covered in detail in Chapter 3 and considerations in Chapter 2.2.2, the assumption that there is a clear source and possessor of agency is problematized and abandoned. Many of the views on agency arise from the assumption that agency is something that someone or something *has* instead of what it *does*, despite agency being tightly tied to action. The abandonment of this view of agency, and instead considering it something co-constituted and emergent in the relationships between actors, such as between people and AI in decision-making, solves many of the tensions. Now, AI does not need to possess moral agency capabilities but can nonetheless have an effect on and perform acts that have a moral element. Indeed, the stance taken in

this dissertation (see Chapter 3) sees agency as emergent and co-constituted. It does not see anything readily possessing agency but allows the subject–object splits to emerge and be practically analyzed as such. Hence, an agentic subject is possible for analysis but not given *a priori*. This view thus mediates both ideas requested by Caldwell (2007) and Lash (2003) of a practice-oriented concept of agency needing to mediate both the ideas of intentional action, decision-making agency and contextual, embodied, and changing decentered agency.

Now, if agency is something co-constituted and it is assumed that it can be analytically split into moral and rational agency, we solve the dilemma of AI moral agency but are faced with a question of co-constituted rational agency. Taking this view of agency and assuming the stance from Raisch and Krakowski (2021) that augmentation and automation are not either/or and that people and AI are always both/and in AI processes, we cannot posit that the rationality behind decisions is inherently situated in either AI or people. Instead, it must be shared in its constitution and mediated by both. This supports the view that formal rationality cannot be separated from substantive rationality, but rather they are inherently intertwined: Rationality, as a constellation of reasons, includes both value-based reasons and means–ends optimization reasons, the inclusion of which is already a value-based choice. However, this relativity of reason along with agency directly approaches the question posed by this dissertation: How does AI mediate rationality in decision-making?

2.3.3 Rationality and explainability

T. Miller (2019) does a thorough review of explainability that encompasses AI, social sciences, and philosophical understandings of the concept. Based on the literature (e.g., D. Lewis, 1986; Lipton, 1990; Overton, 2011), it is assumed that an explanation is an answer to a “why” question that requires “counterfactual reasoning” in the sense that explanations are not given for causal chains leading to an event or decision by themselves, but in relation to other options: Why this and not that or some other thing? Moreover, these “why” questions can either require a process narrative for why something occurred (and not something else) or a reason for the occurrence “implying intentional thought” (Dennett, 2017). An explanation is seen as comprising two factors: The answer to the question and its presuppositions (Overton, 2011), which can be seen to mirror the goal or desire, as well as the beliefs as presuppositions.

Indeed, explanations are related to abductive reasoning in the sense that when looking back on an event, action, or decision, people employ abductive inference to create the best explanation (T. Miller, 2019). Thus, reasons and the subsequent rationality of a reasoning process are related to explainability for decisions. However, the difference is that under the definition for rationality chosen for this dissertation, rationality is the construction process and constellation of reasons for a decision, whereas an explanation is a backward-looking interpretation of a rationality for a decision and the communication of its interpretation. Indeed, like rationality for a decision, an explanation is also both a process and a product (Lombrozo, 2006). This process includes both a cognitive process for abductively inferring the goal, desire, and beliefs or presuppositions for a decision, and the product is the result of this process, but then there is also the process of transferring that product forward from the explainer to the explainee. Both the selection and the explanation against some normative standard are required for a full account of rationality (Kibble, 2017; T. Miller, 2019).

In their explanations, people often rely on the folk psychology belief, desire, and intentionality (BDI) model concepts of belief, desire, and intention (Wallace, 1999; Winikoff et al., 2021). They are folk psychology concepts, like agency, that may not withstand rigorous scientific scrutiny, but remain relevant and used in practice when people do sense-making of their actions and choices (Kashima et al., 1998; Sørensen & Ziemke, 2007). In comparison to the definitions of rationality that specify acting according to reasons as a part of rationality (e.g., Knauff & Spohn, 2021; Strandberg, 2017), the BDI model highlights acting in one's best interests and with desires, intentions, and beliefs that are consistent with each other as a condition for rationality (Kashima et al., 1998; Wooldridge, 2009). Beliefs, desires, and intentions in themselves can be considered reasons for action (Malle, 2006). People use the BDI model concepts to give explanations for both their own and machine behavior (de Graaf & Malle, 2019; Thellman et al., 2017; Winikoff et al., 2021). Hence, it has been argued that machines should then also construct their explanations for their decisions and actions according to this model to build trust with people and to efficiently communicate explanations using the same conceptual frameworks (Winikoff et al., 2021). For instance, Panisson et al. (2021) proposed BDI agents as a way to meet the requests for AI explainability by making software agents translate and explain their decision-making rationalities with natural language so that it resembles

human explanations instead of, for example, highlighting relevant data.

Explainability and rationality are related processes and products in decision-making with or without AI. To study rationalities for decisions, explanations that answer “why” questions require analysis. Thus, while folk psychology concepts like BDI might not withstand rigorous scientific scrutiny, they need to be acknowledged in studying rationality because they are used in explanations by people and in AI explainability. However, while an explanation is an inference of a rationality after an outcome event or a decision, the rationality formation process happens before an outcome. Thus, perfect access to it through explanation is not something that can be taken for granted, because decisions and events also simply happen without conscious deliberation beforehand (Cohen et al., 1972; Haidt, 2001). Indeed, in addition to accessing rationalities through explanations, the process of their formations and their conditions for existence should be considered. While the study of rationalities for decision-making with AI require the study of explanations, the study of AI explainability requires the study of rationality. Hence, with XAI being a field focused on mitigating AI issues and undesired social impacts, such as discrimination, unfairness, and lack of trust, possibly subsequent to the former issues (Baum et al., 2011; Hasan et al., 2021; G. M. Johnson, 2021; O’Neil, 2016), the study of rationality in AI decision-making is part and parcel of addressing such AI issues.

3 METHODOLOGY

3.1 Research philosophy

The nature of agency, especially agency credited to technology, has been developed from classical intentional action theories to embodied and relational. Rationality and agency are inherently connected and, therefore, such development and consideration for rationality as also relational is relevant. This chapter begins with the study of different views on the relative nature of agency with a focus on technology, from which rationality considerations are inferred based on their relation to agency in Chapter 2. Differing views on agency are underpinned by differing ontological and epistemological research philosophies, which is why their coverage is more relevant to research methodology than Chapter 2.

3.1.1 Sociomateriality

The theoretical lens of sociomateriality has been put forward as a sophisticated theory to take the field of MOS forward in technologies' and, subsequently, AI's agency (den Hond & Moser, 2023; R. Weber, 2020). Sociomateriality is the umbrella term used to refer to a strand of research in MOS that considers the technical and social inseparable and co-constituting of each other (Orlikowski, 2007; Orlikowski & Scott, 2008). This is in comparison to the more prominent "instrumentalist" and "social determinist" views of technology (see Chapter 2.2.2) that either consider technology simply a tool (with a lack of autonomy) that is to be managed by people for their purposes and to solve their problems, or completely social, in which technology is inherently value laden (e.g., Bijker et al., 1987; MacKenzie & Wajcman, 1999), and its emergence and use originates entirely solely from the world of the social (den Hond & Moser, 2023).

Different theories are associated with sociomateriality. Bruno Latour's actor-

network theory (ANT) (Latour, 2007; Law & Hassard, 1999; Leonardi et al., 2012; Orlikowski & Scott, 2008) is perhaps the most well known and utilized one. ANT was created in the 1980s (Callon, 1984; Latour, 1987), and newer sociomaterial theories have emerged since the turn of the millennium. An important concept for these theories is “intra-action,” which was originally coined by Barad (2007) in their introduction of agential realism. Intra-action conceptualizes the sociomaterial view beyond the interaction of the social and material. Instead, they co-constitute each other due to their shared, inseparable history.

ANT makes no distinction or separation between social and non-social “actants” and considers all actions symmetrical and equal within a network of non-hierarchical actants (Latour, 2007; Orlikowski & Scott, 2008; Verbeek, 2014; Wise, 1998). For example, a door handle is considered to act just as much as the person turning it. Agential realism adds observation and measurement as fundamental aspects of sociomaterial approaches: “The real” only comes into existence by measurement. Thus, reality exists only by the means of intra-action of the measurement equipment and the measured—the human observation of reality. (Barad, 2007) Both theories highlight the role of agency, ANT as something that happens in and as networks, while agential realism sees agency not to be the attributable anywhere specific or concrete, but to be the “ongoing reconfigurings of the world” (Mahama et al., 2016).

The measurement, or making an “agential cut” in agential realism refers to making an interpretation and assignment of agency that cuts through the ongoing, constantly in a state of becoming agency. For instance, in calculating environmental costs someone may start from thinking about the factory, pollutants, and their relationships, and thus bring into existence an agential cut of the world. Thus, agency does not exist *a priori* as attributable, but is dependent on this “measurement” or “revealing” the world. (Mahama et al., 2016)

It is not often considered under the sociomaterial umbrella, but the first philosophy of technology shares similarities with agential realism. While it predates sociomateriality as a theory of entanglement of technology and the social, Heidegger’s phenomenology is the earliest explicit *philosophy of technology* that surpasses the modernist dualism of splitting reality into subjects and objects—ideas and the material. Phenomenology is the study of conscious phenomena and experiences (Sanders, 1982): the way people make sense of and experience their life-worlds (Saunders et al., 2009). Classic phenomenology is attributed to two continental philosophers’

thought: Husserl's idealism and logic along with Heidegger's existentialism. Husserl called for a return to the things themselves and how they appear to humans, making consciousness key in experience of the life-world. In phenomenology consciousness is always *consciousness of* something; it is plural, grounded, and directed by people's actions and intentionalities towards the world. (de Vaujany et al., 2023)

Heidegger (2010) in his seminal book argues that tools are connections between humans and reality: The concept of "ready-to-hand" describes how technology becomes an invisible way of interacting with the world and visible only when it is not available. The classic example is the hammer and nail—the focus is on the nail and not the technology that enables the interaction with the world. Despite their invisibility, tools facilitate how people exist in reality and how reality exists for them: Heidegger considers technology as a way of revealing the world and truth, in which technology does not come to exist when it is made, but rather it exists first and then is realized materially. The way of being is already present in the world, from which it is possible to conceptualize and bring a technology into being—reveal it. Similarly, technologies reveal the world: A carpenter making a bench is revealing a way of being that was already present in the world. (Heidegger, 1977, 2010) This revealing quality of technology is similar to intra-action in the sense that reality comes to exist only by technology revealing what "is" to us, a measurement of the world bringing it into existence. Moreover, agential cuts are perceptions and interpretations of reality, and thus studying them as human experience (phenomenology), rather than something that objectively exists outside of consciousness, is fitting.

Due to their similarities, sociomaterial theories face same kinds of criticisms. While making actionable "cuts" with agential realism tackles with issues related to ANT related to all agency being delegated forwards and further in the networks until it ceases to exist meaningfully or applicably (Introna, 2014; Mitcham, 2014), agential realism faces the critique that it can be analytically and empirically problematic compared to theories that make a distinction between social and non-social—subject and object (e.g., Tunçalp, 2016). This acknowledged issue is pertinent to also phenomenology, as it considers technology as a monolithic societal force instead of concrete technologies: Phenomenology can only address the conditions of being of technology—not technologies themselves. (Verbeek, 2005, pp. 93–95) Therefore, phenomenology as a philosophy of technology can be considered difficult to benefit from in practice (de Vaujany et al., 2023).

While one issue with the rich theoretical lenses provided by sociomateriality's is their lack of pragmatism, another one is that they get stuck in ontological debates. For instance, Tunçalp (2016) notes that critics of sociomateriality proclaim that its proponents should behave like “normal human beings” and stick with concepts that have been just fine to date. Tunçalp (2016) also proposes actual critique in that the ontology of sociomateriality can be analytically and empirically problematic, and suggests a turn to critical realism, in which the social and material are seen as separate, but becoming entangled in the means of action. However, the argument posed by Tunçalp (2016) is that “by combining the material and the social, we may actually limit our understanding of distinctly material and social issues.” This quote shows that Tunçalp (2016) is not concerned about analytics or empiricism, but about the distinction between the social and material that is “forgotten” in sociomateriality—an approach that rests on the premise that this protected distinction is false.

3.1.2 Pragmatism

Indeed, phenomenology is like ANT in the sense that it is a sophisticated theory, but it often leaves one wanting when moving away from pure theory into practice. Moreover, while agential realism can tackle agency and prevent its dispersion into meaninglessness by making agential cuts, it is caught in ontological and metaphysical debates that can make its application arduous in research. This is expected because discussions around the agency of technology and intentionality are typically suspended in the dualist paradigm, which has been considered outdated for decades (Latour, 2012; Verbeek, 2014; Wise, 1998). Verbeek (2014) suggested that this is because the proposition of applying agency to artifacts, or overcoming the dualism, typically faces wide opposition and even high emotion when proposed because people are afraid that their human freedom and morality are being outsourced to objects that they see as incapable of morality, completely missing the point of any such theories that focus on the co-creation of agency and moral agency through the relations and complex intra-actions of human and nonhuman. Typically, sociomaterial views focus on a co-constitution of agency in which technology directs intentionality, and thus agency, but does not have it. As a result, dualists fight an argument that has not been made (Verbeek, 2014), but rather, they insist on a manufactured human–nonhuman, subject–object split. For AI specifically, the opposite calls are also heard—insistence on dualism and that the technological agency of AI is equal

to human agency (e.g., van Rijmenam & Logue, 2021).

The phenomenon under study in this dissertation is AI as a specific technology and its mediation of decision-making rationality. However, while its agency in decision-making raises interesting sociomaterial questions, sociomaterial theories can suffer from a lack of practical usefulness. ANT disperses agency to the extent that it cannot be meaningfully studied; phenomenology cannot address specific technologies, only their conditions of being; and agential realism gets caught in theoretical debates, which can be theoretically fruitful, but neglect practice. However, the philosophical tradition of pragmatism exists to overcome specifically such issues: Pragmatism is a philosophy focused on problem-solving that is analytical as well as prescriptive (Farjoun et al., 2015). This problem-solving is aimed at the practical, and thus, pragmatism allows addressing research questions from different philosophical positions (Saunders et al., 2009). For pragmatism, only that which supports action and practical consequences is relevant (Kelemen & Rumens, 2008). Indeed, pragmatism is founded on avoiding metaphysical debates while focusing on the concrete, and thus, its stance on ontology allows for multiple possibilities, out of which the one that best addresses the research questions should be chosen (Patton, 2014; Saunders et al., 2009).

Epistemologically, pragmatism emphasizes practical consequences regarding the value of research and, thus, welcomes multiple possible sources for acceptable knowledge (Kelly & Cordeiro, 2020; D. L. Morgan, 2014; Saunders et al., 2009). Pragmatism considers empirical observation, as well as subjective meanings, to be acceptable knowledge, and it is thus compatible with qualitative and interpretative understandings of a constructionist reality of multiple truths (Farjoun et al., 2015; D. L. Morgan, 2014; Saunders et al., 2009). Therefore, pragmatism is typically mixed or multiple methods research that combines both quantitative and qualitative data (Saunders et al., 2009).

Compared to other well-established MOS epistemologies (e.g., rationalism, which focuses on individuals as rational agents), pragmatism emphasizes that knowledge is based on experience and allows, through the study of complexity and diversity, a richer exploration of constantly changing human behavior in organizational contexts (Kelly & Cordeiro, 2020). Pragmatism is always from a certain perspective and oriented toward a perspective—consciousness of something (Moser et al., 2022a).

Pragmatism is well suited to studying change and complexity on multiple levels

of analysis (Farjoun et al., 2015), and as such, it can provide a suitable philosophical approach compared to other options (Kelly & Cordeiro, 2020). Moreover, pragmatism is anti-dualist in that it avoids categorical splits between means and ends, science and arts, and subject and object, and focuses research attention instead on the process of mutual constitution of emergent categories in contexts (Farjoun et al., 2015). Thus, it is well suited as a research philosophy to study the mutual constitution of the social and material, while overcoming the problematized subject–object splits in AI decision-making.

3.1.3 Postphenomenology

While phenomenology is criticized as not pragmatist enough (de Vaujany et al., 2023), it shares with pragmatism the embracing of rich, complex, and unique situations and realities in organizations, making it a fruitful research philosophy for MOS (e.g., Ehrich, 2005; Sanders, 1982). Moreover, phenomenology complements pragmatism in its avoidance of dualism and enables researchers to study the multiple ways agents relate to their life-worlds due to its more rich and complex ontological understanding of agency as embodied, temporal, and embedded (Tsoukas, 2023).

Taking phenomenology and combining it with pragmatism yields an actionable philosophy of technology that honors anti-dualism in sociomateriality. Introduced by Don Ihde in 1990, *postphenomenology* is defined as the combination of phenomenology and pragmatism (Ihde, 2012; Ritter, 2021b). “Post-phenomenologies” can refer to a stream of research in conversation with phenomenology but that are not exactly phenomenology (de Vaujany et al., 2023). However, with technology perhaps one of the most influential mediators of humans’ experience of their life-worlds, postphenomenology as defined by Ihde (1990) and Verbeek (2005) studies the role of technologies in co-constituting lived experiences and in the direction of human-embedded intention and action (Verbeek, 2005, p. 116).

Postphenomenology is the analysis of relations of humans and their life-world, as named by Ihde (1990), and how technology shapes these relations, both in how humans are present in their life-world and how people experience their life-worlds to be present to them (Ihde, 1990; Verbeek, 2006). Postphenomenology centers on specific technologies in their use contexts and empirically studies how those technologies mediate “experiential correlations and associated subject–object constitutions” (Zwier et al., 2016, p. 317). Postphenomenology is nonsubjectivist and interrela-

tional and, in accordance with pragmatism and agential realism, focuses on the process of mutual constitution of the subject–object relationship rather than rejecting it entirely (Ritter, 2021b). Subjects and objects exist as emergent from within intra-action. Postphenomenology, as phenomenology, rejects that technologies could be considered independently of the people who exist in relation to the technologies. Technologies do not have essences, but exist in order to act the same way that consciousness is always conscious of something (Verbeek, 2005, pp. 117–118). Hence, postphenomenology does not consider technology to have a positive or negative moral essence, but they act as the mediums of morality (Rosenberger & Verbeek, 2015, p. 13).

3.1.3.1 Technological mediation

The concept of mediation is a key difference between postphenomenology and other sociomaterial theories: Postphenomenology does not consider all actants symmetrical in their agency, but rather, humans and things create hybrid agency via both interaction and mutual constitution (Verbeek, 2014). Thus, subjects can be seen as subjects and objects as objects—“distinct but not separate” (Greenwood & Wolfram Cox, 2022)—because they are constituted as such through acting and agency in which technologies function as active mediators between people and their life-worlds (Verbeek, 2005). This overcomes the inactionability of ANT. Indeed, postphenomenology is in this regard closer to agential realism, but the unit of analysis in postphenomenology is mediation, specifically (Greenwood & Wolfram Cox, 2022; Verbeek, 2005), and agential cuts are not considered. Postphenomenology is commonly applied to practical ends and used in design disciplines to actively consider how a technology is desired to mediate human experience and action in the life-world. The results of postphenomenological analyses are implemented in a technology’s design (Verbeek, 2006, 2014).

Technological mediation in postphenomenology includes the mediation of action and the mediation of perception (Verbeek, 2005). The mediation of action is the way technology mediates how people exist in the world and can be considered similar to the scripts in ANT presented by Latour (1992): the way the material mediates action and actionability. Scripts encourage some actions and discourage others. The classic example is the speed bump, where a way of acting that encourages slowing down is inscribed in the technology. This can also happen through signifiers that

mediate action, such as traffic signs. Postphenomenology accounts for the mediation of perception, which is not included in the analysis of scripts. The mediation of perception builds from Heidegger's phenomenology and studies how the life-world is present for humans: Technologies offer some interpretation of the world to humans and thus are intentional and not neutral instruments (Ihde, 1990). The classic example by Verbeek (2006) is that of ultrasound: The present possibility of examining a pregnancy with ultrasound mediates what humans consider real and mediates the perception of a potential child to a potential patient, which again mediates how that pregnancy is regarded and related to—which aspects of it become “real.”

An important notion in technological mediations is the concept of multistability, which notes that mediations are not permanent and immutable, but differ depending on context (Ihde, 2012). Thus, the mediation is not an intrinsic property of the technology, but the mediation is also co-constituted. Thus, technologies have no “essences” outside of the contexts of use in which it is interpreted and understood (Verbeek, 2005, pp. 117-118). Multistability is often demonstrated when a technology meant for one goal or action is repurposed upon entering a new context.

Technological mediations in the postphenomenological literature are sometimes placed in different typologies: Verbeek (2005) differentiated mediations of humans being in the world (existential) from how the world is present and perceived by humans (hermeneutical). Moreover, Ihde (1990) originally presented examples of types of mediation: embodiment, in which something becomes a part of you (e.g., eyeglasses); hermeneutic, in which something interprets the world for you (e.g., thermometer); alterity, in which technology is related to as a pseudo-other (e.g., a chatbot); and background, in which the relation has mainly an environmental effect (e.g., central heating).

Kiran (2015) presented a way to categorize technological mediations based on the notion from Ihde (1990, p. 76) that mediation is always two-sided and magnifies one thing at the expense of the reduction of something else. The example used is a telescope: the details of the perceived object are magnified, but its relative position and placement are reduced. Kiran (2015) posited that studying the two-sided dimensions of technological mediation opens up further possibilities, and emphasized that these dimensions are not limited and could include, for instance, legal allowing–prohibiting or political liberating–oppressive dualities of technological mediations. Regardless, Kiran (2015) used four such dualities as examples of how

to analyze technological mediations: epistemological magnification–reduction, existential revealing–concealing, practical enabling–constraining, and ethical involving–alienating.

The magnification–reduction dimension was covered extensively by Ihde (1990), who considered how knowledge and what constitutes knowledge are mediated by technologies. Technology mediates what of the world becomes available to us, but also what becomes removed from us. Specifically, as already noted, it mediates what is magnified in knowledge or as knowledge and what is reduced, for example, in phone calls or emails compared to face-to-face communication: The explicitly stated becomes magnified, while body language or tone become reduced.

Despite not being originally focused on in postphenomenology, Kiran (2015) considered the revealing–concealing dimension fundamental to further analysis of the dual dimensions of mediation. These dimensions begin with Heidegger’s phenomenology and philosophy of technology. In particular, the readiness-to-hand of technologies reveals the world to us through possibilities. The in-order-to of available technologies reveals new ways of being (Heidegger, 1977), for instance, the log reveals itself in the world differently to someone with a log mill than to someone without. Simultaneously, technology entrenches certain ways of being and blinds us to possibilities without that technology (Kiran, 2015).

The enabling–constraining dimension analyzes what technology enables us to do that we otherwise could not and what it rules out of our possible actions. Kiran (2015) highlighted technology’s ability to enable us to do things with it that we could not do without it, such as in assistive technologies. However, it is often forgotten that technology shapes the ways we do things and thus constrains our ways of doing and conceptualizing other ways of doing. In other words, “habits conceal potentiality” (Kiran, 2015).

The involving–alienating dimension considers how technology mediates contexts as ethical by creating pathways and restraints on how we engage with them and how we choose. Technology mediates the moral agency of agents and sets limitations on how ethical choices are formulated or are formulate-able. Self-driving cars and the placement of responsibility in case of an accident is a yet unresolved ethical puzzle of an involving–alienating technology in which the technology alienates the driver from moral decision-making, but it involves the manufacturers of cars in previously unrevealed ways (Kiran, 2015).

Of these dimensions, the three latter ones are of interest because they concern the mediation of decision-making rationality formation processes in decision-makers. The other dimensions mediate the conditions of the process, such as what knowledge is emphasized and what is pushed to the background, but the revealing–concealing, enabling–constraining, and involving–alienating dimensions apply to the unfolding of the rationalization process itself: how it is revealed, its possibilities, and the involvement of the decision-maker. Thus, these dimensions of mediation tackle agency as they are an active, involving process for decision-making.

3.1.3.2 Criticism

Postphenomenology’s view of agency prompts similar criticisms as sociomaterial theories in general. For instance, Illies and Meijers (2014) considered the co-constituted agency of postphenomenology absurd. They understand postphenomenology to posit that in the case of a shooting crime, the combination of man and gun is responsible for actions, and the combination should be “put to jail.” Illies and Meijers (2014) proposed, against their understanding of postphenomenology, that the technologies simply affect an agent’s action scheme of physical, social, and intentional contexts, in which different acts are different levels of attractive, and technologies have “second-order responsibility.” However, Verbeek (2014) clarified that acknowledgment of technological mediation does not imply placement of responsibility onto the mediating technology. While Illies and Meijers (2014) suggested a responsibility of second-order, they conceded that the technologies shape the actions of an agent. However, like Tunçalp (2016), their suggestions for theory improvement rest on enforcing a dualism of subject–object in opposition to postphenomenology, which acknowledges that technologies are not separate variables that shape an action scheme, but that the actor’s scheme and conceptualization of a scheme is, in the first place, constituted by them.

A legitimate criticism of postphenomenology is its insensitivity to broader social and political contexts (Ritter, 2021b). Postphenomenology focuses on specific technologies themselves and their mediations of individual humans being in and perceiving the world, but it is not applicable to extending this analysis to the larger world (Rosenberger, 2014). Critics have doubted postphenomenology’s ability to contribute meaningfully to reflections on the general impact of technology (Ritter, 2021a). Thus, there are calls for expanding postphenomenology as a philoso-

phy of technology beyond individual technologies. Here, Rosenberger (2014) suggested that where postphenomenology fails, ANT excels. Thus, these theoretical and methodological approaches can supplement each other. Moreover, Ritter (2021b) called for postphenomenology to expand what is understood as technology beyond utility and function, into also what is “invisible” about technologies.

This dissertation acknowledges these points regarding postphenomenology and expands analysis beyond the utility and function of the technology in question into also the narratives and conceptualizations of the technology. However, it is acknowledged that with AI, what remains invisible about the technology could reach all the way back to mining materials for computing hardware (Crawford, 2021), the energy required to run complex AI models (Dhar, 2020), and the unseen manual labor and its effects on the workers (Perrigo, 2023). For a meaningful analysis of the technology’s effects and mediations, a scope must be set. Otherwise, postphenomenology will lose its pragmatic edge: the creation more nuanced understanding of human-technology mediations of specific technologies in use (Friedrich et al., 2022). If extended too much postphenomenology will suffer and become too abstract and large and face the highlighted issues of ANT or Heidegger’s phenomenology.

The context of this dissertation is a singular decision-maker and how AI, as a specific technology, mediates the rationality for decisions made. While larger impacts on society or politics are not the unit of analysis, the implications of the results and findings are considered for social impacts. The unit of analysis is the mediation of rationalities for a specific technology. Thus, the raised issues regarding the shortcomings of postphenomenology are not significant. Instead, the merits of postphenomenology’s ability to actionably analyze how a specific technology mediates the way individual people interpret and act in a specific context are pronounced in the research setting of this dissertation. Moreover, the pragmatism of postphenomenology overcomes the debates relevant in other sociomaterial approaches. Thus, it is considered a well-suited research methodology to address the research questions.

3.2 Research context and methods

3.2.1 Postphenomenology as a research methodology

Postphenomenology is a philosophy that always uses empiricism as grounds for philosophical reflection (Crease & Achterhuis, 2001; Mol, 2002; Rosenberger & Verbeek, 2015) centering on understanding the roles of technology in human–world relations and their implications. This reflection and conceptual analysis was used to study the human–world relations and the worlds and subjects co-constituted by the technologies under study. Typically, specific dimensions of human–world relations are studied (Ritter, 2021b; Rosenberger & Verbeek, 2015).

The empirical work used as a basis for postphenomenological work can be secondary empirical work of others, primary self-conducted studies, or first-person reflection. The objective is to investigate the character of technological mediation's dimensions and their implications, not the creation of technology descriptions. Typical postphenomenological work is a conceptual analysis of specific dimensions of human–world relations (Rosenberger & Verbeek, 2015, p. 31). The dimensions in this dissertation are revealing–concealing, enabling–constraining, and involving–alienating of decision rationality.

Despite postphenomenology not being restricted to a certain methodology, the empirical works upon which postphenomenological analysis is based are typically case studies that concern concrete intra-actions between humans and technologies. In case studies, postphenomenological claims are analyzed in the contexts of the studies (Rosenberger & Verbeek, 2015, p. 31). This methodology is seen to add strength to postphenomenological analysis because it generates rich and interesting descriptions of human–technology relationships for analysis (Ritter, 2021a), such as possibilities of identifying advantages, disadvantages, and points of potential expansion (Rosenberger & Verbeek, 2015, p. 31).

Case studies have been increasingly popular and are the basis for the most influential studies in MOS (Eisenhardt & Graebner, 2007). One reason for this is that they can tackle complexity and rich context accessibly (Eriksson & Kovalainen, 2015). Case studies can focus on a single case or on multiple cases, and they focus on empirical descriptions of phenomena of interest in relation to their historical, economic, technological, social, and cultural contexts in order to build propositions and testable

theory based on a variety of data (Eisenhardt, 1989; Eriksson & Kovalainen, 2015; Yin, 2009). Single case studies that are independent analytic units have the potential to paint rich pictures and create deep understanding, but multiple case studies that mimic related or replicated laboratory studies provide a stronger theory-building basis (Eisenhardt & Graebner, 2007; Yin, 2009).

Indeed, case studies typically aim to build empirically testable theory upon which a recursive process of testing and refining generalizable theory will rest. Building theory from cases is pervasive because cases are an excellent bridge from rich qualitative, explorative empirical work to testable theory that emphasizes the development of constructs and testable propositions (Eisenhardt & Graebner, 2007; Yin, 2009). Classic case studies aim to understand the perspectives of the people involved in the embedded contexts and learn the specifics and uniqueness of each case (Eriksson & Kovalainen, 2015). This typically is referred to as thick description of cases, meaning a communicated interpretation that captures the rich and multifaceted aspects of the case (Eriksson & Kovalainen, 2015; Geertz, 1973). Thus, case studies fit well with both pragmatist and phenomenological research philosophies, despite those philosophies being sometimes thought of as unscientific anecdotes. However, MOS research in particular requires the practicality and normativeness enabled by case studies (Eriksson & Kovalainen, 2015).

While the publications included in this dissertation are not published as case studies, phenomenology as a methodology allows secondary empirical work by others, primary self-conducted studies, or first-person reflection as the basis for empirical case analysis. Thus, the empirical studies of the publications are used as case studies for researching mediation with postphenomenology. Hence, the methodological approach is a multiple case study of postphenomenological analysis. The aim is to provide testable propositions about the AI mediation of rationality in decision-making by providing rich descriptions of it in multiple related cases. The propositions are relevant to a variety of AI fields, including XAI, AI ethics, maintenance of organizational learning (Balasubramanian et al., 2022), and unique human knowledge (Fügener et al., 2021).

3.2.2 Data collection and analysis

For postphenomenological analysis, this dissertation considers the empirical studies in publications I–IV each as their own separate but related case studies upon

which the postphenomenological analysis is founded. Publication V is a theoretical paper that gathers recent discussions around AI in decision-making and posits the phenomenological concept of bad faith by Jean-Paul Sartre as a function of AI in decision-making. Thus, publication V provides a point of view for phenomenological analysis and conceptual reflection on the case studies. While publications I–IV do not employ a postphenomenological methodology to study rationality formation, postphenomenology allows for the analysis of technological mediation in the AI decision-making process contained in them to be examples of the technology in use.

The study in publication I used news sentences gathered with a keyword search on camera manufacturers from the LexisNexis database for multiple UML algorithms and preprocessing choices in the AI process. It studied the reproducibility of UML research methodologies as a process that includes various choices related to AI decision-making both before and after the actual implementation of algorithms. Publication II used LexisNexis news pieces collected with the keyword “Statkraft,” which are classified into resource-based categories (based on G. Morgan & Smircich, 1980). Various SML algorithms were taught as classifiers to learn the resource-based classification. The news pieces that the classifiers mislabeled were assessed by a human to note patterns in the mistakes, but the human was not expected to make improvements or corrections to anything. Publication III used similar data in which news sentences were searched for the keyword “Kodak” and retrieved from LexisNexis, then sorted into the same resource-based categories. Multiple SML algorithms and approaches were again employed to teach the labeling to a classifier, but here, two people assessed the outputs and they were given the task of making improvements to the labeling and of retraining the AI based on the improvements. Publication IV used a similar methodology, but the texts extracted with the keyword “European Center for Nuclear Research” (CERN) on LexisNexis included three-sentence snippets from the news. In this publication, the labeling of the snippets according to the mission statements of CERN was performed by humans who were also tasked with assessing the outputs of the classifiers taught to do their labeling as well as those taught with other people’s labelings. The rationality formation processes of those assessing and comparing their labeling rationality to an AI output were compared to those who did the same comparison against other people’s labels. This methodology is summarized in Table 3.1.

Table 3.1 Research design and methodology in the publications

| | Data | Classification | Machine learning | Research design |
|------------------------|---|--|---|---|
| Publication I | News sentences retrieved with the keywords of camera manufacturer names. | Only the number of classes to create for some algorithms. | Unsupervised machine learning, topic modeling and clustering. | Based on discovered unreproducibility of AI methodology in research literature, effects of the undisclosed decisions along the AI implementation processes are studied. |
| Publication II | News sentences retrieved with the keyword "Statkraft". | Resource-based categories: informational, human, organizational, relational, financial, legal, physical. | Supervised machine learning, a variety of algorithms. | Typical AI classification task of label decision-making outcomes are evaluated and assessed by a user of the AI tool. |
| Publication III | News sentences retrieved with the keyword "Kodak". | Resource-based categories: informational, human, organizational, relational, financial, legal, physical. | Supervised machine learning, a variety of algorithms. | Two people are given the task to work as humans-in-the-loop to improve AI decision-making. First separately making corrections to the AI decisions, then together. The effects on AI accuracy and explainability are studied. |
| Publication IV | News sentences retrieved with the keyword "European Organization for Nuclear Research". | European Organization for Nuclear Research mission statement dimensions: technology, science, human resources. | Supervised machine learning, Multilayer Perceptron used. | One group is given decision-making advice by an AI, comparison group is given decision-making advice by other people. The differences in the rationalities they give for their decisions are compared. |
| Publication V | - | - | - | Theoretical paper. |

Thus, the data used and the methodologies were similar across studies in the empirical publications, and they are thus comparable for postphenomenological analysis. Moreover, unstructured text data were deemed suitable for the studies because it is estimated that over 80% of all data available for organizations are in the form of unstructured text (Gandomi & Haider, 2015; Robinson et al., 2020). Thus, the case studies were conducted on data that are representative of the data available to organizations. Moreover, the methodology covers both UML and SML, which represent both the most widely used AI approach (SML) (LeCun et al., 2015; Sindhu Meena & Suriya, 2020) and its complement (UML). Hence, a good scope of AI methodology is accounted for despite some other types of AI being omitted from the scope.

The case studies covered a richness of human–AI relations: Publication II considers human–AI interaction in which the human is tasked only with evaluating outputs. Publication III considers two humans with differing rationalities interacting with AI who were tasked with improving the correctness of the labels taught to the AI. Publication IV makes comparisons of human–AI relations to human–human relations. Thus, various types of human engagement with AI in decision-making are included. Moreover, publication I studied the AI process in terms of the choices present in the lead-up to running AI, as well as the analysis post-AI, on a detailed level. Other publications consider it on a more general level. Based on these case studies, postphenomenological, phenomenological, and conceptual analyses were conducted with the dimensions of technological mediation being revealing–concealing, enabling–constraining, and involving–alienating in the context of AI decision-making. The analysis is similar to phenomenological analysis: First, the cases are described, after which constant themes and subjective relationships to the themes are covered, from which abstractions are then inferred.

4 FINDINGS

4.1 Publication I: Advancing reproducibility and accountability of unsupervised machine learning in text mining: Importance of transparency in reporting pre-processing and algorithm selection

Incentives to use UML rather than SML are increasing along with sizes of data: SML always requires labeled datasets and is associated with human bias and the expensive human labor that goes into labels, while UML avoids this laborious and fickle step (Kobayashi et al., 2018; Tonidandel et al., 2018; Ziegler, 2012). However, this publication demonstrates that despite the lack of acknowledgment, UML and SML require qualitatively different methods and best practices. Hence, studying UML separately is required. This publication addresses AI mediation of rationality in UML, and consequently highlights the mediation effects of rationalities of AI decision-making not only for the output, but also in the process leading up to the output and its analysis. Thus, the publication complements subsequent papers that emphasize the mediation of rationality in the SML analysis and associated required labels.

Pre-processing, algorithm selection, and data analysis are present in both UML and SML and are starting to be emphasized as crucial steps for AI outputs (Hickman et al., 2022). However, studying UML better highlights the impact of the rationality of making these pertinent pre-processing decisions in ML analysis: The construction of the rationality throughout AI decision-making. This publication examines current reporting practices and the accessibility of rationality in making pre-processing, algorithm selection, and output analysis decisions for the implementation of AI decision-making. The publication finds issues with research reproducibility and demands that UML implementations make the decisions apparent. Moreover, the publication provides principles of rationality clarification for these choices.

First, the current unnecessary situation is concealment of the rationality underlying decisions for pre-processing, algorithm selection, and output analysis in reports in the body of literature. Decisions are necessary in all of these phases, and they have crucial impacts on the outputs of the AI tools. Hence, it is curious that the rationality behind them is neither disclosed nor required in scientific publications. This prevents reproducibility of the research. Moreover, the lack of rationalities provided for pre-processing decisions presents such rationalities as arbitrary. Thus, the rationality itself and whether one exists in the first place are obscured and concealed. Regarding outputs, it is not only analysis decisions that are concealed but also what is “looked at” in the analysis to make conclusions and draw inferences. Are the outputs contrasted with the data or, for instance, are the data in the same clusters compared to each other? Or are only some representations of the output clusters used to draw conclusions and make inferences? UML methods, such as the clustering and topic modeling used in the publication, often reveal the most relevant words used in creating subgroups from the data. Looking at the data, it is not only a representation of a subgroup, but rationality is also revealed whether, for instance, specific words are the main reason for placing a data point into a specific subgroup.

It is only in deliberately examining the data that the subgroups reveal a tangible rationality for the outputs. Hence, when a rationality is not provided for conducting this type of analysis or not, even the rationality for accessing rationality is concealed. The concealing of a rationality in the pre-processing decision-making process paints an image of the researcher(s) making the decisions as removed or alienated. The decisions seem to simply happen without someone’s deliberation. Indeed, concealing rationality created alienation, whereas involvement revealed rationality.

The current concealment of rationality for decisions constrains any possibility of assessing the approaches, which in turn makes it impossible to determine whether certain decisions were made for sound reasons or to perhaps tweak outputs until they yield the desired results. Thus, because the *current* conventions make decisions appear arbitrary, they constrain critique and the consequent development of alternative rationalities. Possible rationalities are constrained to formalizing similarities or differences between datapoints based on which data are grouped together or separated. The constraint of rationality to similarity/difference enables the assessment of a potential rationality by creating explanations: “What is different/the same in these UML-created groups?” From there, rationalities can be guessed at. However,

they will remain only guesses since they provide no “truth” about what rationality was employed by the UML algorithm in creating the output.

In summary, analysis rationalities and the decisions leading to them are mostly concealed, but that is not an inherent property of either UML or AI in general. Such concealment was found to be related to alienation from the AI process. Whereas label rationality is not applicable to UML, as it is for SML, the decisions and the rationality for the processes here are applicable to SML. Moreover, in both UML and SML, the constraints on rationality for the outputs are often similarities and differences in data features.

4.2 Publication II: Supervised machine learning for detecting patterns in competitive actions

SML is the most common form of ML used in applications (LeCun et al., 2015) and is seen as a transformative power for Industry 4.0 (Chalmers et al., 2021; Schwab, 2017). SML requires a pre-labeled training dataset that serve as the basis for SML algorithm-constructed maps. They “learn” to label the data similarly to the original categorization as efficiently as possible. Thus, SM-based AI applications can only be as good as the data: bias and incompleteness are acknowledged issues (O’Neil, 2016). This is where the saying “garbage in, garbage out” comes from. Specifically with SML, decision-making can be conceptualized as containing multiple points of decisions: assigning the teaching labels is one decision made. What analysis methods or algorithms are used is a decision. How the analysis is used is a decision. Moreover, if and what actions are taken based on the analysis are decisions.

The purpose of this publication was to study SML in a typical setting: A labeled text dataset was used to train and optimize an SML algorithm in creating an automated tool for data analysis. In this case, the tool is for automatically mapping classified industry player actions onto a map, based on action times and the resources concerned. The SML tool succeeded well and learned to predict action labels according to the way it was taught. To create a realistic case pragmatically, the original data used for teaching the SML was labeled by external people unknown to the end users and evaluators of the output maps. Only AI label prediction accuracy was focused on in the creation of the tool. The mediation of rationality throughout this “basic” SML process was examined. The evaluator and user of the tool looked at the SML-

labeled texts and contrasted them with the original teaching label in cases where they differed. This evaluation does not necessarily apply to every process as it is possible to use the output map uncritically. The impacts of this evaluation on rationality mediation are notable, and studying them brings forth the mediation of rationality in comparing reflective and unreflective uses of AI outputs for decision-making.

First, in the revealing–concealing dimension, the original label’s rationality was black-boxed and concealed from the rest of the process. Going into the process, it was not known to what extent the teaching labels were personal judgments from different labelers and whether the original labeling had been affected. Thus, the rationality upon which the SML tool was taught was concealed from the process. Of course, the interpretable nature of the data labels makes this a more complex case for rationality than labeling different handwritten characters, for instance. However, there is never a perfect case; for example, handwriting can be messy. This means that some interpretation is always employed in labeling. Similarly, in looking at the accuracies of the algorithms, the reasons why certain algorithms yielded better accuracies than others were concealed. AI optimization brings forward the result, but conceals the reasons for the result. In looking at and comparing the AI output labels to the original sentences and original teaching labels, the substantive, subjective rationality was revealed between the end user of the tool and the original labelers: Multiple label options were suitable, and the end user revealed their own differing rationality and reflected on it. For instance, “Statkraft sells minority interest in UK onshore portfolio to reinvest in new renewable energy” was revealed to be either financial or relational depending on the rationality used. Thus, new ways of labeling and associated rationalities were revealed. If the discrepancies were not available or were omitted from scrutiny, this variance in rationality would be concealed. If, in addition to the quantity of mistakes, the qualitative differences of possible labels can also be studied, the differences or possibilities in decision-making rationalities are revealed. Thus, differences of rationality can be searched for and corrective action enabled if an undesired rationality is seen in either the original data or AI outputs.

Second, the original teaching labels in the training data already forced a constraint on rationality: Rationality is constrained to sorting the data into set categories. This enables only certain types of rationality to emerge in categorizing and labeling the data. Due to the SML structural requirement for data in this form, AI constrains rationality into specific locked-in categories. The lack of transparency into AI al-

gorithms constrains possibilities to even begin studying the rationality differences between algorithms. Some algorithms are more transparent than others, but they are often less preferred (Forrest, 2021; Rudin, 2019). Hence, because different algorithms do not reveal their rationalities equally, the direct comparison of rationality between them is constrained. Moreover, looking at the “mistakes” of the AI outputs compared to original teaching labels enabled the reflection of betterment of the classification: “Statkraft has acquired the Irish and UK wind development businesses of the Element Power Group” was in the teaching data labeled “physical resources.” However, some SML algorithms suggested it should be in “organizational.” This prompted scrutiny of the rationality for why one label is better than the other—bringing forth the contrastive nature of rationality (D. Lewis, 1986; Lipton, 1990; T. Miller, 2019; Overton, 2011). Thus, involvement with the differences enables and reveals the possibility of being surprised and exposed to differing decisions and the refining of rationality. AI also therefore enables a greater variety of rationalities *if the chance is taken*.

Third, in the involving–alienating dimension, the original teaching labeling is an involving phase, but its rationality is lost and concealed further along in the AI decision-making: The original labels seem to be given without efforts of involvement, and thus, alienation occurs. In the creation and testing of the algorithms, the developer is alienated from the rationality if they focus on attaining the best accuracy. The rationalities for the algorithm’s and others’ performances are of no concern because the question of why one method is better than another is not asked. In this approach, it is only when the end user looks at both the AI and teaching labels that the rationality behind the labels becomes involving for the user. The evaluator began reflecting and considering whether the labels were correct and which could be placed into other categories.

The rationality behind the teaching labels for the task remained concealed. Regardless, the AI algorithms did their best to learn to label data according to this concealed rationality and to formalize it into calculations. The accuracy honing alienated the developer from studying the differences in rationality for the algorithms, which was added to the constraints of set label requirements and transparency imposed by the algorithms. Moreover, the only revealing and enabling happened with active reflection by the end user—involvement. This allowed for reflection on the quality of rationality and revealed subjectivity and novel rationalities.

4.3 Publication III: Human-in-the-loop: Explainable or accurate artificial intelligence by exploiting human bias?

HITL typically refers to the augmentation of AI decision-making, which is used with AI to advise and enhance human decision-making. In practice, this often means a human (or humans) are assigned to the AI decision-making process to “supervise” the outputs and make corrections and updates on points of improvement. This is in contrast to “automation,” in which no humans are involved in the AI decision-making process. However, this separation is often not possible (Raisch & Krakowski, 2021) and serves to alienate humans and their rationality from the AI decision-making process. HITL is proposed as a solution to managerial and larger societal issues, such as bias (Krügel et al., 2023). HITL is also seen to increase accuracy and explainability. While accuracy and explainability are typically in opposition to each other, their trade-off is mostly a myth (Rudin, 2019).

The purpose of this publication was to extend the research from publication II and study a setting in which explainability is attended to in addition to accuracy. Hence, this publication used a similar setting as publication II, but the difference is that here, there were two evaluators of the AI decision-making outputs making revisions to the output labels and feeding the corrections back into the ML algorithms for retraining. Thus, this publication extends the work into studying how updating data based on revised rationality affects SML performance and also examines the impact of involving two humans with different rationalities in the decision task. The publication tackles the co-constitution of multiple substantive rationalities, with AI mediating the decision-making.

First, with regard to revealing–concealing, this publication found that differences in substantive rationality reveal the formation of rationality when addressed: The revealing of rationality happened when two substantive rationalities were converged and synthesized into a more formal rationality. For instance, what became clear in discussions concerning the differences was that the other human-in-the-loop (HITL-2) did not consider extra clauses in the texts, whereas the first human (HITL-1) included the clauses and made classifications based on actions appearing in extra clauses. In comparison, if the differences in the rationalities of the two HITLs had not been addressed, the conflicting and differing rationalities in the revisions made by the HITLs would have been concealed and formalized into the AI process. This would

lead to decreased accuracy and concealment of rationality. A sole focus on accuracy would have concealed possibilities for refining rationality. Similarly, having only one HITL would have concealed the extent of other possible rationalities. This supports the knowledge that the convergence of a variety of opinions can be detrimental to the performance of a group (Da & Huang, 2020). However, the approach of addressing differences revealed completely new possibilities beyond both substantive rationalities: Previously varied labels on market description data were realized to be relational—to competitors. Previously, HITL-1 had coded them as informational and HITL-2 as financial. The realization was only possible due to the emergent disagreement of rationalities via the HITL process. Thus, AI approaches can both conceal rationality and reveal novel ways of rationalizing that were not available with decision-making that is unmediated by AI.

Second, for enabling–constraining, the differences in rationality are what first revealed its existence. This enabled the development and refinement of the rationalities with which the HITL approach could be implemented for improving the AI tool. Previously agreed-upon labels came under scrutiny as HITLs refined their understandings of the precise task at hand, which led to a more comprehensive understanding of the task, including strengths and points of improvement. After feeding back the revised labels along with the constructed formal rationality shared by the HITLs, it became possible to assess whether the AI output labels were in line with the constructed rationality and points of difference. While the rationality was constructed and revealed in the process, it was also constrained by formalization. Labeling became more abstract, and adherence to rules that did not seem intuitive emerged. For instance, a text that described the opening of a new facility became labeled as an informational action because it included “says” or “announces,” (e.g., “The company announced the construction of new wind farms”).

Third, in the involving–alienating dimension, the HITLs were assigned the task of improving the AI tool. Thus, they were assigned to be involved, but rationality was not an assigned component of that involvement: The emphasis on creation and formalization of rationality was a spontaneous outcome of improvement. Here, HITLs were involved in attending to the labels and data, as well as how the algorithms reflected the rationalities underpinning the data and labels. Indeed, HITL AI as “augmentation” was involving rather than alienating compared to the “automation” view of publication II. Despite similar things remaining unavailable, such

as the original label rationality and the concealment of the algorithms' rationality, the HITLs actively worked with and made changes to the decision-making process. Thus, the involvement revealed *some* aspects of the AI rationality that would have otherwise remained concealed.

In summary, rationality was revealed by employing HITLs, but they were HITLs with differing substantive rationalities of the task. The dialogue between these differing rationalities was what revealed rationality to exist and enabled its formalization and use as an improvement for the AI. Here, the lack of transparency of the AI methods persisted, and the methods remained concealed, but it was still possible to assess whether the AI output labels reflected the HITL's created and revised rationality. Thus, the rationality revealed via dialogue enabled the evaluation of AI rationality on a level that would otherwise have not been attainable. In the process, people began to reflect critically on what, why, and how they know: Rationality. In the context of AI decision-making, this critical reflection would not have been translated into revealing of the AI output rationality without deliberate efforts to do so.

4.4 Publication IV: Exploring the relationships between artificial intelligence transparency, sources of bias, and types of rationality

AI decision-making is often narrated as objective, but it essentially transforms human-originated substantively rational label data into a means–end optimization task. This process is a suppression and presentation of substantive rationality as formal rationality. It is seen as posing threats to possibilities for the freedom of choice in which pluralities of rationalities are concealed under the guise of formal rationality (Lindebaum et al., 2020). Moreover, on an organizational level, groups lose unique human knowledge when advised by such formal rationalities (Fügenger et al., 2021; Keding & Meissner, 2021). This can change, if, for instance, AI advice is considered to be outright bad (Prahl & Van Swol, 2017). Thus, in overrelying on AI, people rely on merged substantive rationalities that were formalized via AI.

The fourth publication set out to tackle a lack of empirical research into the type of rationality employed in AI decision-making. It extend the findings of publications II and III and used a similar methodology in which labels with some possibilities for subjective interpretability are chosen by people during a decision-making task.

Thus, the technological mediation dimensions of publications II and III can be compared with publication IV. The publication studied and qualitatively compared the rationalities of people with decision-making guided by other people versus decision-making guided by AI. Thus, the publication expanded the findings of publication III, in which differing substantive rationalities revealed rationality, and publication II, in which AI labels and rationality guided people in AI decision-making. Differing from publications III and II, the substantive rationality for the teaching labels was available to people in terms of how they had themselves originally coded the labels. However, they had no access to the rationalities of other people's or the AI output labels.

First, regarding the revealing–concealing dimension, in the human-advised group, the formation of formal rationality for the decision-making task was revealed, with decision-makers giving explicit explanations for their decisions. These explanations of their rationality construction were in dialogue with the task's source documents. Here, substantive rationality was transformed not into a purely formal rationality, but a substantive rationality aimed at formalization by referring back to source documents for the task. In the AI group, by comparison, the formation of this type of hybrid rationality was concealed, which created a vacuum of information about the decision rationality. The subjectivity of the rationality for decision-making was not concealed in either group, and everyone highlighted their subjectivities with phrases like “I feel” and “I think.” Moreover, in both groups, the bases for other people's creations of rationality were concealed: The human group did not know the rationality for other people creating their rationalities, and the AI group did not know what rationality AI based its decision on.

In the enabling–constraining dimension, transparency emerged as an influential factor for both groups. In the AI group, the lack of knowledge about the data the AI had access to resulted in decision-makers applying spontaneous assumptions about what the AI did have access to. For instance, it was assumed that the AI had access to the complete news articles rather than just the snippet available to the human decision-maker. This allowed for rationalities like, “I'm leaning on the AI here” for choosing a certain label. In these cases, the decision rationality was that AI was assumed to have more knowledge and was therefore seen as an authority. In the human group, similar spontaneous assumptions about the better subject matter knowledge of other labelers enabled similar authority rationalities. Both cases were

enabled by the lack of transparency into what information and data were available for creating label recommendations. Moreover, the authority of the AI, specifically, was said to feel constraining to the decision-makers: One decision-maker said they wanted to choose differently from the AI, but felt like they could not. Similarly, at the end of the decision-making task, another label decision-maker said they had avoided looking at the AI label suggestions because they believed they would have felt pressure to answer similarly. The presence of AI, and especially its authority, was perceived as something that constrained decision-making and associated rationality.

Perceived AI authority was a significant factor in the involving–alienating dimension. In the AI group, a defensive rationality emerged that was not present in the human group. For instance, decision-makers said they were “sticking” with a label as if it was contested by the AI. Conversely, in the human group, decision-makers offered their rationalities with a conversational approach to differences, using terms such as “rather, this is still” after offering their own reasons. Thus, the presence of AI alienated decision-makers from participating in the discursive creation and formation of rationality, which contributed to concealment of the rationality processes. A vital detail here is that AI did not alienate anyone from the choosing or deciding processes. Freedom was acknowledged, but people were alienated from the creation of the rationality and for the task: In this sense, alienation and concealment of the rationality were closely related.

In summary, in the AI-guided decision-making group, the formation of formal rationality from substantive rationalities conversing with the original task documentation was not revealed. Thus, the rationality for the task remained concealed in this group. The authority given to AI felt constraining to decision-makers, who defended their choices but did not justify or offer their rationalities the human guided group. Moreover, in all instances the role of transparency into the recommendation for the decision-maker became significant: An opaque AI does not offer its own rationality or reasons. Thus, it is not expected that people would start to converse or negotiate with AI as to how the task should be done, unlike with other people. Here, however, the other people’s rationales and reasons were similarly inaccessible for their recommendations, but negotiation and discussions were still initiated by the decision-makers. However, the lack of access into the original rationalities for the recommended labels, while being discussed with, constrained the conclusions and points made in the attempts to create a shared rationality.

4.5 Publication V: Artificial intelligence in the quest for the end of choice: Black boxes as Sartrean bad faith

The types of rationality employed in AI-assisted decision-making and their implications have been prevalent in recent discussions in management (Balasubramanian et al., 2022; Lindebaum et al., 2020; Moser et al., 2022a). AI adoption stems from a rationalist epistemology (Shneiderman, 2022, p. 18). It is assumed to have objectivity and a lack of bias (Claudy et al., 2022; Keding & Meissner, 2021; Parry et al., 2016; Sundar, n.d.), so it is used in a quest for (formal) rationality (Lindebaum et al., 2020). However, as already highlighted in publication IV, AI merely formalizes the substantively rational data that is prone to human bias into universal means–end calculations. Thus, AI has no access to universal truths or purer objectivity in its rationality, but it is still employed to try to achieve such a goal. Contemporary studies warn against the potential unintended consequences of such a project but overlook the potential intentionality of this formalization process that is seen as threatening choice. In other words, the current literature does not question whether the unintended consequences it warns against are unintended in the first place.

The purpose of the final publication was to direct conversation and attention to not only how AI is adopted for a rationality that can result in constrained choice, but also how AI is adopted *for* constrained choice. The fifth publication is a theoretical paper that employed the phenomenological concept of bad faith from the French existentialist philosopher Jean-Paul Sartre. Bad faith is one way that people conceal and deny their freedom of choice to escape from having to choose. The fifth publication takes a theoretical approach and illustrates how in decision-making, AI functions as bad faith. Thus, the publication argues that AI is used to escape choice—making an end to choice not an entirely unintended consequence. The publication brings forward intentional rationality concealment as an enabling attribute with regard to the bad faith role of AI. Moreover, the publication studies the conditions for the concealing of decision rationality that is provided by the lack of AI transparency. Because it starts from the premise that alienation from substantive rationality is aspirational, it contributes to all subquestions.

Regarding the revealing–concealing dimension, the publication argues that AI decision-making has a motivation to conceal rationalities for decisions—particularly the substantive rationality involved in decision-making. In a bad faith manner, when

we do not want to make a decision for some reason, such as regret aversion or fear of potential responsibility and consequences for not making the “correct” choice, we try to pretend we did not have a choice to make. Here, AI conceals the possibilities of other choices by transforming the plurality of substantive rationality into constructed formal rationality as the objective, means–end optimal, correct choice. The decision-maker can more easily ignore the substantive rationalities to choose differently if AI rationality is seen as better. Moreover, because AI rationality is often intentionally black-boxed, as illustrated in this publication, the possibilities for the decision-maker to scrutinize or see points of disagreement with the AI rationality are concealed. Thus, the publication argues a double concealment of rationality. Both the possibility of multiple rationalities and the rationality used in the AI decision are concealed. The concealment of both is necessary for either to exist, because if one were not concealed, the other would be revealed.

In the enabling–constraining dimension, the role of transparency is significant. The intentional black-boxness moves the rationality used by the AI for the decision away from potential scrutiny and understanding. This allows the AI decision to appear superior or more objective as long as these impressions of it are not threatened by perceived flaws. Such views of AI results as objective and superior rest on common narratives about AI’s capabilities (Keding & Meissner, 2021; Lindebaum et al., 2020; Parry et al., 2016; Sundar, n.d.), and the narratives paint AI in a mystical light that can be seen as equivalent to AI reading us truths from the stars or tea leaves (Moser et al., 2022b) that would otherwise remain outside of our bounded—substantive—rationality. Thus, once again, each is required for the existence of the other. Narratives about AI’s superiority rely on us not perceiving their faults, and believing them to be faultless relies on the narratives. Thus, the concealment permits us to believe rationality exists where it is concealed and constrains us from accessing it.

In the involving–alienating dimension, the publication begins with the premise that alienation from substantive rationality can be the intent of AI adoption. It posits that AI can function as bad faith to alienate the decision-maker from the decision-making process. This alienation happens through the dimensions just reviewed. Decision-makers can become alienated from the process by having the possibility of involvement and choosing differently on various premises concealed from them. This concealment is possible only due to involved creation of opacity which con-

strains the decision-maker from judging, changing, or improving the rationality of the AI decision, maintaining alienation.

The fifth publication illustrates how AI is used to avoid and actively become alienated from the decision-making process through the dimensions of technological mediation that are related to each other in creating this alienation. Becoming alienated from the decision-making rationality happens by concealment of the AI rationality, as well as concealment of the possibility of using a rationality other than the AI's. The former happens through sometimes intentional concealment of the rationality used for the AI decisions. The publication gives multiple examples to demonstrate this unnecessary but enforced black-box nature of AI. The latter is seen to stem from perceiving AI as possessing superior rationality or even mystical qualities that give it authority in decision rationality over humans' "bounded" rationality. Studies in the literature mostly attribute this to narratives of AI as a superior, objective rationality. All these relationships are dependent on each other for bad faith to exist. Choice of a rationality different from the AI rationality and the concealment of AI rationality require each other. Moreover, the narratives rely on concealment for people to continue seeing AI as mystical, and opacity requires mysticism for people to believe that the superior rationality exists.

5 DISCUSSION

This chapter begins with coverage of the studied mediations per each dimension corresponding to posed research subquestions. This is followed up with synthesis of the mediations and the relationships between them. Due to the already discursive nature of this synthesis between the posed research subquestions, discussion with other literature is focused on only later on in the chapter. This is to ensure clarity and the full development of the ideas of the synthesis before relating them to larger contexts. The earlier sections will *mostly* focus on the empirical publications I–IV. While the theoretical publication V is relevant to all mediation dimensions, it connects to discussion only well in coverage of the relationships of the dimensions to each other.

5.1 Technological mediation: Revealing–concealing

This chapter discusses the answer to the dissertation’s first posed research subquestion: “How does AI in decision-making mediate the revealing–concealing of rationality?” This subquestion was addressed in all publications. The discussion takes the premise of the current literature, in which the formal rationality employed by AI is problematized in decision-making, and compares it to substantive rationality—the transformation of substantive rationality into formal means–end calculations (Balasubramanian et al., 2022; Lindebaum et al., 2020; Moser et al., 2022a). The concealment of rationality is important to AI discussions around XAI transparency, fairness, and accountability, and associated demands for AI decisions that can be explained, understood, and audited for social responsibility (T. Miller, 2019; O’Neil, 2016; Rai, 2020).

Publications I–IV all found AI in decision-making to have a dimension of concealing with regard to rationality, in both UML and SML. In UML, this was present in the concealment of decisions made in the AI implementation processes as well as

the data used in them. In SML, the rationality used for the decisions on the original teaching labels was concealed from further points in the AI decision-making process in publications II–IV. However, in UML, the rationality for selecting an AI process is concealed, whereas in SML, it is often less so. The best quantitative accuracy or similar measure is used to compare a wide variety of methods, and the best performing one is selected. However, the concealment of decision-making rationality in UML processes is not inherent, and not revealing it is a decision made in itself. Moreover, optimizing for accuracy in SML is a decision, but the rationality for it is concealed. For instance, why is accuracy chosen instead of required computational capacity? Publication IV found that AI concealed rationality for decision-making in label selection, as compared to decision-making without AI. Specifically, with AI in decision-making, when explaining the rationality for their decisions, people omitted relevant presuppositions (Overton, 2011) and explanations of rationality formalizations based on interacting with task materials.

In all publications, rationality was not something that was looked for or searched unless contrasting happened with either output labels and the original teaching labels, or between different people or AI decisions. This suggests, in accordance with T. Miller (2019), that explanations and rationalities are always in relation to “why this and not something else?” The existence of a rationality was revealed in these moments of contrast. Furthermore, publications III and IV posit that what a rationality is *and could become* is revealed in dialogue—a process of contrasts—and compromise while it is being built by a human decision-maker. This, however, did not apply to the original teaching labels that had been labeled by someone other than the decision-makers in the tasks for these studies: Original label rationality remained concealed. However, in publications I and III, the revealed existence of a rationality in the first place prompted some guesses as to what these “original” rationalities could possibly be. In other words, the *possibilities of rationalities* were revealed, but their contents remained concealed. Hence, there are levels of concealment: At the base level is the concealment that a rationality exists in the first place, which is the presentation of a decision as self-evident, or what Vesa and Tienari (2022) referred to as “ideology.” At the second level of concealment, it is revealed a rationality exists but its contents remain concealed in any meaningful level of detail.

Publications III and IV go into further detail than publication II and show that the formalization process of the decision-making rationality from substantive ra-

tionality to formalized is revealed in dialogue and explanation in comparison to another (concealed or revealed) rationality. Thus, formalization and means–end abstractions—happened even without AI, but AI concealed this process. In publication II, subjectivity remained concealed even when looking at the AI “mistakes” that reveal multiple possible “correct” labels. In assessing the mistakes, decision-makers took note that AI could be “more correct” than the original labels, but did not consider the possibility of multiple, substantively correct labels. Instead, they focused on improving the formal rationality of the AI. The fact that the AI was taught on substantive rationalities remained concealed. Indeed, formalization happened regardless of whether AI was involved, and instead of the core issue being the formal rationality’s use for decision-making (Lindebaum et al., 2020; Moser et al., 2022a), the concealment of rationality formation from substantive to formal was brought forward.

In comparison, in publication III, the dialogue to find compromise for a decision between two substantively rational humans revealed the formalization process of the rationality. As an example, say that people agree to put all data containing a certain word into a specific category and establish a formal rule. The concealment of AI rationality in the algorithm and original teaching labels persists, but with the formalized rationality of the HITLs, it could be assessed whether the AI outputs and original labels reflected the formalized rationality or not. This revealed AI rationality on a level not otherwise attainable. These results support prior studies (Lebovitz et al., 2021), in which it was noticed that in the process, people began to reflect critically on what, why, and how they know—rationality. In the AI decision-making process, this critical reflection would have not translated to the explainability of the AI outputs without deliberate efforts made to do so. Moreover, dialogue and reflection—involvement—revealed completely new possibilities of rationalization that did not previously exist in any label, AI output, or substantive rationality.

Publication IV supports this finding and adds that, while the compromise of rationality led to formalization with reference to other points of formalization, such as documents, AI concealed the formalization process. The revealed substantive quality of people’s rationalities remained with AI, but the process of how the rationality is and was formalized was concealed with the AI decision “partner.” Thus, concern about the formalization of rationality with AI (Balasubramanian et al., 2022; Lindebaum et al., 2020; Moser et al., 2022a) might be missing a level of detail in

the problematization of AI use. Formalization of rationality was present with or without AI, making the problematized aspect not specific to AI decision-making. However, the concealment of the formalization was present only in AI, which can have detrimental effects on explainability. In general, in AI decision-making, the final decision is highlighted, but the process for reaching that decision was concealed.

5.2 Technological mediation: Enabling–constraining

This subchapter discusses how this dissertation answers the second posed research subquestion: “How does AI in decision-making mediate the enabling–constraining of rationality?” This subquestion was studied in all of the publications, and the results are covered in this section.

In both SML and UML, as noted in publication I, the form of data and labeling for an AI decision-making task constrains the ability to rationalize about differences or similarities. This is reminiscent of how science was mainly conducted during the Renaissance (Foucault, 2005). In UML, AI often tries to group together data points that are similar, while separating the similar within a group from other groups. In SML, AI is often optimized to find similarities within the data grouped under a certain label, and the groups are compared to find differences to other labels. Thus, both force a constraint on what types of rationalities can emerge, but UML has a level of freedom compared to SML in that the labels do not have to be predefined for groups (Ziegler, 2012). Moreover, SML often constrains usable data to ready-labeled datasets that are constrained and structured because creating a new set of categorization for labeling may be impossible in terms of labor resources (Kobayashi et al., 2018; Kuang et al., 2015). Within the constraint of similarity criteria for UML and SML, people in the processes making judgments and reflecting enabled different rationalities to be created and explained.

In publications II–IV, formalization was a constraint, whether through AI or humans co-creating a formalization by compromising among themselves or with reference to documentation material. In publication II, the formalization happened with the SML algorithms used, in which the performance of the algorithm was a part of the rationality for its choice, as well as in the formalization the algorithm then performs. However, in publications III and IV, the dialogue and compromise between human decision-makers was formalizing and constraining: certain rules were set in

place about the rationality even if a variety of rationalities were sensible. Between publications III and IV, the main difference was that in III, there was no documentation or suggested decisions, but the dialogue happened between two substantively rational people who had access to AI output and original labels. In publication IV, people discussed against and between already made decisions, making this process resemble more an abductive inference (T. Miller, 2019) of rationality rather than following from its emergence. However, in both cases, the formalizations of rationality required abstraction beyond singular decisions per data point, for which many rationalities would have been sensible. This led to decision rationalities that would suit multiple datapoints but that frequently became counterintuitive. Thus, formalization is constraining with or without AI, but AI constrains access to the formalization and formalized rationality. Details of the enabling–constraining dimension’s relationship to the revealing–concealing dimension are covered in Chapter 5.4.1.

In publication IV, it was revealed that some decision-makers found seeing AI decisions for their task constraining: If they looked at the AI decision, they felt they could not choose differently from the AI suggestion. This constrains free rationalization, but also allows a specific, different type of rationality to emerge: relying on perceived authority. In other words, this constraint on rationality enables the rationality “I choose this because AI said so.” A similar type of rhetoric emerged, but to a lesser degree, in the group where decision-makers were shown other people’s decisions. In accordance with prior research (e.g., Elson et al., 2018), this reliance reasoning appeared more in cases of uncertainty.

5.3 Technological mediation: Involving–alienating

This chapter discusses how this dissertation answers the third posed research subquestion: “How does AI in decision-making mediate the involving–alienating of rationality?” This subquestion was mainly studied in publications II, III, IV, and V and the findings are covered in this section.

While not key for answering this subquestion, a minor related point from publication I is that the reader of AI research in UML literature was alienated from the research decision process. This mirrored SML in publication II in how “self-evident” accuracy optimization alienated the decision-maker from the decisions made and the rationality of the methodology in the first place. It does not matter how or why

the methodological choices are made as long as the key measure is optimized. This externalizes the goal to outside of the user/developer, who is alienated from the rationality of the task at hand. Similarly, the use of readily labeled data makes the task appear to be already given, and there is no need to concern oneself with the hows or whys. A clear example of such alienation is van Rijmenam and Logue (2021), who did not see human involvement in AI decision-making: Goals set for AI tasks appear as givens, without social connections.

However, in publication III, the HITL approach was involving rather than alienating. HITL is inherently an active involvement with the AI process in order to improve it, but here, it was notable that the involvement of the HITLs was mostly with the decision rationality. Indeed, it was the involvement with the rationality itself that aided with explainability, not HITL as a general manifestation of the “augmentation” heralded as the cure-all for AI issues (Arrieta et al., 2020; Haque et al., 2023; Krügel et al., 2023). In publication IV, people were not alienated from making decisions, but they were alienated from decision rationality with AI-aided decisions compared to human-aided ones. In publication III, the HITLs began reflecting on and making guesses as to whether the AI was following the rationality they created for the task. Thus, they became involved with the AI rationality and felt they could participate and guess on it. The results of publication III prompt the question whether the user in the case of publication II would become more involved in the decision-making process if tasked with making improvements to the tool.

Publication IV results showed that people engaged with rationality formation more with other humans even if the other people hid behind decision suggestions on paper. In contrast, decision-makers withheld their participation and activity and instead simply stood their ground passively with AI guidance. This supports the difference between publications II and III regarding the level of involvement. The finding is in line with prior studies that found people interacted and participated less in decision-making when interacting with machines (Amalberti et al., 1993; Shaikh & Cruz, 2019).

All together, AI can be involving or alienating depending on how it is implemented. The HITL approach with the goal of decision-making improvement for AI was more involving than alienating compared to an automation approach without human participation in the decision-making. However, referring to AI as automated when, in reality, it is often impossible to have no augmentation aspect, only serves

to alienate people from the process of which they are inherently a part (Raisch & Krakowski, 2021). Publication V makes the argument that AI decision-making is intentional alienation from choice and suggests a reason for the automation narratives to be intentional alienation, which happens, as suggested by all publications, through alienation from the rationality. However, how that alienation plays out occurs in relation to the other mediation dimensions considered. Thus, this will be expanded upon in the next chapters.

5.4 Relationships of technological mediation dimensions

This section of the chapter discusses how this dissertation answers the posed overarching research question, “How does AI mediate rationality in decision-making?” The answer to this question as by this dissertation is structured with the research subquestions mediation dimension relationships as follows: First, how AI mediates rationality in decision-making by the relationship of the revealing-concealing and enabling-constraining mediation dimensions. Second, how AI mediates rationality in decision-making by the relationship of the enabling-constraining and involving-alienating mediation dimensions. Third, how AI mediates rationality in decision-making by the relationship of the involving-alienating and revealing-concealing mediation dimensions. Lastly, it is discussed how the interplay of all the considered dimensions mediates rationality when AI is used in decision-making.

5.4.1 Revealing–concealing and enabling–constraining

The constraining of rationality happened in abstraction and formalization, particularly in publications III and IV, but that constraint was also revealing when people discussed and contrasted rationalities. However, formalization with AI was concealing in comparison to revealing human formalization, as suggested by both publications II and IV. The existence of differing rationalities was not concealed in AI formalization-constraining. In the publications concerning SML, despite the revealing of a rationality’s existence, AI black-box algorithms did not reveal what that rationality was and, thus, constrained possibilities to question or expand on it.

Regardless, this constrainin is not a complete disability of questioning, and in publications II and III, the revealed existence and differences in rationality—despite not knowing the rationality—enabled the development and refinement of rationality. In

publication III, previously agreed-upon label classification came under scrutiny as HITLs refined their understanding of what the exact task at hand was, which led to a more comprehensive understanding of the classification task. After retraining the AI with revised labels based on the constructed shared formal rationality from the HITLs, assessment was enabled both where the AI outputs were in line with the created rationality and where they differed. This conversation between substantive rationalities and AI outputs and original labels enabled new points of view to be revealed by the AI when it differed, and supports the view of Moser et al. (2022a) in which human decision-making informs and steers AI decision-making, and vice versa. Thus, revealing was also enabling. But also, in the reverse, concealed choices—such as in publication I—constrained the access to points in the used rationality from which other rationalities could emerge. Moreover, while the concealment of rationality constrains us from rationalizing differently, it also allows us to assume the level of formality and “goodness” of the AI rationality. Thus, in authoritativeness and concealment, our rationalities can become constrained to “because AI said so.” In other words, concealment allows us to become complacent, which echoes Jarrahi (2019), who discovered that people become detached and ceded decision-making in certain situations of uncertainty.

Indeed, concealing was also enabling of certain types of rationalities to emerge. For instance, “I’m leaning on the AI” and reliance on other people’s assumed better knowledge for the decision were enabled by the concealed rationality for the decisions made by AI and people in publication IV. Thus, the spontaneous assumptions about the rationalities that sparked this authoritative rationalization would have not been possible without concealment. Indeed, in support of Schneider and Leyer (2019), with lower situational awareness or concealed rationality, people are more likely to delegate decision-making to AI.

5.4.2 Enabling–constraining and involving–alienating

The constraining formalization in dialogue to find compromise among humans was actively involving, whereas in AI decision-making, in which no dialogue between humans and substantive rationalities existed, this degree of active involvement and participation was not present. Thus, the constraining by collaborative formalization was involving. Because in publication III, the HITL approach spontaneously led to formalizing rationality, it can be suggested that involvement can have a constraining

mediation, but this requires further research. In publication IV, the AI authority that felt constraining to people with regard to rationalization, which alienated people from free engagement in the construction of rationality as well as subdued active dialogue engagement, supported previous findings (Amalberti et al., 1993; Shaikh & Cruz, 2019). Moreover, in the constraint to pre-existing labeled datasets and their label categorizations in SML, even in exploration and contrasting of different manners of rationality for the decisions, the constraint of pre-existing categories as a given was not engaged with by people in any publication. Thus, the methodological constraints of an AI type of task of similarities and differences is suggested by this dissertation to alienate from engagement with asking whether the type of task is appropriate for the goals of the context or whether the similarity–difference constraint is something that should be looked beyond.

5.4.3 Involving–alienating and revealing–concealing

Publications I–III suggest that the concealment of decision rationality is alienating. On one hand, the revealing of rationality to exist in comparisons and contrasts happened along with involvement in its creation: Differences that were spotted in decisions were compromised and a new rationality was formed with active participation by people. In publication I, the revealing of rationality and putting it under question happened when outputs were contrasted with the data that generated them. In publication II, this was the comparison of original labels with AI output labels. In publication III, this was between two people and the AI.

However, the degrees of involvement are different between contrasting original labels and AI outputs, as in publication II, and substantive rationalities between humans either through interaction or on paper, as in publications III and IV, respectively. In publication IV, people explained and revealed their rationality to other people, but did not offer much to AI except for their final decision. Thus, with AI, the possibility of different rationalities is noted but not actively created, as in publication III between two people conversing with the AI. Here, decision-makers also engaged and challenged the AI decisions and original labels in this process. Future research should address whether AI authority is as prominent when there are other substantive rationalities involved along with support from other people for decision-making. However, it is possible, as suggested by Krügel et al. (2023), that people will act as partners in crime for morally questionable AI decisions. The level

of involvement and number of people involved in the decision-making could change or perpetuate the findings for decision-makers not in groups.

While the sample size for publication IV was small, its positions are echoed in other studies. For example, Elson et al. (2018) found that in situations with either low or high levels of uncertainty, people's decisions complied more with a perceived AI decision recommendation. In this dissertation, the same was seen to happen due to assumed better knowledge, which was caused by concealed rationality. Shaikh and Cruz (2019) and Amalberti et al. (1993) found that, with AI present, people became more alienated and interacted less, while also concealing their thought processes, which are alienation and the concealing of rationality.

Publications I and IV together suggest that alienation is the goal and that it is reached through concealment. However, the concealed—what appears as a given—requires involvement, as in the lack of reporting methodological choices in publication I. Thus, people are actively involved in alienating themselves with AI decision-making. AI is not a new phenomenon in this regard; managers have been known to avoid information to intentionally impair and alienate themselves from fully informed decision-making (Golman et al., 2017). This is addressed by publication V: This involving alienation is bad faith with which people try to ease their inherent discomfort at making choices by actively making an effort to convince themselves there is no choice. This is also posited in publication V as a reason for intentional—involved—use of black-box AI and methodology for AI.

5.4.4 Revealing–concealing, enabling–constraining, and involving–alienating

This section of the chapter covers the interrelations between all of the studied mediation dimensions. It is structured so that inferences from the previous sections covering the relationships between two mediation dimensions are complemented with the insights from the third mediation. Taken together, alienation from the process of rationality formation happened with AI, which created a “vacuum of conversation/interaction” with other potential rationalities from other sources, which were found to be the key to revealing the rationality and its formation in AI decision-making. This is depicted in Figure 5.1. Moreover, this section engages publication V more than the previous ones, as it provides a theoretical background against which the discovered mediations can be contrasted.

The discovered mediation of concealment of rationality and constrained access

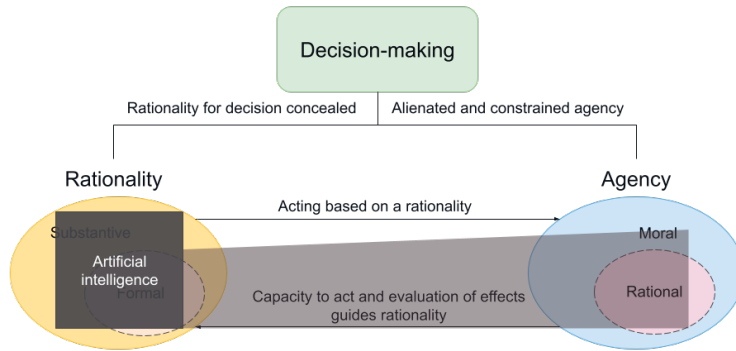


Figure 5.1 Concealment of decision rationality with artificial intelligence decision-making

to rationalities without active human participation in the AI decision-making process was present in both UML and SML. However, the places of concealment and constraint were different. In SML, the constraint and concealment were present in the algorithms used and the rationality for the original teaching labels. In UML, the concealment and constraint were present in undisclosed inherent decision rationalities in the methodologies and the intentional lack of comparison of AI outputs to the original data. This paints AI decision-making as something that simply happens and that hides—alienates—the inherent decision-maker from the process. This concealment of any rationality to exist in the first place is depicted in Figure 5.2. Thus, in UML, the choice to conceal the involvement in concealing the rationality was more notable than in SML. In SML, the black-box nature of algorithms appears as something inscribed into the methods themselves rather than as decisions made about the process. However, Forrest (2021) suggested that preference for AI with concealed rationality is an intentional decision because it is questionable whether there are any benefits to choosing the prevalent concealed rationality AI over others, except for said concealment of rationality. Indeed, the publications suggest the alienation that happens due to rationality concealment can only happen by involvement and active participation in the constraining of rationality and selection of concealing ap-

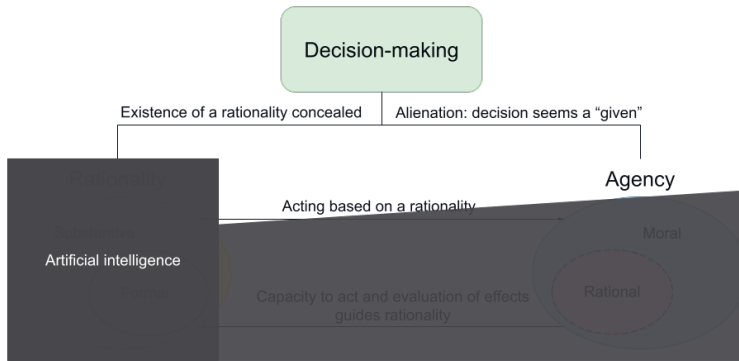


Figure 5.2 Concealment of rationality with artificial intelligence decision-making

proaches to AI decision-making. Thus, the interplay of the mediation dimensions is delicate and complex: Through involvement, we attempt to reach alienation from decision-making rationality. Indeed, with rationality inherently tied to agency, the concealment of rationality and possibilities of different rationalizations can suppress, or conceal our agency. Moser et al. (2022a) posited that AI decision-making and human decision-making co-constitute each other, which is supported by the findings of this dissertation. However, the findings also suggest that people attempt to alienate themselves from this intra-action and aim for either AI guiding humans or people remaining in control and designing AI agency (D. G. Johnson, 2006), as represented respectively by A and B in Figure 5.3.

Publication V provides the theoretical lens for this relationship: Bad faith is an intentional lying to oneself by “the resolution ‘not to ask for too much, to consider itself satisfied when it is poorly persuaded, to force through, by means of a decision, its adherence to uncertain truths,’” meaning that bad faith is “resigned in advance not to be fulfilled by [apprehended evident facts]” (Sartre, 2021, p. 114). This alienation of the self of the decision-maker becomes easier the more the facts are apprehended, which is why the involvement happens through constraining and concealing rationalities as the “facts” the choices and decisions are made on. For instance, being

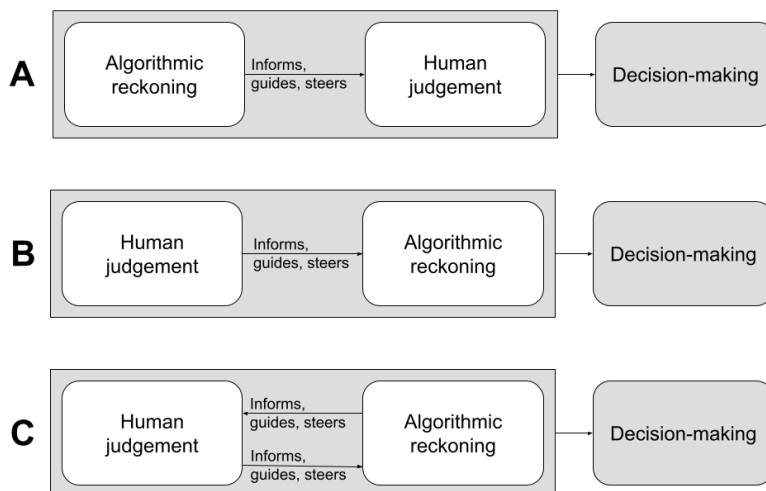


Figure 5.3 Three scenarios of inter- or intra-action between human judgement and artificial intelligence in decision-making (modified from Moser et al., 2022a)

alienated from participating in the rationality for decision-making allows, as considered in publication I, the easier lie to the decision-maker (themselves and others) that they chose this because it is popular—not because it created the outputs they wanted.

The found relationships between rationality formalization-constraint and its revelation were tied to involvement. The formalization of rationality was an involving process of people into the decision-making tasks, and it both constrained and revealed the rationality through that involvement. Thus, the involvement in the decision rationality constrained and revealed said rationality, when it remained concealed, with human alienation. This is depicted in Figure 5.4. However, alienation enabled and revealed alienated rationality that relied on authority, but this was also revealed only through asking people about their rationalities. Thus, the involvement and engagement of people was a prerequisite for this alienated rationality to be discovered. Alienated rationality, however, was not unique to AI involvement in decision-making, but also resulted when decision advice was received from other people in publication IV. It was more prevalent in ambiguous decisions, and it has been noted that people comply more with AI in situations with high levels of uncertainty (Elson et al., 2018).

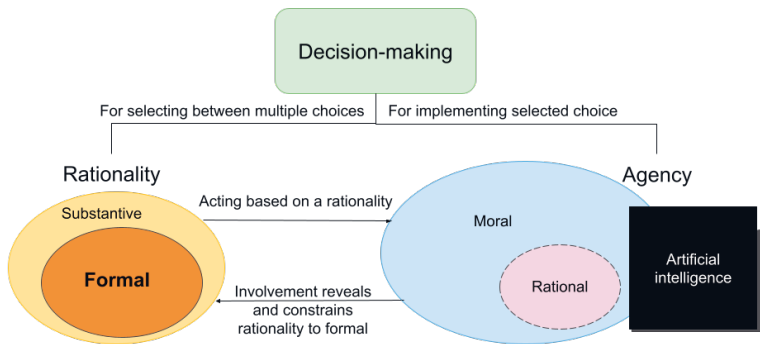


Figure 5.4 Involvement with decision-making rationality with artificial intelligence revealing and constraining rationality to formal

However, while the formalization-constraining of involvement in rationality allowed the revelation of what the rationality was, formalization was not a requirement for the revelation of rationality to exist. This result, however, also required involvement. Publications II and III found that people involved with making comparisons between labels from different decision-making sources revealed the existence of rationality. Despite the involved people not knowing what these rationalities were, they were able to propose development ideas and points of refinement to the rationalities they assumed were there. In the conceptualization of rationality in MOS, AI rationality is not novel, but it appears to be a natural next step. The formal, universal optimization rationality brought into decision-making with AI fits neatly with the trajectory of conversations embedded in bounded rationality (K. D. Miller, 2008). Human decision-making has been seen as bounded and “simplified” in the face of the complexity of the world (Simon, 1957), and it has been assumed that artificial simulation of human rationality could replicate human decision-making and overcome its limitations (Simon & Newell, 1958). AI has been considered able to access unbounded rationality devoid of human simplifications. The belief that human rationality could be separated from the body and put into coded form in ma-

chines demonstrates the basis of this thought in Cartesian mind–body dualism (K. D. Miller, 2008; Patokorpi, 2008). Such assumptions have guided the development of AI as a field (Mabaso, 2021).

The type of detached logic assumed by this influential view of rationality is often an ideal and not how decision-making happens or can happen in practice (Elliot et al., 2020; Y. Li et al., 2014). The workings and details of how rationality behaves with AI decision-making discovered in the studies of this dissertation show that formal rationality and substantive rationality cannot be separated. The substantive rationalities in publications III and IV create formalizations. Formal rationality is justified by substantive rationality. Decisions for the use of formal rationality in algorithms in publication II and leanings on AI formal rationality in publication IV demonstrate that, while the rationality may not be very detailed or conscious of itself, the decision is a substantively rational decision. Indeed, this supports views that emphasize that AI as a pure formal rationality is unrealistic and impractical.

Moreover, AI relies on the premise of yielding decision-making to this type of formal rationality and is often tested in laboratories removed from the contexts of their use (Raisch & Krakowski, 2021). This can be one reason why real-life applications often fall short of expectations (e.g., Fountaine et al., 2019; Lebovitz et al., 2021; Van den Broek et al., 2021) or even cause issues of complacency in decision-making and overreliance on AI (Jarrahi, 2019; Keding & Meissner, 2021). Disappointment is expected because expectations for technological achievements are typically too high to begin with due to overoptimism (Clark et al., 2016; Zaitsava et al., 2022), but this optimism comes with downplaying the details of the dynamics at play (Luoma, 2016). Indeed, people associate new technology with a type of mystery and complexity that sets a premise of overlooking details and forgoing attempts to critically analyze and understand how the technology works “under the hood” (Elsbach & Stigliani, 2019; Vishwanath & LaVail, 2013). To get to the inner workings, the black boxes of AI must be opened (Orlikowski & Scott, 2008). However, the results of this dissertation suggest the opposite: Rationality is concealed and tied to alienation from the decision-making process with AI, in line with Jarrahi (2019), who discovered that people become complacent—alienated.

Conversations and understandings about agency have evolved into more relational, co-constituted agencies between people and technologies intra-acting (Introna, 2014; Orlikowski, 2007; Orlikowski & Scott, 2008). AI has a focal position and is

attributed agency on a level unprecedented for technologies (e.g., Kaplan & Haenlein, 2020; Murray et al., 2021). However, while rationality is inherent in agency and both are inherent in decision-making, such considerations of co-constituted rationality for decision-making are lacking. The concept of tacit knowledge makes a possible exception to this. Tacit knowledge rejects mind–body dualism and sees decision-making rationalities embedded into contexts and actions that escape formalizations (Blackler, 1995; K. D. Miller, 2008; Polanyi, 2012).

However, tacit knowledge is prone to being bastardized into the mind-body dualism: Tsoukas (2005) narrates how tacit knowledge is often seen as something located in the head—the mind—of a practitioner that needs to be extracted into explicit knowledge, creating a dichotomy out of tacit and codifiable knowledge instead of interdependent them as dimensions of knowledge. Unsurprisingly, due to its origins in the dualism, AI is now used as something that could, better than people, translate tacit knowledge into explicit (for example, see, X. Li et al., 2023). In opposition to this, the results of this dissertation show that rationality, like and *with* agency, is co-constituted in the active intra-action of humans and technology. Thus, it both forms like tacit knowledge and participates in the formation of tacit knowledge. Alienating intra-action results in and is a cause of concealment of rationality that exists nonetheless and becomes more tacit, whereas the involvement of intra-action—tied to rationality—increased explainability, which can be seen as the more codifiable dimension of knowledge.

Regarding explainability, in publications III and IV, the involvement of people in changing and accessing rationalities was the difference in whether the rationalization process could be followed as it happened or only accessed through abductive inferences later on upon questioning. While AI concealed rationality in its progression and as explanations looking back, in the explanations in particular, presuppositions were also concealed. This demonstrates details and workings of how rationality becomes concealed depending on the type of involvement of people in the AI process—whether they look back or are involved in the creation of rationality. This difference is relevant for HITL approaches that study how to ensure AI transparency, fairness, and accountability (Arrieta et al., 2020; Binns et al., 2018; Haque et al., 2023; Shin, 2020; Teodorescu et al., 2021). As noted by Friedrich et al. (2022), a study of mediation dimensions overcomes a dichotomy of explainable versus unexplainable, and tackles AI in decision-making to understand the particulars of how technology me-

diates the situation. Moreover, for XAI, it is not only explanations that are relevant, but also fundamentally how AI mediates rationality for decisions. Especially interesting is consideration of intentional concealment and alienation from rationality as a way to trust AI and ignore its possible flaws, while simultaneously, the revealing of these is seen as a premise for trust (Baum et al., 2011; Hasan et al., 2021). To trust AI, we do not want to see its biases and issues, but to trust AI, we nonetheless need to see them. We only want to see that AI has no flaws, but because that is unlikely, we choose to not see.

These findings prompt some “why” questions. Why are people alienated from rationality with AI in decision-making? Why are people involved in becoming alienated from rationality in decision-making? It could be argued that one goal of AI is to let people tend to other matters—a question of the use of resources. However, that does not answer the tension of calls after XAI (Haque et al., 2023; Saeed & Omlin, 2023) in unison with our rejection of it (Forrest, 2021). A potential answer comes from the mind–body dualism embedded in AI development. Applying the dualistic view to machines gives them a “mind” separate from their physicality. Simon’s paradox is created by seeing human rationality as bounded and imperfect, but also seeing the possibility of the rationality of machines made by humans to human goals as unbounded (Patokorpi, 2008). By concealing rationality, the paradox does not need to be confronted and the boundedness remains concealed (Patokorpi, 2008; Zeleny, 2001). The belief in unbounded rationality can be maintained by beliefs in mysterious, alien qualities of technology (Elsbach & Stigliani, 2019), and concealing any boundedness and embodiedness becomes a relevant task. If revealed, we may see the nooks and crannies of the embodied reckoning happening from the data and its faults. Moreover, the lack of autonomy of AI can be concealed by concealing the rationality. If rules are not seen to originate from human-generated data, they can be considered detached from human origins, and thus autonomy becomes attributable to AI (Castro-Manzano, 2010). The concealed human origin and design lets artifacts become “autonomous” and removed from human goals (Kibble, 2017). All together, the concealment of rationality maintains the view that AI has a mind that is required for autonomy. The prevalence of a mind stems from Descartes’ philosophy of the mind, which is still the most frequently used among people not educated in philosophy (Kenny, 2018, pp. 206–207). This theory of the mind comes with the baggage of seeing the mind as something in the image of God—the body, not

quite so much (Kenny, 2018, p. 201). AI research and development rarely employs philosophy (T. Miller, 2019).

While AI has business benefits, but also pitfalls, the consequences of employing optimization and reckoning in decision-making—or in other words, increasing the amount of formal rationality in decision-making—is seen as posing consequences for the morality of decision-making. Other schools of thought see AI as an improvement to morality due to AI’s “objectivity” (Martinho et al., 2021), while others see AI as detrimental to moral choice and a possible route for encoding unwanted and unreflective morality into decision-making (e.g., Lindebaum et al., 2020; Moser et al., 2022a). Both rely on a distinction between formal and substantive rationality with one having the advantage over the other, but formal rationality and substantive rationality are both always present. Hence, the type of rationality may be a false dichotomy upon which to build studies of AI’s impacts. This dissertation shows that concealment, constraint, and alienation from rationality provide a more nuanced and detailed view of the issue regarding AI morality: Concealment and alienation of rationality allow us to forget that AI rationality may be prone to the same issues in its data as those who created that data, or to conceal the formalization of substantive data to the extent that “substantive” is forgotten. Indeed, two levels of concealment were identified: One in which it was concealed that a rationality was present in the first place, presenting decisions as givens, and another in which the details of a revealed rationality are concealed.

The end of choice is happening, not due to the formalization of rationality, but due to the alienation and concealment of rationalized choice that is presented as a given. Vesa and Tienari (2022) saw AI painted as a tool of inevitability and people being coerced into having *faith in its rationality*. Moreover, they highlighted AI as a tool for creating self-evident truths—to establish AI rationality as an ideology. While Vesa and Tienari (2022) covered this at a higher level, this dissertation offers an answer to how this ideology arises: The presence of a rationality is concealed and people are alienated from the creation of rationality. Concealment and alienation affect and constrain human agency by concealing choice. The concealment of rationality is alienating. On one hand, the revealing of rationality as existing in the first place required contrasted involvement: Differences that were spotted in decisions were points to notice where different decisions—rationalities—were possible. Rationalities were formed in active participation by people in the formation of the

rationalities. The results of the mediations highlighted that they happen throughout the AI decision-making process—not only with the algorithms themselves. The choice to employ AI, which type of AI to employ, how the data are chosen, how it is optimized, how it is analyzed, and how it is used are all decisions within AI decision-making that influentially mediate the entire process of when and where the rationalities are concealed–revealed, constrained–enabled, or alienated–involved.

6 CONCLUSIONS

6.1 Summary of findings

The main research question for this dissertation was “How does AI mediate rationality in decision-making?” The main research question was split into three subquestions according to three identified dimensions of mediation. Technological mediation is a core concept in the methodology of postphenomenology that was used, which centers specific technologies as the starting points for study. The first subquestion was “How does AI in decision-making mediate the revealing–concealing of rationality?” All of the publications found that AI concealed decision rationality at various stages of AI decision-making, while some publications discovered AI also revealed specific types of possibilities for novel rationalities. Moreover, two levels on concealment were discovered: The contents of a rationality could become concealed, but the presence of a rationality in the first place could also be concealed.

Regarding the second subquestion, “How does AI in decision-making mediate the enabling–constraining of rationality?” publications II–IV all found a transformation of rationality from substantive to formal regardless of whether the rationality was constructed with AI. This formalization was constraining because it crowded out the possible plurality of other valid rationalities. Publications I–IV, covering both SML and UML, discovered a constraining mediation due to the nature of AI decision-making tasks: classification. Rationalities necessarily take the form of searching for similarities and differences in the data, which limits other potential rationalities. However, these constraints of AI enable the benefits of data analysis in quantities only approachable by AI (Cao & Duan, 2017; McAfee et al., 2012; Olabode et al., 2022).

The third posed subquestion was “How does AI in decision-making mediate the involving–alienating of rationality?” Publications I–II and IV found alienating mediations in AI decision-making. Alienation happened in UML and SML through

overlooking decisions in the first place, by accepting the tasks of optimization and the given label categories as arbitrary truths—something to avoid involvement with. Publication IV found AI to have an authority that alienated people from engaging in the rationality formalization process, and publication V theoretically argued that alienation from decision-making is the goal of AI decision-making. Taken together, the results suggest that people can be involved in their alienation from rationality in AI decision-making.

To answer the main research question, the interplay and relationships between the dimensions emerged as important factors. In SML and UML, concealment of rationality constrained the use of other possible rationalities. Conversely, apparent differences in decisions brought forward the possibilities of different rationalities in publications I, II, and IV. Hence, involvement with the decision-making revealed both that a rationality existed (publications I–II) and the contents and processes of those rationalities (publications III–IV). Publications I–IV suggest that the constraint of formalization was revealing with involvement. In other cases, the constraint of formalization was concealed because of AI in decision-making and also resulted in alienation from it. Publication V suggested a theoretical framework of Sartrean bad faith to explain why the concealment of decision rationality is aspired to. Why be involved in intentional alienation via rationality concealment? Kiran (2015) saw that with modern technology, revealing leads to alienation in which we forget to doubt, we *forget* “things, the world, could be revealed in different manners” (Kiran, 2015, p. 128). This echoes Vesa and Tienari (2022), who saw AI as creating an ideology: self-evident truths that we *forget* to question. The discoveries of alienation through concealment of rationalities in this dissertation paints a detailed picture of *how* we forget.

6.2 Contribution to theory

This dissertation has three main contributions to theory. First, it provides new insights into and levels of analysis for conversations surrounding AI moral agency and the moral panic over AI’s formal rationality in decision-making (e.g., Lindebaum et al., 2020; Moser et al., 2022a). Second, it contributes to XAI theory by emphasizing the importance of rationality considerations and differentiates them from explainability. Moreover, the dissertation provides testable claims about the techno-

logical mediation relationships that can be used to develop theory on explainability. Third, the dissertation contributes to theory on rationality and agency in organizational decision-making. Agency of technology, especially AI, has been an issue in the field, with AI being attributed very different degrees of agency. However, while rationality is a key component of agency, such considerations have largely ignored the role of rationality in AI agency. This dissertation shows that the research methodologies currently used to contend with relative artifact agency are applicable to rationality considerations as well.

For the first contribution regarding the discussions around AI impacts on morality in decision-making, this dissertation found that the contrast of formal and substantive rationality may not catch important nuances. Lindebaum et al. (2020) characterized AI as a *supercarrier of formal rationality* because it is exceptional at applying formal rationality in quantities and at velocity. While it is true that AI applies formal rationality on a different scale than other types of decision-making, this dissertation's results indicate that the conceptualization of formal rationality as itself the issue may lack important details. The concealment of and alienation from the used rationality created a lack of choice. Indeed, when it was not apparent that alternative rationalities were possible, the choices made were concealed, and people often became alienated from the decision-making. In contrast, when the formal rationality was revealed, people actively involved themselves in making decisions and perceived corrections to the formal rationality (publications III–IV).

Moreover, this dissertation found levels of rationality concealment relevant to an end of choice: The existence of a rationality for a decision to exist in the first place alienated people and made decisions seem given and self-evident. This provides a level of concrete detail to the considerations from Vesa and Tienari (2022), who saw AI as capable of turning into unquestioned ideology. Moreover, publication V directs theoretical discussions to consider the involvement in alienation from decision-making with AI, whereas previous considerations addressed alienation only.

Regarding the second theoretical contribution, the dissertation brings a focus on rationality mediations as something to consider in XAI. It posits that technological mediations can be useful in theorizing how to achieve the goals associated with XAI, such as fairness, transparency, and accountability (Shin & Park, 2019). For instance, in accordance with the findings of this dissertation, Binns et al. (2018) found that comparisons of rationalities reveals differences, which made people react strongly

to perceived issues in fairness. In other words, revealing involved people in fairness judgments of the rationality. This dissertation posits more complex relationships between technological mediations that can be studied to develop theory in their effects on XAI.

For the third theoretical contribution on agency and rationality, this dissertation helps connect previously separate streams of research: artifact agency and AI decision rationality. While artifact agency theory has developed with sociomateriality and related agential views (den Hond & Moser, 2023; Orlikowski, 2007; Orlikowski & Scott, 2008), the theory of AI rationality remains embedded in the dualism (Mabaso, 2021) that sociomaterial theory overcomes. With rationality and agency tied together in decision-making, finding common theoretical ground between artifact agency and the views of rationality in AI can be seen as contributing to some of the issues in AI decision-making. This dissertation makes the theoretical contribution of studying rationality with a relative agency methodology, and shows that it is well applicable. This suggests that rationality studies in AI would benefit from more relational theories of rationality in addition to agency. Moreover, studying rationality as either/or ignores that formal rationality is a substantive rationality choice already. Studying this relationship with AI helps expand thought and see and make connections that have been previously ignored.

This dissertation highlights the issues with the assumptions of bounded rationality (Simon, 1957, 2000, 2013), showing that technology can indeed further constrain rationality in organizational decision-making while simultaneously opening up completely new rationalities. These discoveries challenge conceptualizations of rationality in MOS. Especially the successful application of treating decision-making rationality as relational supports a move in MOS decision-making theory into territories similar to Polanyi's tacit knowledge (Polanyi, 2012), in which rationality is seen as embedded and situated. The findings of this dissertation show that formal, coded rationality is not separable from more substantive, situated rationality in decision-making. The concretization of the assumptions of established theories like bounded rationality via AI can demonstrate their problematic nature, which requires the field of MOS and decision-making to begin a search for new directions regarding decision-making rationality.

den Hond and Moser (2023) called for more nuanced and comprehensive research agendas that take seriously relational agency in technology. They noted that while

these views on agency have been developed, there is still not much recognition of how technology exercises its agency. While this dissertation focused on rationality, it contributes to nuanced understandings of how AI mediates rationality in decision-making with a methodology that takes relational agency seriously. Because agency and rationality are tied in decision-making, this dissertation contributes to efforts to help views of rationality catch up with views on agency.

6.3 Contribution to practice

AI is being widely adopted in practice to aid with decision-making (Haque et al., 2023), and its use is associated with possibilities of gaining competitive advantages and financial performance (Cao & Duan, 2017; Forrest, 2021; McAfee et al., 2012; Rudin, 2019). Thus, its adoption has desirable benefits for organizations. However, realizing those benefits faces acknowledged practical issues. People do not trust AI enough to employ it, due to algorithm aversion (Dietvorst et al., 2015, 2018; Prahll & Van Swol, 2017) and known biases in AI (Baum et al., 2011; Hasan et al., 2021; G. M. Johnson, 2021; O’Neil, 2016). Moreover, practical application of AI implementations can be unsuccessful (e.g., Lebovitz et al., 2021; Van den Broek et al., 2021), and even with seemingly functional applications, there is the risk of managerial overreliance on them and the threat of losing unique human knowledge in organizations (Fügener et al., 2021; Keding & Meissner, 2021).

In response to issues with drawing benefits from AI, the field is searching for answers and fixes. XAI as a field deals with how to build trust in AI decision-making and how to ensure AI fairness and accountability (Binns et al., 2018; Shin, 2020; Shin & Park, 2019; Teodorescu et al., 2021). This dissertation posits rationality for a decision and the explainability of a decision as different points of view of the same thing: answers to “why” questions about decisions (T. Miller, 2019). The difference is that explainability focuses on providing explanations of the rationalities for decisions, while rationality concerns itself more with its emergence and content, not with its communication. A core concept of explainability is transparency (Shin & Park, 2019), which in practical terms means access to the decision rationality.

This dissertation contributes to XAI practices by demonstrating where and how rationality becomes concealed and how it can be revealed (i.e., how to increase transparency). It found that rationality is revealed through contrasting a variety of pos-

sible decisions, and that the active involvement of people in specifically the rationality and its creation is revealing. Practitioners can apply this knowledge to increase transparency by involving people in the creation of the rationality for decisions, not only judging and correcting AI decisions, making for effective placement of human resources in AI decision-making. Moreover, this dissertation contributes to knowledge about how black boxes form, which can be used to increase transparency and consequential trust in AI.

Regarding successful implementations of AI decision-making, there is ongoing discussion about discovering whether and what decision-making to augment or automate, despite such a separation being paradoxical (Raisch & Krakowski, 2021). This dissertation provides insight and a methodology to look past automation or augmentation to discover how and where people become involved in AI decision-making and when they are alienated from it along rationality revealment or concealment. A level of detail into the interlinked mediations is provided, from which practitioners can identify the most relevant for ensuring successful AI adoption and mitigating its pertinent issues in their specific contexts.

Moreover, this dissertation demonstrates that postphenomenological analysis can provide insights into AI decision-making. It can be applied in even more specific contexts in practice to address issues practitioners may have with very specific AI and ML technologies, to identify issues and points of improvement in applications. Indeed, Friedrich et al. (2022) found that postphenomenological analysis helped gain a better understanding of AI applications to medical decision-making beyond splits of explainable versus unexplainable AI and to discover that discovering and attending to the technology's mediation is key for successful practical application. Thus, in addition to the insights already provided into AI mediation in decision-making, the approach taken in this dissertation can be applied to identifying and mitigating issues for AI implementations. Indeed, it needs to be emphasized that postphenomenology does not consider any mediation or technology good or bad, but rather emphasizes the importance of mindful attendance to the mediations.

Moreover, by positing that formal and substantive rationality always co-exist, this dissertation poses an important question about practical applications with regard to formal rationality. AI implementations often rely on an assumption of formal rationality (Shneiderman, 2022) and are developed as removed from contexts (Raisch & Krakowski, 2021). Because formal, purely rational decision-making is often not

possible in practice and is possibly even detrimental to decision-making (Elliot et al., 2020; Y. Li et al., 2014; M. Weber, 2019), this dissertation suggests that practitioners pay active attention and study the assumptions held about decision-making rationality. In this way, AI applications can hopefully become more successful while also mitigating known issues posed by it.

6.4 Assessing the research

Reliability and validity are common measures of research quality, but they mostly apply to quantitative research. While they can be applied to some extent to qualitative research as well, misapplied rigor assessments can harm qualitative inquiry. Conforming qualitative research to fit quantitative criteria can create replicable yet trivial research. (Creswell, 2014; Denzin & Lincoln, 2018) Social sciences often consider it a high aspiration to meet assessment criteria associated with the natural sciences (Schultz, 2010), which creates a dilemma for qualitative research assessment. Denzin and Lincoln (2018) outline a framework for research rigor assessment depending on the “softness” or “hardness” of data, in which the softest data refers to interpretative, experiential data that cannot be compared to any concrete, permanent phenomena: the data is simply what the people say it is. Hard data on the other hand is often replicable, numerical data of permanent phenomena (Denzin & Lincoln, 2018).

The research in this dissertation sits on two different levels of softness of data. The publications in themselves depict studies with hard, numerical data in them, as well as clear coding categories that can be assessed for interrater reliability. The postphenomenological analysis represents the softest type of data, in which, due to the phenomenological nature, subjective, lived experiences are studied. Hence, two different levels of research rigor assessment are relevant. For the harder level of data in the publications, reliability assessments are more relevant than the softest levels (Denzin & Lincoln, 2018).

Reliability assesses whether the same results would be obtained if a study was repeated methodologically (Saunders et al., 2009; Yin, 2009). In the publications, all used quantitative methodology is documented in detail and is replicable to the extent possible. However, due to the black-box nature of some of the used algorithms and the random sampling for test and training data for the AI, the outputs can have some

variance. Empirical publications I–III used a variety of samples to overcome this issue to confirm the results. Publications that use human labeled datasets were, as noted, concealing of the original rationalities for the labels, and thus no clear coding scheme exists for them. Hence, due to the subjective nature of the labels, the results are not reliable across varied datasets. However, publication III, in which a coding scheme was formed between different people in rationality formalization, and publication I, interrater agreement rates have been addressed in the publications for reliability assessment.

Reliability lends itself better to assessment on the level of the publications, from which reliability for the dissertation can be reflected upon. Since the publications are the data used for postphenomenological analysis, and the quality of data is key to research rigour (Denzin & Lincoln, 2018), reliability of the data requires mindful assessment. However, the subjective, experiential aspect of the research does not lend itself well to reliability assessment. Indeed, reliability easily becomes an issue in qualitative research, which usually includes interpretations and drawing inferences from subjective data. Especially a methodology based in phenomenology that focuses on subjective experiences of the world cannot be replicable with any methodology, unless it is assumed people’s experiences are uniform. Because meanings vary from people to people, phenomenological views on reliability sometimes focus on if the discovered “essences” can be applied consistently (Beck et al., 1994; Giorgi et al., 1971).

However, the validity of the research can be assessed regarding its rigor not only on publication level. Validity phenomenologically refers to whether the discovered phenomenon truly captures its vital characteristics, i.e., nothing cannot be added or removed without changing the phenomenon (Beck et al., 1994; Giorgi et al., 1971). Validity refers to the comprehensiveness of the topic, and the inclusion of varied datasets and types of research settings (Denzin, 2012). On a more general level Denzin and Lincoln (2018) outlines rigor assessment of research with soft data to include saturation, methodological cohesion, and theoretical coherence as key features for rigor. Saturation refers to more than seeking the replicability of findings. Instead, it seeks concordance at the conceptual level of the analysis, which can mean finding support in other research and literature, and also finding the same concepts within varied, rich settings. (Denzin & Lincoln, 2018)

This dissertation found saturation of the findings with varied research settings of

semi-structured and unstructured interviews, theoretical approaches, and literature reviews with different data and different settings of decision-makers among people and AI, as well as build concordance of the findings with other literature. Still, further rigor could be attainable by increasing the number of participants in the studies. However, the mediations discovered in the publications were consistently discovered also in some of the other publications, and in previous literature, for instance, people offering less information (Amalberti et al., 1993; Shaikh & Cruz, 2019) on their decision-making or becoming alienated from the decision-making with AI (Jarrahi, 2019). The findings remained logical and coherent among the research set-ups as well as other literature, speaking for the rigor of the findings. Hence, the found essences “applied constantly”, providing support for the phenomenological definition of reliability.

However, it must be noted that this dissertation uses only text data, but also uses different datasets and two different labelling methods for the datasets. Hence, data poses no issues to saturation and triangulation of findings, which supports research validity (Creswell, 2014; Yin, 2009). Moreover, the research set-ups varied in the studies and in publication IV attention was paid to careful selection of varied decision-makers for the task to study mediation. The research uses rich description to communicate the found mediations in accordance with validity recommendations (Creswell, 2014; Yin, 2009). However, a sole researcher is definitionally methodologically constrained (Cheek, 2008). Hence, it poses an issue for rigor that postphenomenological analysis was performed by one person. Personal preconceptions and biases were mitigated as well as possible, but they can be unconscious and, thus, their impact can never be fully ruled out. The researcher focused on mediation discovery without presumptions as suggested for reliable phenomenological methodology (Giorgi et al., 1971), and found the supporting literature after discovering the mediations. A made trade-off for mitigating researcher bias is that the type of data required was not initially determined for the dissertation (Denzin & Lincoln, 2018), but rather true to the explorative approach, emergent directions and concepts from findings directed the research, while the researcher tried to not hold any precognitions about the topic and mediations. Moreover, in the end, due to its nature, phenomenological analysis of the rationality mediations necessarily puts the reader in the place of a critical evaluator of the researcher’s conclusions (Giorgi et al., 1971).

Validity is related to generalization, which refers to the extent the findings of the

research are applicable and can be extended to outside of the used research settings. (Aguinis & Edwards, 2014; Yin, 2009) Here it must be again noted, that the research philosophy has its roots in phenomenology, which sees people as having multiple realities and living multiple truths shaped by their lived experiences (Saunders et al., 2009). Hence, a fully generalizable “truth” is irrelevant. However, on a publication level each publication ensures its transparency into its limitations regarding generalizability, from which their generalizability can be assessed. The findings are not specific to a certain industry, but they are limited as to the type of data used and the studied AI methodology. This poses obvious constraints on generalizability, but the consistencies of the mediations in different research set-ups and support from other literature suggest the discovered mediations extend beyond their specific research settings. For exploratory research, instead of generalizability, the worthiness of the topic and contributions can be a more meaningful aspect to study (Tracy, 2010). This dissertation contributes to many relevant and ongoing discussions around AI impacts, and poses synthesis and testable claims that can help take advantage of while mitigating the issues of AI.

6.5 Limitations and future research avenues

The research in this dissertation, as with all research, has limitations to be acknowledged. Some limitations are posed by the research context and scope, while others are posed by the research philosophy and methodology. Regarding the context, this dissertation studied and compared decision-making rationality with and without AI present and related the findings to MOS conversations and discussions around this topic. Regarding its contribution to MOS, a clear limitation is the lack of coverage on collective decision-making. All of the publications are concerned with individual decision-making, even when individuals act together. The transferability of the findings of such an approach to larger collectives making decisions is problematic. Moreover, the decision tasks used in the studies were simple, and thus longer chains of decisions that impact each other were not studied.

Individual decision-making is associated with Herbert Simon’s bounded rationality, and a move to studying rationality and agency more in line with tacit knowledge is suggested (K. D. Miller, 2008). Tacit knowledge emphasizes communal emergence on rationality (Polanyi, 2012). Indeed, in the future, research into AI rationality me-

diations should be extended into studying larger collective decision-making. Despite postphenomenology being criticized as not applicable to the study of technology in a wider context, it can be used to study specific technologies, even with larger collectives (see Chapter 3.1.3.2).

Another limitation is the sole focus on rationality. The dissertation conceptually clarifies the relationships between rationality, agency, and decision-making to contribute to research on those aspects. However, to get a fuller scope of AI impacts on decision-making, agency should be afforded dedicated research in the future. Possible studies could include the mediations of levels of activity and passivity in decision-making, not only regarding rationality. Moreover, what emerged in this dissertation as AI's authoritative agency that alienated people from decision-making could be extended into research of how much people feel their agency becomes constrained in relation to an AI in decision-making. In the light of recent discussions on AI moral agency, it becomes a fascinating question of whether people feel an AI moral suggestion affects their actions, for instance, the impact on agency between the cases in figure 2.2.

A clear limitation is posed by the studied AI methods and data. Only SML and UML were studied using specifically text data for simple classification tasks. While the methods were chosen for their ubiquity, and thus aiming for the usefulness and generalizability of conclusions drawn from them, the mediations with reinforcement learning and, for instance, image and tabular data should be conducted to confirm the generalizability and external validity of the conclusions. Moreover, the specific phases of AI decision-making should be awarded sole focus in future research. For UML pre-processing and analysis decisions were given a level of detail that was not present for SML: UML choices require justification, whereas SML uses whatever yields the best accuracy or other similar quantitative measure. However, in UML, the rationality was concealed, while it was revealed in SML. However, the revelation of rationality in SML in this aspect was found to be alienating. Further detail on this level of AI decision-making is required.

Of course, the used research philosophy has its limitations. Postphenomenology, as acknowledged, does not do well with studying technology in broader social contexts. Moreover, its roots in pragmatism and phenomenology emphasize the limits of knowledge to multiple lived truths and the practical value of research. To fully bring home a postphenomenological study, the work is not yet finished. The found

mediations here should be formulated into empirical research settings and verified therein. Moreover, the AI mediations of rationality in decision-making should be applied in practice by designing AI that takes the mediations into account to realize benefits and wanted impacts while mitigating what is unwanted in the mediations. Moreover, postphenomenological methodology can be further applied in MOS for intentional design of technology.

REFERENCES

- Abney, K. (2012). Robotics, ethical theory, and metaethics: A guide for the perplexed. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics* (pp. 35–52). MIT Press, Cambridge, MA.
- Aguinis, H., & Edwards, J. R. (2014). Methodological wishes for the next decade and how to make wishes come true. *Journal of Management Studies*, 51(1), 143–174.
- Alonso, E. (2014). Actions and agents. In K. Frankish & W. M. Ramsey (Eds.), *The cambridge handbook of artificial intelligence* (pp. 232–46). Cambridge University Press Cambridge, UK.
- Al-Surmi, A., Bashiri, M., & Koliouisis, I. (2022). Ai based decision making: Combining strategies to improve operational performance. *International Journal of Production Research*, 60(14), 4464–4486.
- Amalberti, R., Carbonell, N., & Falzon, P. (1993). User representations of computer systems in human–computer speech interaction. *International Journal of Man–Machine Studies*, 38(4), 547–566.
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115.
- Balasubramanian, N., Ye, Y., & Xu, M. (2022). Substituting human decision-making with machine learning: Implications for organizational learning. *Academy of Management Review*, 47(3), 448–465.
- Barad, K. (2007). *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*. Duke University Press.

- Baum, S. D., Goertzel, B., & Goertzel, T. G. (2011). How long until human-level AI? results from an expert assessment. *Technological Forecasting and Social Change*, 78(1), 185–195.
- Beck, C. T., Keddy, B. A., & Cohen, M. Z. (1994). Reliability and validity issues in phenomenological research. *Western Journal of Nursing Research*, 16(3), 254–267.
- Bermúdez, J. L. (2009). *Decision theory and rationality*. Oxford University Press Oxford.
- Bertsimas, D., & Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science*, 66(3), 1025–1044.
- Bettis, R. A. (2017). Organizationally intractable decision problems and the intellectual virtues of heuristics. *Journal of Management*, 43(8), 2620–2637.
- Bijker, W. E., Hughes, T. P., & Pinch, T. (Eds.). (1987). *The social construction of technological systems: New directions in the sociology and history of technology*. MIT Press.
- Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). "it's reducing a human being to a percentage": Perceptions of justice in algorithmic decisions. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14.
- Blackler, F. (1995). Knowledge, knowledge work and organizations: An overview and interpretation. *Organization Studies*, 16(6), 1021–1046.
- Borges, A. F. S., Laurindo, F. J. B., Spinola, M. M., Gonçalves, R. F., & Mattos, C. A. (2021). The strategic use of artificial intelligence in the digital era: Systematic literature review and future research directions. *International Journal of Information Management*, 57, 102225.
- Broome, J. (2021). Reasons and rationality. In M. Knauff & W. Spohn (Eds.), *The handbook of rationality* (pp. 129–136). MIT Press.
- Bryson, J. J. (2018). Patience is not a virtue: The design of intelligent systems and systems of ethics. *Ethics and Information Technology*, 20(1), 15–26.
- Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2), 220–239.
- Caldwell, R. (2007). Agency and change: Re-evaluating Foucault's legacy. *Organization*, 14(6), 769–791.

- Callon, M. (1984). Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St. Brieuc Bay. *The Sociological Review*, 32(1_suppl), 196–233.
- Cao, G., & Duan, Y. (2017). How do top- and bottom-performing companies differ in using business analytics? *Journal of Enterprise Information Management*, 30(6), 874–892.
- Castelfranchi, C., & Falcone, R. (2003). From automaticity to autonomy: The frontier of artificial agents. In H. Hexmoor, C. Castelfranchi, & R. Falcone (Eds.), *Agent autonomy: Multiagent systems, artificial societies, and simulated organizations* (pp. 103–136). Springer.
- Castro-Manzano, J. M. (2010). The argument from autonomy revisited. *2010 Ninth Mexican International Conference on Artificial Intelligence*, 67–72.
- Chalmers, D., MacKenzie, N. G., & Carter, S. (2021). Artificial intelligence and entrepreneurship: Implications for venture creation in the fourth industrial revolution. *Entrepreneurship Theory and Practice*, 45(5), 1028–1053.
- Cheek, J. (2008). Researching collaboratively: Implications for qualitative research and researchers. *Qualitative health research*, 18(11), 1599–1603.
- Choudhury, P., Allen, R. T., & Endres, M. G. (2021). Machine learning for pattern discovery in management research. *Strategic Management Journal*, 42(1), 30–57.
- Clark, B. B., Robert, C., & Hampton, S. A. (2016). The technology effect: How perceptions of technology drive excessive optimism. *Journal of Business and Psychology*, 31, 87–102.
- Claudy, M. C., Aquino, K., & Graso, M. (2022). Artificial intelligence can't be charmed: The effects of impartiality on laypeople's algorithmic preferences. *Frontiers in Psychology*, 13, 898027.
- Cohen, M. D., March, J. G., & Olsen, J. P. (1972). A garbage can model of organizational choice. *Administrative Science Quarterly*, 17(1), 1–25.
- Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press. <https://books.google.fi/books?id=XvEdEAAAQBAJ>
- Crease, R. P., & Achterhuis, H. (2001). *American philosophy of technology: The empirical turn*. Indiana University Press.

- Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th). Sage.
- Csaszar, F. A., & Eggers, J. P. (2013). Organizational decision making: An information aggregation view. *Management Science*, 59(10), 2257–2277.
- Cui, R., Gallino, S., Moreno, A., & Zhang, D. J. (2018). The operational value of social media information. *Production and Operations Management*, 27(10), 1749–1769.
- Da, Z., & Huang, X. (2020). Harnessing the wisdom of crowds. *Management Science*, 66(5), 1847–1867.
- Davidson, D. (1982). Rational animals. *dialectica*, 36(4), 317–327.
- Davis, F. D., Lohse, G. L., & Kottemann, J. E. (1994). Harmful effects of seemingly helpful information on forecasts of stock earnings. *Journal of Economic Psychology*, 15(2), 253–267.
- Debenham, J., & Sierra, C. (2010). Ecologically rational agency. *IFIP International Conference on Artificial Intelligence in Theory and Practice*, 3–12.
- de Graaf, M. M., & Malle, B. F. (2019). People’s explanations of robot behavior subtly reveal mental state inferences. *2019 14th ACM/IEEE International Conference on Human–Robot Interaction (HRI)*, 239–248.
- den Hond, F., & Moser, C. (2023). Useful servant or dangerous master? Technology in business and society debates. *Business & Society*, 62(1), 87–116.
- Dennett, D. C. (2004). *Freedom evolves*. Penguin UK.
- Dennett, D. C. (2017). *From bacteria to Bach and back: The evolution of minds*. W W Norton & Company.
- Denzin, N. K. (2012). Triangulation 2.0. *Journal of mixed methods research*, 6(2), 80–88.
- Denzin, N. K., & Lincoln, Y. S. (Eds.). (2018). *The SAGE handbook of qualitative research* (5th ed.). SAGE Publications.
- de Vaujany, F.-X., Aroles, J., & Perézts, M. (2023). Phenomenologies and organization studies: Organizing through and beyond appearances. In F.-X. de Vaujany, J. Aroles, & M. Perézts (Eds.), *The oxford handbook of phenomenologies and organization studies*. Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780192865755.013.2>
- Dhar, P. (2020). The carbon impact of artificial intelligence. *Nature Machine Intelligence*, 2(8), 423–425.

- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126.
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170.
- Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of big data—evolution, challenges and research agenda. *International Journal of Information Management*, *48*, 63–71.
- Edwards, J. S., Duan, Y., & Robins, P. C. (2000). An analysis of expert systems for business decision making at different levels and in different roles. *European Journal of Information Systems*, *9*(1), 36–46.
- Ehrich, L. (2005). Revisiting phenomenology: Its potential for management research. In K. Grint (Ed.), *Challenges of organisations in global markets: Bam-2005* (pp. 1–13). British Academy of Management.
- Einola, K., & Khoreva, V. (2023). Best friend or broken tool? Exploring the co-existence of humans and artificial intelligence in the workplace ecosystem. *Human Resource Management*, *62*(1), 117–135.
- Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review*, *14*(4), 532–550.
- Eisenhardt, K. M., & Graebner, M. E. (2007). Theory building from cases: Opportunities and challenges. *Academy of Management Journal*, *50*(1), 25–32.
- Elder, A. (2020). The interpersonal is political: Unfriending to promote civic discourse on social media. *Ethics and Information Technology*, *22*(1), 15–24.
- Elliot, V. H., Paananen, M., & Staron, M. (2020). Artificial intelligence for decision-makers. *Journal of Emerging Technologies in Accounting*, *17*(1), 51–55.
- Elsbach, K. D., & Stigliani, I. (2019). New information technology and implicit bias. *Academy of Management Perspectives*, *33*(2), 185–206.
- Elson, J. S., Derrick, D., & Ligon, G. (2018). Examining trust and reliance in collaborations between humans and automated agents. *Proceedings of the 51st Hawaii International Conference on System Sciences*, 430–439.
- Emirbayer, M., & Mische, A. (1998). What is agency? *American journal of sociology*, *103*(4), 962–1023.

- Enholm, I. M., Papagiannidis, E., Mikalef, P., & Krogstie, J. (2022). Artificial intelligence and business value: A literature review. *Information Systems Frontiers*, 24(5), 1709–1734.
- Eriksson, P., & Kovalainen, A. (2015). *Qualitative methods in business research: A practical guide to social research*. Sage.
- Etzioni, A., & Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *The Journal of Ethics*, 21, 403–418.
- Farjoun, M., Ansell, C., & Boin, A. (2015). Perspective—pragmatism in organization studies: Meeting the challenges of a dynamic and complex world. *Organization Science*, 26(6), 1787–1804.
- Forrest, K. B. (2021). *When machines can be judge, jury, and executioner: Ujustice in the age of artificial intelligence*. World Scientific.
- Foucault, M. (2005). *The order of things*. Routledge.
- Fountaine, T., McCarthy, B., & Saleh, T. (2019). Building the AI-powered organization. *Harvard Business Review*, 97(4), 62–73.
- Frederick, W. C. (1998). Creatures, corporations, communities, chaos, complexity: A naturological view of the corporate social role. *Business & Society*, 37(4), 358–389.
- Friedrich, A. B., Mason, J., & Malone, J. R. (2022). Rethinking explainability: Toward a postphenomenology of black-box artificial intelligence in medicine. *Ethics and Information Technology*, 24(1), 1–9.
- Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *MIS Quarterly*, 45(3), 1527–1556.
- Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137–144.
- Gao, J., Koronios, A., & Selle, S. (2015). Towards a process view on critical success factors in big data analytics projects. *AMCIS 2015 Proceedings*, 16.
- Geertz, C. (1973). Thick description: Toward an interpretive theory of culture. In *The interpretation of cultures: Selected essays* (pp. 3–30). Basic Books New York.
- Ghasemaghaei, M. (2020). Improving organizational performance through the use of big data. *Journal of Computer Information Systems*, 60(5), 395–408.

- Giddens, A. (2013). *The constitution of society: Outline of the theory of structuration*. Polity Press. <https://books.google.fi/books?id=YD87I8uPvnUC>
- Giorgi, A., Fischer, W. F., & Von Eckartsberg, R. (Eds.). (1971). *Duquesne studies in phenomenological psychology* (Vol. 1). Duquesne University Press Pittsburgh.
- Glaser, V. L., Pollock, N., & D'Adderio, L. (2021). The biography of an algorithm: Performing algorithmic technologies in organizations. *Organization Theory*, 2(2), 26317877211004609.
- Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2), 627–660.
- Golman, R., Hagmann, D., & Loewenstein, G. (2017). Information avoidance. *Journal of Economic Literature*, 55(1), 96–135.
- Greenwood, M., & Wolfram Cox, J. (2022). Seduced by technology? How moral agency is mediated by the invisibility of everyday technologies. *Organization Studies*.
- Haenlein, M., & 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.
- Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4), 814–834.
- Hao, K. (2020a). The two-year fight to stop Amazon from selling face recognition to the police. *MIT Technology Review*.
- Hao, K. (2020b). We read the paper that forced Timnit Gebru out of Google. Here's what it says. *MIT Technology Review*.
- Hao, K. (2021). The race to understand the exhilarating, dangerous world of language AI. *MIT Technology Review*.
- Haque, A. K. M. B., Islam, A. K. M. N., & Mikalef, P. (2023). Explainable artificial intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technological Forecasting and Social Change*, 186.
- Hasan, R., Shams, R., & Rahman, M. (2021). Consumer trust and perceived risk for voice-controlled artificial intelligence: The case of Siri. *Journal of Business Research*, 131, 591–597.
- Heaven, W. D. (2022). Why Meta's latest large language model survived only three days online. *MIT Technology Review*.

- Heidegger, M. (1977). The question concerning technology. In *The question concerning technology and other essays* (pp. 3–35). Harper Row New York.
- Heidegger, M. (2010). *Being and time*. Suny Press.
- Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2022). Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, 25(1), 114–146.
- Ihde, D. (1990). *Technology and the lifeworld: From garden to earth*. Indiana University Press. <https://books.google.fi/books?id=u2pRAAAAMAAJ>
- Ihde, D. (2012). *Experimental phenomenology, second edition: Multistabilities*. State University of New York Press. <https://books.google.fi/books?id=UBf9r-7tKkcC>
- Illies, C. F. R., & Meijers, A. (2014). Artefacts, agency, and action schemes. In P. Kroes & P.-P. Verbeek (Eds.), *The moral status of technical artefacts* (pp. 159–184). Springer.
- Indhul, S. (2022). Towards a constructor theory conception for wicked social externalities: Delineating the limits and possibilities of impactful pathways to a better world. In A. Thakhathi (Ed.), *Transcendent development: The ethics of universal dignity* (pp. 43–52). Emerald Publishing Limited.
- Introna, L. D. (2014). Towards a post-human intra-actional account of sociomaterial agency (and morality). In P. Kroes & P.-P. Verbeek (Eds.), *The moral status of technical artefacts* (pp. 31–53). Springer.
- Izenman, A. J. (2008). Springer.
- Jarrahi, M. H. (2019). In the age of the smart artificial intelligence: AI's dual capacities for automating and informing work. *Business Information Review*, 36(4), 178–187.
- Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology*, 8, 195–204.
- Johnson, D. G., & Verdicchio, M. (2019). AI, agency and responsibility: The VW fraud case and beyond. *AI & Society*, 34, 639–647.
- Johnson, G. M. (2021). Algorithmic bias: On the implicit biases of social technology. *Synthese*, 198(10), 9941–9961.
- Kalberg, S. (1980). Max Weber's types of rationality: Cornerstones for the analysis of rationalization processes in history. *American Journal of Sociology*, 85(5), 1145–1179.

- Kaplan, A., & Haenlein, M. (2020). Rulers of the world, unite! The challenges and opportunities of artificial intelligence. *Business Horizons*, 63(1), 37–50.
- Kashima, Y., McKintyre, A., & Clifford, P. (1998). The category of the mind: Folk psychology of belief, desire, and intention. *Asian Journal of Social Psychology*, 1(3), 289–313.
- Keding, C., & Meissner, P. (2021). Managerial overreliance on AI-augmented decision-making processes: How the use of AI-based advisory systems shapes choice behavior in R&D investment decisions. *Technological Forecasting and Social Change*, 171, 120970.
- Kelemen, M. L., & Rumens, N. (2008). *An introduction to critical management research*. Sage.
- Kelly, L. M., & Cordeiro, M. (2020). Three principles of pragmatism for research on organizational processes. *Methodological Innovations*, 13(2), 1–10.
- Kenny, A. (2018). *An illustrated brief history of western philosophy*. John Wiley & Sons.
- Kibble, R. (2017). Communication breakdown? Reasoning about language and rational agents. *AISB 2017: Computing and Philosophy Symposium*.
- Kiran, A. H. (2015). Four dimensions of technological mediation. In R. Rosenberger & P.-P. Verbeek (Eds.), *Postphenomenological investigations: Essays on human–technology relations* (pp. 123–140). Lexington Books London.
- Knauff, M., & Spohn, W. (2021). *Psychological and philosophical frameworks of rationality: A systematic introduction* (M. Knauff & W. Spohn, Eds.).
- Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihok, G., & Den Hartog, D. N. (2018). Text classification for organizational researchers: A tutorial. *Organizational Research Methods*, 21(3), 766–799.
- Kottemann, J. E., Davis, F. D., & Remus, W. E. (1994). Computer-assisted decision making: Performance, beliefs, and the illusion of control. *Organizational Behavior and Human Decision Processes*, 57(1), 26–37.
- Krakowski, S., Luger, J., & Raisch, S. (2022). Artificial intelligence and the changing sources of competitive advantage. *Strategic Management Journal*.
- Kraut, R. (2022). Aristotle's ethics. In E. N. Zalta & U. Nodelman (Eds.), *The stanford encyclopedia of philosophy* (Fall 2022). <https://plato.stanford.edu/archives/fall2022/entries/aristotle-ethics/>

- Krügel, S., Ostermaier, A., & Uhl, M. (2023). Algorithms as partners in crime: A lesson in ethics by design. *Computers in Human Behavior*, 138, 107483.
- Kuang, L., Yang, L. T., Chen, J., Hao, F., & Luo, C. (2015). A holistic approach for distributed dimensionality reduction of big data. *IEEE Transactions on Cloud Computing*, 6(2), 506–518.
- Langley, P., Meadows, B., Sridharan, M., & Choi, D. (2017). Explainable agency for intelligent autonomous systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(2), 4762–4763.
- Lash, S. (2003). Reflexivity as non-linearity. *Theory, Culture & Society*, 20(2), 49–57.
- Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Harvard University Press.
- Latour, B. (1992). Where are the missing masses? The sociology of a few mundane artifacts. In W. E. Bijker & J. Law (Eds.), *Shaping technology/building society: Studies in sociotechnical change* (pp. 225–258). MIT Press Cambridge, MA.
- Latour, B. (2007). *Reassembling the social: An introduction to actor-network-theory*. Oxford University Press Oxford.
- Latour, B. (2012). *We have never been modern*. Harvard University Press.
- Law, J., & Hassard, J. (1999). *Actor network theory and after*. Wiley-Blackwell.
- Lebovitz, S., Levina, N., & 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, 1501–1525.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- Lee, D., & Van den Steen, E. (2010). Managing know-how. *Management Science*, 56(2), 270–285.
- Leonardi, P. M., Nardi, B. A., & Kallinikos, J. (2012). *Materiality and organizing: Social interaction in a technological world*. Oxford University Press on Demand.
- Lewis, D. (1986). Causal explanation. In *Philosophical papers vol. II* (pp. 214–240). Oxford University Press.
- Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science*, 6(2), 279–311.

- Li, X., Chen, D., Xu, W., Chen, H., Li, J., & Mo, F. (2023). Explainable dimensionality reduction (XDR) to unbox AI "black box" models: A study of AI perspectives on the ethnic styles of village dwellings. *Humanities and Social Sciences Communications*, *10*(1), 1–13.
- Li, Y., Ashkanasy, N. M., & Ahlstrom, D. (2014). The rationality of emotions: A hybrid process model of decision-making under uncertainty. *Asia Pacific Journal of Management*, *31*, 293–308.
- Liao, S. M. (2020). The moral status and rights of artificial intelligence. In S. M. Liao (Ed.), *Ethics of artificial intelligence* (pp. 480–503). Oxford University Press.
- Lichtenthaler, U. (2019). An intelligence-based view of firm performance: Profiting from artificial intelligence. *Journal of Innovation Management*, *7*(1), 7–20.
- Lindebaum, D., Vesa, M., & den Hond, F. (2020). Insights from "The Machine Stops" to better understand rational assumptions in algorithmic decision making and its implications for organizations. *Academy of Management Review*, *45*(1), 247–263.
- Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplements*, *27*, 247–266.
- Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103.
- Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464–470. <https://doi.org/https://doi.org/10.1016/j.tics.2006.08.004>
- Lukes, S. (2006). *Individualism*. ECPR Press. <https://books.google.fi/books?id=IEOg9yiNx7kC>
- Luoma, J. (2016). Model-based organizational decision making: A behavioral lens. *European Journal of Operational Research*, *249*(3), 816–826.
- Ma, L., & Sun, B. (2020). Machine learning and AI in marketing—Connecting computing power to human insights. *International Journal of Research in Marketing*, *37*(3), 481–504. <https://doi.org/https://doi.org/10.1016/j.ijresmar.2020.04.005>
- Mabaso, B. A. (2021). Computationally rational agents can be moral agents. *Ethics and Information Technology*, *23*(2), 137–145.

- MacKenzie, D., & Wajcman, J. (1999). *The social shaping of technology*. Open University Press.
- Mahama, H., Elbashir, M. Z., Sutton, S. G., & Arnold, V. (2016). A further interpretation of the relational agency of information systems: A research note. *International Journal of Accounting Information Systems*, 20, 16–25.
- Mahmoud, M. S. (2020). *Multiagent systems: Introduction and coordination control*. CRC Press.
- Malle, B. F. (2006). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT Press.
- Manna, R., & Nath, R. (2021). The problem of moral agency in artificial intelligence. *2021 IEEE Conference on Norbert Wiener in the 21st Century (21CW)*, 1–4.
- Marler, J. H., Fisher, S. L., & Ke, W. (2009). Employee self-service technology acceptance: A comparison of pre-implementation and post-implementation relationships. *Personnel Psychology*, 62(2), 327–358.
- Martinho, A., Poulsen, A., Kroesen, M., & Chorus, C. (2021). Perspectives about artificial moral agents. *AI and Ethics*, 1(4), 477–490.
- McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: The management revolution. *Harvard Business Review*, 90(10), 60–68.
- Miller, F. D. (1984). Aristotle on rationality in action. *The Review of Metaphysics*, 37(3), 499–520.
- Miller, K. D. (2008). Simon and Polanyi on rationality and knowledge. *Organization Studies*, 29(7), 933–955.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38.
- Mitcham, C. (2014). Agency in humans and in artifacts: A contested discourse. In P. Kroes & P.-P. Verbeek (Eds.), *The moral status of technical artefacts* (pp. 11–29). Springer.
- Mol, A. (2002). *The body multiple: Ontology in medical practice*. Duke University Press.
- Möller, K., Schäffer, U., & Verbeeten, F. (2020). Digitalization in management accounting and control: An editorial. *Journal of Management Control*, 31, 1–8.

- Morgan, D. L. (2014). Pragmatism as a paradigm for social research. *Qualitative Inquiry*, 20(8), 1045–1053.
- Morgan, G., & Smircich, L. (1980). The case for qualitative research. *Academy of Management Review*, 5(4), 491–500.
- Moser, C., den Hond, F., & Lindebaum, D. (2022a). Morality in the age of artificially intelligent algorithms. *Academy of Management Learning & Education*, 21(1), 139–155.
- Moser, C., den Hond, F., & Lindebaum, D. (2022b). What humans lose when we let AI decide. *MIT Sloan Management Review*, 63(3), 12–14.
- Murray, A., Rhymer, J., & Sirmon, D. G. (2021). Humans and technology: Forms of conjoined agency in organizations. *Academy of Management Review*, 46(3), 552–571.
- Muslea, I., Minton, S., & Knoblock, C. A. (2006). Active learning with multiple views. *Journal of Artificial Intelligence Research*, 27, 203–233.
- Ngai, E. W. T., & Wu, Y. (2022). Machine learning in marketing: A literature review, conceptual framework, and research agenda. *Journal of Business Research*, 145, 35–48.
- Olabode, O. E., Boso, N., Hultman, M., & Leonidou, C. N. (2022). Big data analytics capability and market performance: The roles of disruptive business models and competitive intensity. *Journal of Business Research*, 139, 1218–1230.
- O’Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Penguin Books Limited. <https://books.google.fi/books?id=60n0DAAAQBAJ>
- Orlikowski, W. J. (2007). Sociomaterial practices: Exploring technology at work. *Organization Studies*, 28(9), 1435–1448.
- Orlikowski, W. J., & Scott, S. V. (2008). Sociomateriality: Challenging the separation of technology, work and organization. *Academy of Management Annals*, 2(1), 433–474.
- Overton, J. (2011). Scientific explanation and computation. *Explanation-Aware Computing ExaCt 2011*, 41–50.
- Özemre, M., & Kabadurmus, O. (2020). A big data analytics based methodology for strategic decision making. *Journal of Enterprise Information Management*, 33(6), 1467–1490.

- Panisson, A. R., Engelmann, D. C., & Bordini, R. H. (2021). Engineering explainable agents: An argumentation-based approach. *International Workshop on Engineering Multi-Agent Systems*, 273–291.
- Parry, K., Cohen, M., & Bhattacharya, S. (2016). Rise of the machines: A critical consideration of automated leadership decision making in organizations. *Group & Organization Management*, 41(5), 571–594.
- Parthemore, J., & Whitby, B. (2013). What makes any agent a moral agent? Reflections on machine consciousness and moral agency. *International Journal of Machine Consciousness*, 5(02), 105–129.
- Parthemore, J., & Whitby, B. (2014). Moral agency, moral responsibility, and artifacts: What existing artifacts fail to achieve (and why), and why they, nevertheless, can (and do!) make moral claims upon us. *International Journal of Machine Consciousness*, 6(02), 141–161.
- Paschen, J., Wilson, M., & Ferreira, J. J. (2020). Collaborative intelligence: How human and artificial intelligence create value along the B2B sales funnel. *Business Horizons*, 63(3), 403–414.
- Patokorpi, E. (2008). Simon’s paradox: Bounded rationality and the computer metaphor of the mind. *Human Systems Management*, 27(4), 285–294.
- Patton, M. Q. (2014). *Qualitative research & evaluation methods: Integrating theory and practice*. Sage Publications.
- Perrigo, B. (2023). OpenAI used Kenyan workers on less than \$2 per hour to make ChatGPT less toxic [Accessed on 30.3.2023]. *Time*. <https://time.com/6247678/openai-chatgpt-kenya-workers/>
- Polanyi, M. (2012). *Personal knowledge*. Routledge.
- Prahl, A., & Van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, 36(6), 691–702.
- Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, 48, 137–141.
- Rai, A., Constantinides, P., & Sarker, S. (2019). Next generation digital platforms: Toward human–AI hybrids. *MIS Quarterly*, 43(1), iii–ix.
- Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review*, 46(1), 192–210.

- Ritter, M. (2021a). Philosophical potencies of postphenomenology. *Philosophy & Technology*, 34(4), 1501–1516.
- Ritter, M. (2021b). Postphenomenological method and technological things themselves. *Human Studies*, 44(4), 581–593.
- Robinson, T. J., Giles, R. C., & Rajapakshage, R. U. (2020). Discussion of “experiences with big data: Accounts from a data scientist’s perspective”. *Quality Engineering*, 32(4), 543–549.
- Rometty, G. (2016). Digital today, cognitive tomorrow. *MIT Sloan Management Review*, 58(1), 28.
- Rosenberger, R. (2014). Multistability and the agency of mundane artifacts: From speed bumps to subway benches. *Human Studies*, 37(3), 369–392.
- Rosenberger, R., & Verbeek, P.-P. (2015). *Postphenomenological investigations: Essays on human–technology relations*. Lexington Books.
- Rouleau, T. (2020). What are the types of machine learning? [Accessed on 30.3.2023]. *Sama*. <https://www.sama.com/blog/types-of-machine-learning>
- 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.
- Russell, S. (2010). *Artificial intelligence: A modern approach*. Pearson Education, Inc.
- Russell, S. (2016). Rationality and intelligence: A brief update. In V. C. Müller (Ed.), *Fundamental issues of artificial intelligence* (pp. 7–28). Springer.
- Russell, S., & Norvig, P. (1995). *Artificial intelligence: A modern approach*. Prentice-Hall.
- Sado, F., Loo, C. K., Liew, W. S., Kerzel, M., & Wermter, S. (2023). Explainable goal-driven agents and robots—A comprehensive review. *ACM Computing Surveys*, 55(10), 1–41.
- Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263, 110273.
- Sanders, P. (1982). Phenomenology: A new way of viewing organizational research. *Academy of Management Review*, 7(3), 353–360.
- Sartre, J.-P. (2021). *Being and nothingness: An essay in phenomenological ontology*. Atria Books. <https://books.google.fi/books?id=Xj5qDwAAQBAJ>

- Saunders, M., Lewis, P., & Thornhill, A. (2009). *Research methods for business students*. Pearson Education.
- Schneider, S., & Leyer, M. (2019). Me or information technology? Adoption of artificial intelligence in the delegation of personal strategic decisions. *Managerial and Decision Economics*, 40(3), 223–231.
- Schultz, M. (2010). Reconciling pragmatism and scientific rigor. *Journal of Management Inquiry*, 19(3), 274–277.
- Schwab, K. (2017). *The fourth industrial revolution*. Currency.
- Shaikh, S. J., & Cruz, I. (2019). "Alexa, do you know anything?" The impact of an intelligent assistant on team interactions and creative performance under time scarcity. *arXiv*, 1912.12914.
- Shapira, Z. (2002). *Organizational decision making*. Cambridge University Press.
- Shead, S. (2022). Machines are getting better at writing their own code. But human-level is "light years away". *CNBC*. <https://www.cnbc.com/2022/02/08/deepmind-openai-machines-better-at-writing-their-own-code.html>
- Shin, D. (2020). User perceptions of algorithmic decisions in the personalized AI system: Perceptual evaluation of fairness, accountability, transparency, and explainability. *Journal of Broadcasting & Electronic Media*, 64(4), 541–565.
- Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*, 98, 277–284.
- Shneiderman, B. (2022). *Human-centered AI*. Oxford University Press. <https://books.google.fi/books?id=YS9VEAAAQBAJ>
- Shrestha, Y. R., Ben-Menahem, S. M., & Von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review*, 61(4), 66–83.
- Shrestha, Y. R., Krishna, V., & von Krogh, G. (2021). Augmenting organizational decision-making with deep learning algorithms: Principles, promises, and challenges. *Journal of Business Research*, 123, 588–603.
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, 550(7676), 354–359.
- Simon, H. A. (1957). *Models of man: Social and rational*. Wiley.

- Simon, H. A. (1960). The new science of management decision.
- Simon, H. A. (2000). Bounded rationality in social science: Today and tomorrow. *Mind & Society*, 1, 25–39.
- Simon, H. A. (2013). *Administrative behavior*. Simon; Schuster.
- Simon, H. A., & Newell, A. (1958). Heuristic problem solving: The next advance in operations research. *Operations Research*, 6(1), 1–10.
- Sindhu Meena, K., & Suriya, S. (2020). A survey on supervised and unsupervised learning techniques. In L. A. Kumar, L. Jayashree, & R. Manimegalai (Eds.), *Proceedings of international conference on artificial intelligence, smart grid and smart city applications: Aisgsc 2019* (pp. 627–644).
- Sørensen, M. H., & Ziemke, T. (2007). Agents without agency? *Cognitive Semiotics*, 1, 102–124.
- Strandberg, C. (2017). A puzzle about reasons and rationality. *The Journal of Ethics*, 21(1), 63–88.
- Sturm, T. (2021). Theories of rationality and the descriptive-normative divide: A historical approach. In M. Knauff & W. Spohn (Eds.), *The handbook of rationality* (pp. 71–86). MIT Press.
- Sun, R. (2014). Connectionism and neural networks. In K. Frankish & W. M. Ramsey (Eds.), *The cambridge handbook of artificial intelligence* (pp. 108–127). Cambridge University Press Cambridge.
- Sundar, S. S. (n.d.). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger & A. J. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 73–100). MIT Press.
- Teodorescu, M. H. M., Morse, L., Awwad, Y., & Kane, G. C. (2021). Failures of fairness in automation require a deeper understanding of human-ML augmentation. *MIS Quarterly*, 45(3), 1483–1500.
- Thellman, S., Silvervarg, A., & Ziemke, T. (2017). Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots. *Frontiers in Psychology*, 8, 1962.
- Tonidandel, S., King, E. B., & Cortina, J. M. (2018). Big data methods: Leveraging modern data analytic techniques to build organizational science. *Organizational Research Methods*, 21(3), 525–547.
- Tracy, S. J. (2010). Qualitative quality: Eight “big-tent” criteria for excellent qualitative research. *Qualitative Inquiry*, 16(10), 837–851.

- Tsoukas, H. (2005). Do we really understand tacit knowledge? In S. Little & T. Ray (Eds.), *Managing knowledge: An essential reader* (Second edition, pp. 1–18). Sage.
- Tsoukas, H. (2023). Afterword: Why and how phenomenology matters to organizational research. In F.-X. de Vaujany (Ed.), *The oxford handbook of phenomenologies and organization studies*. Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780192865755.013.38>
- Tunçalp, D. (2016). Questioning the ontology of sociomateriality: A critical realist perspective. *Management Decision*, 54(5), 1073–1087.
- Van den Broek, E., Sergeeva, A., & Huysman, M. (2021). When the machine meets the expert: An ethnography of developing AI for hiring. *MIS Quarterly*, 45(3), 1557–1580.
- van Kraalingen, I. (2022). Theorizing technological mediation in the outdoor classroom. *Postdigital Science and Education*, 1–23.
- van Rijmenam, M., & Logue, D. (2021). Revising the "science of the organisation": Theorising AI agency and actorhood. *Innovation*, 23(1), 127–144.
- Verbeek, P.-P. (2005). *What things do: Philosophical reflections on technology, agency, and design*. Penn State Press.
- Verbeek, P.-P. (2006). Materializing morality: Design ethics and technological mediation. *Science, Technology, & Human Values*, 31(3), 361–380.
- Verbeek, P.-P. (2014). Some misunderstandings about the moral significance of technology. In P. Kroes & P.-P. Verbeek (Eds.), *The moral status of technical artefacts* (pp. 75–88). Springer.
- Vesa, M., & Tienari, J. (2022). Artificial intelligence and rationalized unaccountability: Ideology of the elites? *Organization*, 29(6), 1133–1145.
- Vishwanath, A., & LaVail, K. H. (2013). The role of attributional judgments when adopted computing technology fails: A comparison of Microsoft Windows PC user perceptions of Windows and Macs. *Behaviour & Information Technology*, 32(11), 1155–1167.
- Volkmar, G., Fischer, P. M., & Reinecke, S. (2022). Artificial intelligence and machine learning: Exploring drivers, barriers, and future developments in marketing management. *Journal of Business Research*, 149, 599–614.
- Wallace, R. J. (1999). Three conceptions of rational agency. *Ethical Theory and Moral Practice*, 2(3), 217–242.

- Weber, M. (2019). *Economy and society: A new translation*. Harvard University Press.
- Weber, R. (2020). Taking the ontological and materialist turns: Agential realism, representation theory, and accounting information systems. *International Journal of Accounting Information Systems*, 39, 100485.
- Williams, M. (2007). Towards a better understanding of managerial agency: Intentionality, rationality and emotion. *Philosophy of Management*, 6(2), 9–26.
- Wilson, H. J., & Daugherty, P. R. (2018). Collaborative intelligence: Humans and AI are joining forces. *Harvard Business Review*, 96(4), 114–123.
- Winikoff, M., Sidorenko, G., Dignum, V., & Dignum, F. (2021). Why bad coffee? Explaining BDI agent behaviour with valuations. *Artificial Intelligence*, 300, 103554.
- Wise, J. M. (1998). Intelligent agency. *Cultural Studies*, 12(3), 410–428. <https://doi.org/10.1080/095023898335483>
- Wooldridge, M. (2009). *An introduction to multiagent systems*. John Wiley & Sons.
- Yin, R. K. (2009). *Case study research: Design and methods* (Vol. 5). Sage.
- Zaitsava, M., Marku, E., & Di Guardo, M. C. (2022). Is data-driven decision-making driven only by data? When cognition meets data. *European Management Journal*, 40(5), 656–670.
- Zeleny, M. (2001). Obituary: Herbert A. Simon (1916–2001). *Human Systems Management*, 20(1), 3.
- Ziegler, C.-N. (2012). *Mining for strategic competitive intelligence*. Springer.
- Zwier, J., Blok, V., & Lemmens, P. (2016). Phenomenology and the empirical turn: A phenomenological analysis of postphenomenology. *Philosophy & Technology*, 29(4), 313–333.

PUBLICATIONS

PUBLICATION

|

Advancing reproducibility and accountability of unsupervised machine learning in text mining: Importance of transparency in reporting preprocessing and algorithm selection

Valtonen, L., Mäkinen, S. J., and Kirjavainen, J.,

Organizational Research Methods, 10944281221124947

Publication reprinted with the permission of the copyright holders.

Advancing Reproducibility and Accountability of Unsupervised Machine Learning in Text Mining: Importance of Transparency in Reporting Preprocessing and Algorithm Selection

L. Valtonen¹ , Saku J. Mäkinen²,
and Johanna Kirjavainen¹

Abstract

Machine learning (ML) enables the analysis of large datasets for pattern discovery. ML methods and the standards for their use have recently attracted increasing attention in organizational research; recent accounts have raised awareness of the importance of transparent ML reporting practices, especially considering the influence of preprocessing and algorithm choice on analytical results. However, efforts made thus far to advance the quality of ML research have failed to consider the special methodological requirements of unsupervised machine learning (UML) separate from the more common supervised machine learning (SML). We confronted these issues by studying a common organizational research dataset of unstructured text and discovered interpretability and representativeness trade-offs between combinations of preprocessing and UML algorithm choices that jeopardize research reproducibility, accountability, and transparency. We highlight the need for contextual justifications to address such issues and offer principles for assessing the contextual suitability of UML choices in research settings.

Keywords

unsupervised machine learning, clustering, topic modeling, pattern discovery, exploratory data analysis, data preprocessing

Organizational Research Methods
1–26

© The Author(s) 2022



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/10944281221124947

journals.sagepub.com/home/orm



¹Industrial Management, Faculty of Management and Business, Tampere University, Tampere, Finland

²Department of Mechanical and Materials Engineering, Faculty of Technology, University of Turku, Turku, Finland

Corresponding Author:

L. Valtonen, Industrial Management, Faculty of Management and Business, Tampere University, Korkeakoulunkatu 10, P.O.Box 527, 33014 Tampereen yliopisto, Finland.

Email: laura.valtonen@tuni.fi

Introduction

By some estimates, over 80% of all data available for organizations are in the form of unstructured text (Gandomi & Haider, 2015; Robinson et al., 2020). Organizations that employ these data in their decision-making have been shown to be more successful than those that do not (Cao & Duan, 2017; McAfee et al., 2012), which makes exploiting the vast quantities of available text data tempting. As a potential means of benefiting from this data, machine learning (ML) has increasingly attracted attention from both industry and academia. In organizational research, ML offers vast potential through its ability to identify patterns that researchers can apply for hypothesis development grounded in data, further exploratory inductive or abductive research, or post hoc analyses of regression results for previously undetected patterns, among other applications (Choudhury et al., 2021). However, the task of turning text into data analyzable by computers is not straightforward, as machines understand only numbers, and transforming language into numbers involves many potential pitfalls (Cambria & White, 2014).

In terms of ML applications, supervised machine learning (SML) has predominantly been used in previous research (LeCun et al., 2015). SML algorithms construct a way of mapping inputs to assigned outputs based on human-labeled training datasets. Although SML can discover useful data patterns that have gone unnoticed by more traditional methods (Choudhury et al., 2021), SML, through its use of predetermined labels, subjects data inferences to human presumptions about what is to be and what can be discovered, resulting in increased concern over the lack of accountability and transparency in such methods (Agrawal et al., 2020; Jain, 2017; Rosso, 2018; Tonidandel et al., 2018). Moreover, labeling a dataset for use in SML is a slow, error-prone, and costly process of human coding (Abney, 2007; Kobayashi et al., 2018b; Muslea et al., 2006), and current data analysis methods already struggle to manage data's exponential growth (Kuang et al., 2015). In the face of such limitations, unsupervised machine learning (UML) becomes an increasingly lucrative option for data analysis, as UML discovers data characteristics and patterns based purely on the data itself, without preassigned labels (Ziegler, 2012). Thus, UML enables the discovery of patterns independently of human presumptions, while also reducing the manual labor required (Kuang et al., 2015).

With increasing use of ML for text mining in organizational studies, researchers have a greater need to preprocess the natural language texts they intend to analyze. Preprocessing refers to the decisions made prior to the analysis itself that determine how the words will be converted into numbers in a way that decreases the complexity of the inputs in the analysis while also maintaining the interpretability and reliability of the results (Denny & Spirling, 2017). As preprocessing choices are gradually beginning to be recognized as crucial steps in ML with potentially radical impacts on the analysis results (Denny & Spirling, 2017), the search for best practices has begun (Hickman et al., 2022; Kobayashi et al., 2018a, 2018b; Schmiedel et al., 2019). For instance, Hickman et al. (2022) attend to improving the reproducibility, validity, and transparency of text mining practices in organizational SML research by creating preprocessing recommendations for text data, since heretofore the reporting of preprocessing methodology has been ununiform and obscure (Fokkens et al., 2013; Hickman et al., 2022).

However, recommendations based on SML methodology do not automatically translate into UML as such (Denny & Spirling, 2017). While there are similarities between SML and UML methodologies and their preprocessing steps and algorithms, the key difference between them lies in the fact that it is not possible to reliably evaluate the validity of UML results in a numerical, objective manner appropriate for evaluating SML results (Chang et al., 2009). In SML, objective numerical measures exist for how well a predetermined task is performed; any assortment of preprocessing choices will yield statistics on how well the data were categorized according to the predefined conceptualized categorization. The inference always remains a task performance measurement from which it is possible

to objectively ascertain the best combination of both preprocessing techniques and algorithms through testing and evaluating the performance values. On the other hand, despite existing quantitative measures for UML evaluation, it remains an inherently subjective task of interpretation (Denny & Spirling, 2017; Friedman et al., 2001). Hence, engaging in quantitative evaluation practices relevant to SML in an inherently qualitative UML research setting may result in arbitrary or even possibly cherry-picked UML methodology selections. Currently there persists a lack of differentiation between best practices for SML and UML in the literature. Thus, it is critical for the development of the field to identify and discuss the issues of UML research separately from those of SML.

The possibility of biased practices raises healthy suspicion, considering the persistent lack of transparency in contemporary UML research (Fokkens et al., 2013), as explicit consideration of the impact that preprocessing has on UML results is frequently omitted (in, e.g., Bellstam et al., 2021; Jeong et al., 2019; Kim & Chen, 2018; Westerlund et al., 2018; White et al., 2016). Similar transparency concerns persist regarding algorithm choices (in, e.g., Agrawal et al., 2020; Bellstam et al., 2021; Hannigan et al., 2019; Huang et al., 2018; Jeong et al., 2019; Westerlund et al., 2018; Zhong & Schweidel, 2020), which are no less significant. Even when preprocessing is considered, the impacts that the researchers' choices have on the results are overlooked (in, e.g., Ashton et al., 2020; Choudhury et al., 2021; Lee & Kang, 2018; Talafidaryani, 2021; Zhong & Schweidel, 2020), a practice that causes such choices to appear arbitrary. To tackle these issues, this article demonstrates the requirements for accountability and reproducibility in UML. We empirically explore how different preprocessing methodologies and algorithm choices affect UML analysis results on a common dataset in organizations—a large set of relatively short, unstructured texts (Schmiedel et al., 2019).

In this study, we build on the few exceptional studies that have considered the effects of using different UML algorithms (Erzurumlu & Pachamanova, 2020; Lee & Kang, 2018; Talafidaryani, 2021) and the effects that the preprocessing measures utilized have had on the results (Erzurumlu & Pachamanova, 2020; Huang et al., 2018; Schmiedel et al., 2019). We demonstrate the effects of both preprocessing and UML algorithm choices and show that the decisions made on both fronts have major impacts on the reproducibility, transparency, and accountability of UML research. Our results demonstrate that the best practice in UML research is meticulous contextual justification of methodological choices.

We investigate the outputs of UML data analysis regimes (i.e., combinations of preprocessing and algorithm choices) in terms of their interpretability, representativeness (Ashton et al., 2020), and computational time requirements. The qualitative differences we discover in the UML data analysis regime outputs highlight existent trade-offs between the three dimensions, and how negligence of one over the others can cause issues in research accountability, transparency, and reproducibility. In summary, we aim to alleviate the prevalent vague methodological descriptions of UML data analysis regimes that prevent future scholars from reproducing research to confirm, utilize, or improve upon the results (Haibe-Kains et al., 2020; Zhang & Shaw, 2012).

Theoretical Background

Preprocessing: Let There be no Fishing

Preprocessing is the application of various techniques to reduce data complexity and size (Denny & Spirling, 2017; Hardeniya et al., 2016). Data preprocessing decisions significantly affect the interpretability and validity of UML results, and what is applicable in one research setting may not be applicable to the text data of another (Denny & Spirling, 2017). Thus, choices need to be justified specifically in the context of the research discipline and setting (Hickman et al., 2022). Contemporary organizational research using UML lacks transparency for preprocessing choices and does not consider the theoretical and contextual factors pertinent to making these choices

(in, e.g., Jeong et al., 2019; Kim & Chen, 2018; Westerlund et al., 2018; White et al., 2016). Merely copying the preprocessing used in the previous literature without offering contextual justifications may lead to unsuitable methodological decisions (Denny & Spirling, 2017) and result in unwarranted inferences being drawn from the analysis.

Moreover, the lack of transparency regarding the choices made allows researchers to simply report only the preprocessing techniques that yield the expected or desired analytical results. The need for reproducible research and transparent data cleansing, especially in the big datasets associated with ML, is greater than ever, with instances of questionable and outright fabricated papers coming to light, as discussed by Braun et al. (2018). The lack of contextual justification for crucial data analysis steps undertaken in current UML research—steps that may significantly alter results and analytical inferences, even with a single dataset (Denny & Spirling, 2017)—allows researchers to risk data splicing (Covin & McMullen, 2019; Kirkman & Chen, 2011) and to propose potentially fished and cherry-picked data analysis regimes to achieve specific results at will.

Vague methodological descriptions undermine the reproducibility of UML research, thereby limiting the potential to ascertain its validity and reproducibility and hindering other researchers' possibilities of building on the results (Haibe-Kains et al., 2020; Zhang & Shaw, 2012). Although quantitative performance measurements exist for UML, they are often contrary to actual human evaluation (Chang et al., 2009); hence, evaluating and drawing inferences from UML results is often a time-consuming task of interpretation and heuristic argumentation (Denny & Spirling, 2017; Friedman et al., 2001). This ambiguity of analysis, combined with results that vary drastically depending on preprocessing choices (Denny & Spirling, 2017), allows researchers to make preprocessing choices favorable to a specific interpretation of the results.

There is no limit to how creative one can get while preprocessing a dataset, as arbitrary decisions can be freely made. Thus, to investigate the effects of preprocessing choices on UML data analysis regime outputs, we chose to study the set of common preprocessing steps depicted in Table 1. To further elucidate various preprocessing techniques, see Hickman et al. (2022).

In preprocessing, choices are always inherent: stop words are removed or not, and data are either stemmed, lemmatized, or neither (Hardeniya et al., 2016). These techniques can reduce data dimensionality and make computation easier, but some information will always be lost in the process of taking these steps (Hickman et al., 2022). The step that finally turns the processed text into a numerical representation for computation is vectorization. Usually, this representation is a matrix in which documents are represented as rows, and the tokens occurring per document are represented as columns. This matrix of token counts is commonly called the “bag-of-words” (BOW) representation, since it simply counts the “words” (tokens) and omits positional information (Zhang et al., 2010). The BOW document-term matrix for the three sample documents is represented in Table 2.

To emphasize the importance of rarer tokens, “term frequency–inverse document frequency” (TF-IDF) vectorization can be chosen over BOW. TF-IDF compares the frequency of tokens in individual documents to their inverse frequency over all documents in a corpus, which results in larger impacts for tokens that appear in fewer documents (Manning et al., 2008; Salton & Buckley, 1988). An example of TF-IDF vectorization is presented in Table 3. TF-IDF ignores semantics or positional information and can therefore be considered a form of BOW vectorization. Both vectorizations assume that terms are more important to a document the more frequently they appear in it, and TF-IDF assumes that rare tokens are more meaningful than common ones (Manning et al., 2008; Salton & Buckley, 1988).

A common, yet possibly overemphasized (Landauer et al., 1997), criticism of BOW is that semantic information about the text data is lost (Fu et al., 2018; Sinoara et al., 2019; Zhao & Mao, 2018). If retaining some semantic information is prioritized, then token order information can be acquired with word sequences, such as chunks of words or n-grams (Hickman et al., 2022; Kobayashi et al., 2018b; Zhong & Schweidel, 2020). N-grams and chunks combinatorically increase the data size. The order

Table 1. Text Preprocessing Choices Present in Machine Learning, Following Denny and Spirling (2017) and Hardeniya et al. (2016).

| Choice | Function | Execution | Example |
|---------------------------|--|--|---|
| Data cleansing | Dataset-specific actions such as removing noise, outliers, and specific characters, and unifying encoding. | Dataset specific. Large datasets often need more cleansing measures; for guidance, see Braun et al. (2018). | Consider “Yesterday I ate many #apples. 🍏” To avoid encoding errors, the emoji and # are removed, resulting in: “Yesterday I ate many apples.” |
| Tokenization | Splitting a string of text into smaller substrings called tokens. Most commonly, single words. | Can include optional lowercasing and punctuation removal. Can be customized if the data have special requirements. | “Yesterday I ate many apples.” becomes the following list: “yesterday,” “i,” “ate,” “many,” “apples,” or “Yesterday,” “i,” “ate,” “many,” “apples,” “. . .” |
| Stopword removal | Removing tokens that do not contribute information to reduce data dimensionality, e.g., common tokens and rare tokens. | Unnecessary, but can be done with lists of stop words, by removing a percentage, or counts of common/rare tokens. | The list would become “yesterday,” “ate,” “apples,” if a stop word list is used for removal, since “many” and “i” are common stopwords. |
| Stemming or lemmatization | Reducing the tokens to their base form. Stemming removes endings in a rule-based manner. Lemmatization is a more sophisticated, data-based approach. | Unnecessary to perform, in which case all different token forms remain in the dataset, maintaining high dimensionality. | Stemming removes endings such as “-s” and “-ing,” and the list becomes: “yesterday,” “ate,” “apple.” The lemmatized list is: “yesterday,” “eat,” “apple.” |
| N-grams or chunks | Merging tokens together based on proximity (n-grams) or by grammatical rules (chunks). Retains some positional and relational information about tokens in the data. | Unnecessary, in which case, tokens are often single words called unigrams. N-grams can consist of as many adjacent tokens as specified, and chunks can be extracted as desired. | N-grams/bigrams of the original sentence: “yesterday i,” “i ate,” “ate many,” “many apples.” Chunking would likely extract “yesterday,” “i,” “ate,” and “many apples” as noun chunks. |
| Vectorization | Turning text into a vector representation for computation. Usually, there is a “document-term” matrix in which documents are represented as rows and tokens occurring per document as columns. | Often either bag-of-words, in which token counts are used as is, or term frequency inverse document frequency, in which tokens are weighted by their relative importance to a specific document. | See Tables 2 and 3. |
| Phase order | The order in which the preprocessing steps are executed. | Most steps can be performed in different orders. Vectorization is usually the final step. | If n-grams were extracted after stopword removal and lemmatization, the result would simply be: “yesterday eat,” “eat apple.” |

Table 2. Example of bag-of-words Vectorization.

| | this | is | a | sentence | another | there | also | third |
|------------------------------------|------|----|---|----------|---------|-------|------|-------|
| This is a sentence. | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| There is also another sentence. | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 |
| There is also this third sentence. | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 |

Table 3. Example of term-frequency-inverse document frequency Vectorization.

| | this | is | a | sentence | another | there | also | third |
|------------------------------------|------|-----|-----|----------|---------|-------|------|-------|
| This is a sentence. | 1/2 | 1/3 | 1/1 | 1/3 | 0 | 0 | 0 | 0 |
| There is also another sentence. | 0 | 1/3 | 0 | 1/3 | 1/1 | 1/2 | 1/2 | 0 |
| There is also this third sentence. | 1/2 | 1/3 | 0 | 1/3 | 0 | 1/2 | 1/2 | 1/1 |

in which preprocessing steps are applied also has a significant impact on the final processed data (Denny & Spirling, 2017). For instance, making chunks only after removing stop words can compound and associate tokens that originally had an insignificant connection to each other by removing the tokens between them. The preprocessed data were then fed into a UML algorithm. Differences in preprocessing choices already yield diverging results with a single algorithm (Denny & Spirling, 2017), let alone with different UML algorithms, which yield even further diverging results specific to each. Suddenly, the possibilities for proposing fished results increase combinatorically: different preprocessing outputs can be passed to different UML algorithms to consider and choose which combination yields the desired results.

Algorithm Selection: Following the Herd

In contemporary research, careful consideration of the choice of the UML algorithm itself is uncommon. Instead, most implementations of UML in organizational research literature use a specific algorithm, namely the latent Dirichlet allocation (LDA) topic model (Banks et al., 2018; Blei et al., 2003), without further consideration of other UML algorithms (in, e.g., Choudhury et al., 2020; Hannigan et al., 2019; Huang et al., 2018; Jeong et al., 2019; Westerlund et al., 2018; Zhong & Schweidel, 2020). LDA is a *topic model*; topic models are a group of probabilistic algorithms that create probability distribution-based lists of text documents based on how often tokens appear together in the same contexts (Mohr & Bogdanov, 2013). One such token distribution list is called a topic. Topic models attempt to discover “substantively meaningful categories” (Mohr & Bogdanov, 2013) as well as abstract, latent topics that exist within text data (Blei et al., 2010).

This arbitrary use of LDA as a default algorithm poses transparency issues similar to failing to report the preprocessing steps, allowing for the possibility that results will be biased or overlook important contextual considerations. This is also problematic, since the dominance of LDA has no basis in extant research comparing possible topic-modeling algorithms (Ashton et al., 2020), not to mention other types of methodologies. Choosing between UML algorithms should require careful consideration and contextual justification (Ashton et al., 2020; Schmiedel et al., 2019), rather than using LDA as a seemingly arbitrary default method.

If topic models are probabilistic ways of splitting a body of texts into subsets of topics, a deterministic way to accomplish the same is through *clustering*. Clustering algorithms create clusters of similar text documents based on the similarities of the tokens used in the documents. A text document

can be any piece of text. The aim of clustering is to represent a dataset as smaller, distinct groups within which the characteristics of the data points (texts) are alike and different from those in other groups. These smaller, separate groups of data are called clusters (Aggarwal & Zhai, 2012; Srivastava & Sahami, 2009). The main difference between topic modeling and clustering is that topic modeling is *probabilistic*, whereas clustering results are *deterministic*—a document either belongs to a certain cluster or does not. In topic models, all words found in a corpus have a probability of belonging to a topic, and all text documents have a probability of belonging to all topics.

On the rare occasion when the topic modeling selection is explained in the literature (in, e.g., Lee & Kang, 2018; Talafidaryani, 2021), the reason given for choosing topic modeling over a deterministic methodology such as clustering is the possibility of a text document containing multiple topics. However, with the most common organizational data for short texts, contextual justification is required for the assumption that multiple topics exist within a single document in a research setting (Ashton et al., 2020; Schmiedel et al., 2019).

Moreover, since topic models are supposed to discover *latent* topics, it is possible that topics with no documents strongly affiliated with them will arise. In such situations, it is still possible that a researcher might unjustly infer the existence of such a topic from the results, even if its existence is uncertain according to the topic model itself. Topic modeling’s probabilistic vagueness can, at times, make it difficult to pinpoint why certain documents reflect a specific topic. In contexts that require data analysis interpretability for the purposes of decision-making (Jain, 2017; Lee & Shin, 2020), the ability to explain results and offer transparency is important (Lee & Shin, 2020), and a probabilistic methodology just might not be up to the task. However, such contextual considerations are rare in current research.

Topic models—and clustering algorithms—can be evaluated by their *interpretability* (Ashton et al., 2020), which is a task of heuristic argumentation (Denny & Spirling, 2017; Friedman et al., 2001). The interpretability of UML outputs is mainly evaluated in contemporary research by token list representations of each topic *without looking at the documents* themselves. This readily creates confusion and transparency issues that need to be resolved, as shown in recent examples (Ashton et al., 2020; Schmiedel et al., 2019). Failing to show the links between the generated topics and the corpus creates possible transparency issues in which top tokens represent nonexistent or misleading topics when compared to the actual data. To evaluate this aspect, *representativeness* can be assessed (Ashton et al., 2020). Here, representativeness refers to what Ashton et al. (2020) defined as a measure of “when evaluating a selection of documents, do they reflect the topic that was understood based on the keywords?” (p. 111). Here, the keywords indicate the top token list representation of a cluster or topic.

Moreover, the computational requirements of algorithms can differ radically and become impractical with increasing amounts of data (Xu & Tian, 2015), while some UML data analysis regimes behave combinatorically with regard to data dimensionality, for instance, with n-grams. This imposes constraints on UML data analysis regime choices due to computational requirement limits. Hence, an assessment of the possible trade-offs between quick, interpretable, or representative results may be required. As we highlight the degree of transparency required for reproducible research by demonstrating the interpretability and representativeness effects of different preprocessing choices with two types of UML methodologies—deterministic and probabilistic—we also call for transparency regarding the computational requirements of the UML data analysis regimes used.

Methodology

Data

Since we want our research to be as generalizable as possible, we use typical and realistic unstructured data from news sentences as our corpus. This corpus was acquired by retrieving news on digital camera manufacturers using keyword retrieval over company names from the LexisNexis database.

The resulting corpus was then fed into further preprocessing phases, allowing us to collect a corpus of over a million news sentences. The requisite computational capacity of some implemented UML algorithms is very sensitive to sample size (Xu & Tian, 2015); hence, to make computation plausible in terms of time, random samples of 581, 1,163, 2,907, 5,813, and 11,627 news sentences were extracted. For reference, the LDA performance, as a ubiquitous benchmark, has been reported to stabilize after 1,000 documents (Mohr & Bogdanov, 2013; Schmiedel et al., 2019). We limited our data sample size to 11,627, since we deemed this sufficient, and increasing data samples beyond this extended the computational requirements of the most computationally demanding algorithm to over 72 h without any perceivable further benefit.

Preprocessing

Table 4 presents the preprocessing choices we made and their justifications in terms of the context and aims of the research. All preprocessing was performed using the spaCy library for Python with its standard methods, except for n-gram and chunk extraction and vectorization. For the n-grams and chunks, a Python library called textacy was used due to its compatibility with spaCy. Vectorization was performed using the scikit-learn library for Python. These fast and robust libraries were chosen because ready-to-use statistical packages are commonly employed and recommended in the field of organizational research (Kobayashi et al., 2018a, 2018b; Schmiedel et al., 2019), and we want our research to be as accessible as possible.

UML Algorithm Selection: Topic Modeling vs. Clustering

To explore topic modeling, three different algorithms were considered and compared: LDA (Blei et al., 2003, 2010) as the ubiquitous benchmark method (Mohr & Bogdanov, 2013); latent semantic indexing (LSI) (Deerwester et al., 1990) as the predecessor and usual comparison to LDA (Ashton et al., 2020); and the hierarchical Dirichlet process (HDP) as a newer proposition to overcome the limitation of having to predefine the number of topics to be created in the parametric algorithms, since the HDP autonomously defines the number of topics to create (Blei et al., 2010). All topic model algorithms were imported and used in their standard form from the Gensim topic modeling library for Python. Gensim is dedicated to topic modeling and has all the algorithms under consideration ready to use. Default parameters were used to keep the approaches general. The mathematical details of the methods are well covered in the papers mentioned above and are therefore not covered in detail here.

To study clustering, the popular and straightforward K-means algorithm was chosen as a parallel to LDA, since it both creates a similar output (Ziegler, 2012) and requires the number of clusters to be created to be set as a predefined parameter (Xu & Tian, 2015). The K-means algorithm is iterative: the first set of potential cluster centers is a random guess, after which all data points are assigned to the center closest to them. Cluster centers are updated to be the average “position” of all the data points that were closest to the previous center until a set convergence criterion is met (Friedman et al., 2001).

We compare the affinity propagation (AP) clustering algorithm to K-means in a manner similar to that in which HDP is compared to LDA. AP does not require the number of clusters to be created as a preset parameter; however, AP is complex timewise, sensitive to its required set of parameters, and not well suited to large datasets. AP regards every data point as a potential cluster center and a specified distance measure between data points as their affinity. In practice, this means that the higher the number of data points that are similar to a certain data point, the higher the probability of that data point being a cluster center (Xu & Tian, 2015).

The density-based mean shift (MS) clustering algorithm was studied because it can be compared by cluster centers to the other chosen clustering algorithms. The idea of density is simple: points close to each other—constituting a “dense” area of data points—are grouped together as a cluster.

Table 4. This Study’s Preprocessing Choices.

| Phase | Actions performed |
|----------------------------|---|
| Data cleansing | The text was uniformed into Unicode Transformation Format – 8-bit encoding to avoid encoding errors. News pieces were split into sentences with spaCy’s tokenizer. The first few sentences of each news piece were retained in the dataset. This was judged not to subvert data informativeness. With spaCy’s ready methods, we removed obvious defining tokens (company names, named entities, special characters, numbers, and emails) from each text to study the ability of the methods to discover unnoticed patterns. Results indicating that the data are split by the search terms used or company actor names are <i>a priori</i> information and not a new discovery. |
| Tokenization | Three tokenizations were used: unigram, bigram, and chunks. SpaCy used unigrams by default. Punctuation was removed in all cases because it played no role in the data. If one were studying data in which punctuation is used to structure information, it would be important and would require more attention. |
| Stopword removal | SpaCy-defined stop words were removed in all tokenizations because testing the set without stopword removal yielded topics and clusters containing mostly stopwords. We could not specify the correct percentage of common or rare words to use to avoid losing potentially important common words, such as “patent,” so using such methods was implausible. |
| Stemming or lemmatization | Except for chunks, spaCy’s default lemmas were used. We chose lemmatization because it performs more reliably than stemming (Hardeniya et al., 2016) and better reduces data dimensionality, which is important for studying n-grams that combinatorically expand the data size. Chunks were tested with lemmatization and mostly became obscure. |
| N-gram or chunk extraction | Noun chunks were extracted with ready spaCy methods. Textacy methods were used to extract bigrams for n-grams and verb chunks that matched the regular expressions pattern: r“<VERB>*<ADV>*<PART>*<VERB>+<PART>*”. Extracting bigrams was deemed sufficient to demonstrate n-gram behavior, leaving longer token combinations for chunk tokenization. To observe the behavior of chunk tokenization in general, both noun and verb chunks are valid, since neither has any significance over the other for our general purposes. |
| Vectorization | Both term frequency-inverse document frequency and bag-of-words vectorizers were used and compared for different tokenizations. Other vectorizations were judged to be too niche for our purposes, such as Word2Vec, which builds on bag-of-words. |
| Phase order | First, punctuation was removed. For extracting chunks, stopwords were not removed and tokens were not lemmatized, since chunks become easily unintelligible and undetectable with lemmatization and stopword removal. For tokens outside of chunks and in both uni- and bigrams, stopwords were removed and the remaining tokens were lemmatized, after which unigrams and bigrams were formed. |

Density-based clustering methods are very sensitive to their required set of parameters and require much computational memory (Xu & Tian, 2015).

All clustering algorithms were available on scikit-learn for Python and were used as such. AP needed an affinity measure to be provided; Euclidean distance was chosen for this purpose since it is the default in K-means. Otherwise, all scikit-learn default parameters were used to keep the approaches general.

Preprocessing and UML Algorithm Combinations

In section “Preprocessing”, we presented the procedures that left us with two preprocessing steps to explore: tokenization and vectorization. For tokenization, three different methods were explored: unigrams, n-grams, and chunks. For vectorization, two different methods are explored: BOW and TF-IDF. Thus, we have six different preprocessing combinations. In Section “UML Algorithm Selection: Topic Modeling vs. Clustering”, we presented two UML approaches: topic modeling and clustering. Three topic modeling algorithms were explored—LDA, LSI, and HDP—together with three clustering methods—K-means, AP, and MS. All six UML algorithms were fed with all preprocessing combinations to generate an output. All 36 possible UML data analysis regimes are depicted in Figure 1.

Each output was then evaluated in terms of three dimensions: interpretability, representativeness, and time requirements. A detailed time requirement analysis methodology is provided in the supplemental material. The interpretability and representativeness of all UML data analysis regimes were evaluated according to Figure 1 by first comparing the effects of the different tokenizations made within the vectorizers. Tokenizer effects were studied separately for the BOW and TF-IDF vectorizations. Tokenization comparisons were followed by vectorization effect comparisons, considering how tokenization was affected as well. Interpretability refers to how well a human reader can conceptualize an overall sensible theme from the resulting token lists of the topics and clusters created, while representativeness refers to whether a selection of data documents reflects the topic that was understood based on the topic/cluster representation (Ashton et al., 2020). Clustering and

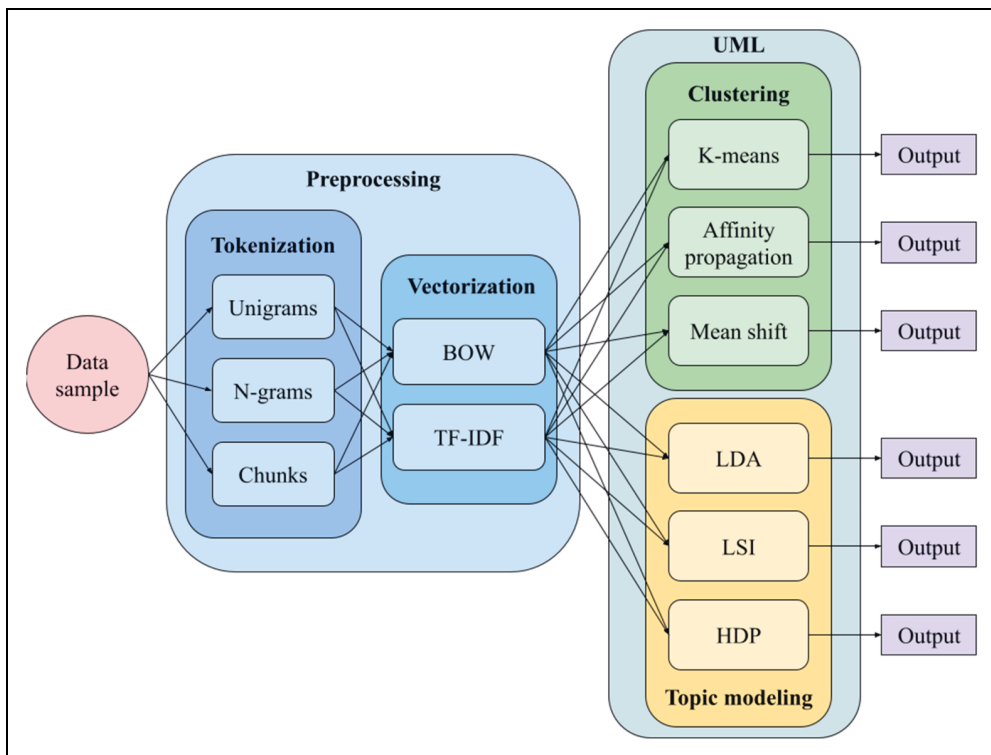


Figure 1. Description of the preprocessing and UML algorithm combinations and UML data analysis regimes used in this study. Note: BOW = bag-of-words; TF-IDF = term frequency-inverse document frequency; LDA = latent Dirichlet allocation; LSI = latent semantic indexing; HDP = hierarchical Dirichlet process.

topic modeling create similar outputs (lists of the most representative tokens of a cluster/topic) and can be assessed similarly.

Interpretability Assessment. We used a similar evaluation method to that proposed by Ashton et al. (2020), by which two researchers independently coded each topic and cluster output. For output interpretability, all 36 regimes' topics and clusters were coded based on the top token lists into the following categories: "interpretable," "uninterpretable," and "uncertain." For a topic or a cluster to qualify as interpretable, the answer to the question "Does this represent a coherent and understandable concept?" had to be positive. To demonstrate the assessment process, we show some samples of our results below. For instance, the K-means clustering algorithm with unigram tokenization and TF-IDF vectorization discovered the following cluster:

"model, new, market, price, launch, sell, announce, plan, business."

This cluster was assessed as interpretable, and the concept it was interpreted to represent was "new model launches." For uninterpretable results, the answer to the interpretability question had to be negative. For instance, HDP with unigram tokenization and TF-IDF vectorization discovered the following topic:

"p5, guru, cesthe, biness, capital, rearrangement, leaderinwait, ic, target."

This topic was assessed as uninterpretable since it was not possible to interpret any coherent concept from it. For the uncertain results, no certain answers existed for the interpretability question. For instance, LDA with unigram tokenization and TF-IDF vectorization discovered the following topic:

"analyst, grow, printer, business, sale, technology, estimate, company, equipment,"

which can be judged to either concern analysts making estimates about printer sales or analysts estimating the business and sales of a company that also happens to be in the printer business. The same UML data analysis regime also discovered the following topic:

"tv, right, material, lawsuit, team, seek, subsidiary, expensive, head,"

which might be assumed to concern lawsuits regarding material rights to television. However, to include later tokens, such as "seek" and "expensive," the concept would have to concern "teams seeking expensive lawsuits regarding material rights to television." Perhaps such a topic exists, but interpreting it is not clear, and it requires guesswork. From both examples, interpreting a concept seems like a leap of faith. In summary, uncertain results were not clear and required guessing at either the concept itself or between different possibilities.

The aggregate output of each UML data analysis regime was then coded into the following categories: poor, moderate, and good. For "poor" outputs, the clear majority of all topics or clusters were uninterpretable. For "moderate" outputs, no clear majority appeared for either interpretable or uninterpretable clusters or topics. This was the case when most results were uncertain. For "good" outputs, the clear majority of all topics and clusters were interpretable.

The same coding scheme was implemented by both coders on the whole dataset and discussed afterwards to resolve all discrepancies and ensure that the coders understood the scheme similarly before independently coding the complete dataset again according to the revised scheme. After coding the interpretability results, the differences in the final results were discussed. Intercoder agreement on the aggregate output interpretability assessments was 86%, which is acceptable

(Lombard et al., 2002). Among the 36 results, the coders only clearly disagreed on one output. They also differed in the coding of four others, but it became apparent through discussion that these were borderline cases straddling the line between poor and moderate.

Conflicts were resolved by the coders explaining to each other why they saw certain topics or clusters as interpretable or not and finding a compromise; as is typical for UML (Denny & Spirling, 2017; Friedman et al., 2001), all assessments are inherently subjective. There is no objective truth as to how or whether a cluster or topic can be interpreted. For instance, a topic represented by the tokens “job, cut, plant” could be interpreted to concern factory layoffs by one coder and the gardening profession by another. Both coders would correctly judge the topic to be interpretable despite their interpretations being different. Similarly, a topic of “biennial, bolt, medium, variety” could equally validly be uncertain for one coder and interpretable for the other (who is assumed to be more acquainted with gardening). Explaining the gardener coder’s perspective to the uncertain coder may prompt them to agree with the interpretable assessment. Interpretability comparisons were made based on the dataset with 5,813 documents because it was the largest set that could be run for all methods within a reasonable amount of time (a more detailed example of interpretability coding is presented in the supplemental material). Altogether, 8,564 topics and clusters were covered in the interpretability coding process.

Representativeness Assessment. Representativeness was assessed in a similar fashion to interpretability assessment after the results for the latter were attained. To assess representativeness, the outputs from all 36 UML data analysis regimes were coded into the following categories: “representative,” “nonrepresentative,” and “uncertain.” For a document assigned to a topic or a cluster to qualify as representative, the answer to the question “Do the contents of this document represent the concept interpretable from the topic or cluster to which it has been assigned?” had to be positive. To evaluate topic modeling, the topics to which documents were assigned with the highest probability were studied. To again use samples from our results for demonstration purposes, for the previous K-means cluster,

“model, new, market, price, launch, sell, announce, plan, business,”

the document “*to introduce online only models words to tackle the conflict between online and offline retailers over the pricing of its products will introduce new models to be sold exclusively through ecommerce portals said president and CEO of*” was assessed to qualify as representative, since its content matches the interpreted context of the cluster, namely new model launches.

For nonrepresentative results, the answer to the representativeness question had to be negative. For instance, the document “*The of the itself solidifying with words reinventing its business model for longterm growth has announced the creation of a cuttingedge broadcast solutions package and a significant expansion of inhouse television production capabilities*” assigned to the same cluster was assessed as nonrepresentative, since although it touches on the topic of creating something new, it does not concern a new model launch. On the other hand, the document “*We would continue to avoid the stock as smartphone sales are falling off faster than expected and we are sceptical that new models will be able to replace lost profits said analyst*” was assessed as uncertain because while the document cannot be said for certain to represent a new model launch, it clearly does touch on the concept of launching new models. Uncertain representativeness results required making similar leaps of faith as were seen in the interpretability evaluation.

Uninterpretable topics or clusters were naturally nonrepresentative as well, since a document cannot represent an uninterpretable concept. For instance, for the previous HDP topic,

“p5, guru, cesthe, binsess, capital, rearrangement, leaderinwait, ic, target,”

the document “*Following the announcement in of the creation of a new global company structure has continued to integrate its operations under distinct organisations and*” was assessed as nonrepresentative. For topics and clusters of uncertain interpretability, all types of representativeness can be present. For instance, for the previous LDA topic:

“analyst, grow, printer, business, sale, technology, estimate, company, equipment,”

The document “*My prediction is that overnight there should be a good market for secondhand’s as current users upgrade and less demanding firsttime users scout for a cheap laser printer*” was assessed as representative, since it clearly represents the concept of printer business estimates—a concept that may be interpretable from the topic. The document “*the consensus estimate for may also be lowballing the company again*” assigned to this topic was assessed as nonrepresentative, since the document is too abstract to be connected to the core concept of the topic: the printer technology business. The document “*Analysts said the sale might have been accelerated by’s woes and ongoing weakness in hardware sales after the biggest technology services company reported a percent drop in revenue from on*” assigned to this topic was assessed as uncertain because, while it touches on estimates of technology business, printers are specifically important to the topic and are not present in the document.

The aggregate output of each UML data analysis regime was then coded into the following categories: poor, moderate, and good. Here, “poor” indicated a result in which the clear majority of documents did not “reflect the topic that was understood based on the keywords” (Ashton et al., 2020, p. 111); in other words, they were not representative. Conversely, “good” indicated a result in which the majority of documents assigned to topics and clusters were representative. A “moderate” result implied an output that could not be said to have a clear majority of topics, either representative or nonrepresentative. For instance, this happened whenever the majority of documents were of uncertain representativeness.

To ensure agreement on the coding scheme, the coders followed a similar process as in the interpretability assessments. In their final analysis, both coders studied the same 1,000 documents for each UML data analysis regime for representativeness (a detailed example of representativeness coding is presented in the supplemental material). Altogether, 36,000 documents were covered in the representativeness coding process. Intercoder agreement on the aggregate representativeness assessments was 81%, which was deemed sufficient (Lombard et al., 2002). The differences in coding were discussed, and all differences were borderline results that could have been coded into proximate categories according to both coders. Conflicts were resolved via discussion to determine the final categories in a similar manner to interpretability evaluation because representativeness evaluation is also an inherently subjective task. Using the previous example from section “Interpretability Assessment”, the coder who interpreted the topic “job, cut, plant” to represent gardening would rate documents assigned to the topic differently from the coder who interpreted the topic to concern factory layoffs. The former may assess the document “Compost can be used to replenish soil nutrients” as being representative, while the latter would likely assess it as nonrepresentative. Both are equally valid assessments, unless the original data are consulted.

In the interpretability and representativeness assessments, the number of topics or clusters that needed to be created was set to 50 for the parametric algorithms that required such a value; this number was used based on iterative testing, which showed that it produced non-repetitive topics for LDA. LDA behaved worse as more topics were created. When the number of topics was set to 150, all topics had the same top tokens. Other methods were not this volatile in terms of the parameter, and the output quality remained stable.

The Python code we used for our methodology is provided in the supplemental material, along with a sample of the data. However, the supplemental code retrieves n-grams and chunks using

spaCy's methods only, because the support was ceased for the originally employed textacy Python library. The two Jupyter Notebook files created for the supplemental material have been heavily commented on, and the code is simple enough for a novice to play around with. Despite the textacy-free method for attaining chunks and n-grams, the results were similar to those we drew from the original code. It is extremely important to note that in light of the results presented in the following section, the supplemental code in its present form should not be considered applicable to anything other than reproducing this study's methodology.

Results

Tables 5 and 6 present summaries of the interpretability, representativeness, and computational time requirement (speed) results for all UML data analysis regimes explored. Table 5 presents the results for the UML data analysis regimes using BOW vectorization, while Table 6 presents those using TF-IDF vectorization. The complete time requirements analysis used to yield these results is provided in the supplemental material.

We now concentrate on the major differences among the UML data analysis regime outputs for the three aspects mentioned above and leave detailed descriptions of the outputs for the supplemental material.

Interpretability

No common trends are identifiable among the tokenizations from Tables 5 and 6. Most UML data analysis regimes had a minor or no effect on tokenization changes. Vectorization, however, was more influential on UML data analysis regime interpretability than tokenization. For topic modeling methods, interpretability could be increased or retained using BOW instead of TF-IDF. HDP with n-grams was the only exception. There were no general trends among clustering methods, and the effects of vectorization were algorithm specific. However, in multiple cases, TF-IDF vectorization degraded n-gram and chunk tokenization interpretability due to overly specific or nonsensical tokens in the outputs.

Any interpretability differences due to preprocessing choices were overshadowed by those due to algorithm choice. The output could be tweaked with preprocessing, but the algorithm itself mostly determined whether the UML data analysis regime was rated good or poor on the evaluation scale. Topic modeling—and especially LDA—discovered interesting and thought-provoking patterns when the results made sense in the interpretability assessments. LDA appeared to split and merge otherwise clear topics, and often, the effort required to interpret these topics sparked realizations. For instance, LDA with unigram tokenization and TF-IDF vectorization discovered the following topic:

“loss, forecast, fall, global, demand, hit, job, rise, cut.”

This topic implies the presence of a potentially interesting relationship between the demand forecasts and job cuts concepts. In certain research settings, it may be inferred that falling global demand and job cuts are correlated or that perhaps even a causal relationship may exist between the two. Clustering, as a deterministic method, discovered crude, simple clusters compared to topic modeling. For example, K-means (also with unigram tokenization and TF-IDF vectorization) discovered the following cluster around the theme of job cuts:

“job, cut, production, plan, plant, facility, say, announce, company, manufacture.”

Table 5. Results with bag-of-words Vectorization.

| Bag-of-words (BOW) vectorization | | | | | | | |
|----------------------------------|----------------------|------------------|--------------------|----------|------------------|--------------------|----------|
| | | Interpretability | Representativeness | Speed | Interpretability | Representativeness | Speed |
| Unigram | K-means | Good | Moderate | Moderate | LDA | Poor | Fast |
| | Mean shift | Moderate | Poor | Slow | LSI | Poor | Fast |
| | Affinity propagation | Moderate | Moderate | Slow | HDP | Poor | Moderate |
| N-gram | K-means | Good | Poor | Moderate | LDA | Poor | Fast |
| | Mean shift | Moderate | Poor | Slow | LSI | Poor | Fast |
| | Affinity propagation | Poor | Moderate | Slow | HDP | Poor | Moderate |
| Chunk | K-means | Good | Poor | Moderate | LDA | Poor | Fast |
| | Mean shift | Moderate | Poor | Slow | LSI | Poor | Fast |
| | Affinity propagation | Poor | Poor | Slow | HDP | Poor | Moderate |

Note: LDA = latent Dirichlet allocation; LSI = latent semantic indexing; HDP = hierarchical Dirichlet process

Table 6. Results with term frequency-inverse document frequency (TF-IDF) vectorization.

| | | Term frequency-inverse document frequency (TF-IDF) vectorization | | | | | | |
|---------|----------------------|--|--------------------|----------|------------------|--------------------|----------|--|
| | | Interpretability | Representativeness | Speed | Interpretability | Representativeness | Speed | |
| Unigram | K-means | Good | Good | Moderate | LDA | Moderate | Fast | |
| | Mean shift | Good | Poor | Slow | LSI | Good | Fast | |
| | Affinity propagation | Poor | Moderate | Slow | HDP | Poor | Moderate | |
| N-gram | K-means | Moderate | Poor | Moderate | LDA | Poor | Fast | |
| | Mean shift | Good | Poor | Slow | LSI | Moderate | Fast | |
| | Affinity propagation | Poor | Poor | Slow | HDP | Poor | Moderate | |
| Chunk | K-means | Moderate | Moderate | Moderate | LDA | Moderate | Fast | |
| | Mean shift | Good | Poor | Slow | LSI | Good | Fast | |
| | Affinity propagation | Poor | Poor | Slow | HDP | Poor | Moderate | |

Note: LDA = latent Dirichlet allocation; LSI = latent semantic indexing; HDP = hierarchical Dirichlet process

The cluster is straightforward; jobs are being cut at a facility. Unless “job cuts” and “facility” are treated as separate concepts, no potential correlations or causal relationships between the concepts can be inferred. In either case, the correlation’s abstraction level between the event and implied location may be clearer to interpret than the correlation implied by topic modeling. To summarize, data patterns can be inferred from clustering results, but they are more obvious and less intricate than topic modeling results. This difference was consistent across all the UML data analysis regimes studied, albeit to different degrees.

Representativeness

For poorly interpretable results, representativeness was also naturally poor. If most topics or clusters are non-interpretable, they cannot contain documents that represent the concept interpretable, as given by the definition of representativeness. Hence, the focus on representativeness results is on the regimes that yielded good or moderate interpretability results.

The effects of tokenization were clearer for representativeness than for interpretability. Generally, unigrams created the most representative results. Depending on the chosen vectorization, either n-grams or chunks degraded the representativeness more than unigrams. Using n-grams or chunks to emphasize rarer tokens in topics and clusters can result in notably poorer result representativeness, while not necessarily impacting the interpretability of the results.

Similar trade-offs were present in the vectorization choices. While the vectorization effects varied by algorithm for clustering, topic modeling *representativeness* was notably worse with BOW vectorization than with TF-IDF. This contrasts with interpretability, which improved when using BOW with topic models. However, despite vectorization having clearer impacts on representativeness than interpretability, the chosen algorithm was the major determinant of how the UML data analysis regime was rated. In general, representativeness was better for clustering than topic modeling.

Of the parametric methods, the LDA and LSI topic modeling algorithms’, lower overall representativeness compared to K-means clustering was expected. While clustering simply groups similar documents together, topic modeling is an exploratory method (Schmiedel et al., 2019) meant to discover latent, hidden, topics and patterns. The nature of topic models allows for the emergence of topics with documents assigned to them with varying probabilities. Topics that had documents assigned to them with a high probability were noticeably more representative than topics that lacked high probabilities for any document — degrading the overall representativeness of topic modeling.

For example, for the same topic from section “Interpretability”.

“loss, forecast, fall, global, demand, hit, job, rise, cut,”

the three documents assigned to it with the highest probability were: “*Operating profit totalled compared with a loss in while revenues inched higher to*” with 67%, “*market tracker cut its forecast from to growth for global spending on information technology*” with 52%, and “*In Q2 it added the and expanded into global markets*” with 51%. To compare this topic modeling behavior to clustering, using the same cluster from section “Interpretability”.

“job, cut, production, plan, plant, facility, say, announce, company, manufacture,”

the three documents assigned to it were “*The massive reorganization which will cut jobs of the workforce revamp retirement benefits and restructure internal business units is expected to save the company beginning in,*” “*The global cuts are to take place over as part of its integration of in its operations,*” and “*On the world’s largest consumerelectronics manufacturer said it would cut jobs and close plants across the world in and abroad.*” Contrary to the LDA topic, no document was assessed as nonrepresentative.

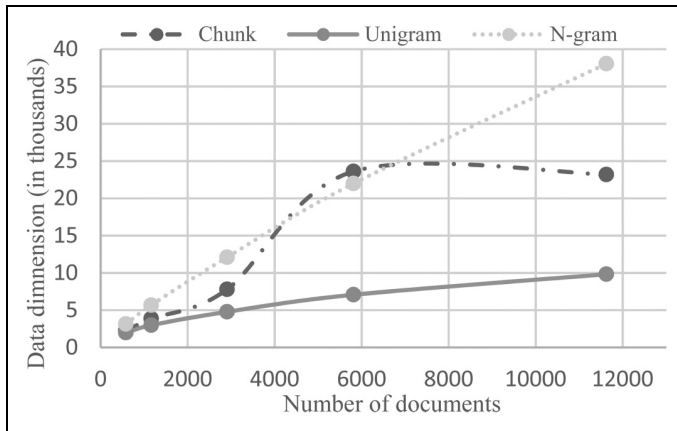


Figure 2. Vectorization matrix dimension for different tokenizations.

This demonstrates the main difference discovered between probabilistic topic modeling and deterministic clustering. While the presented topic interpretation suggested that a potential relationship between changes in global demand and job cuts was present in the data, none of the documents assigned to this topic (over other topics) suggested such a relationship. Therefore, while topic modeling can function as a tool to explore hitherto unnoticed data patterns, their existence cannot necessarily be justified based on the actual documents in the data. Clustering had no such issue. Located patterns in clustering are often cruder than topic modeling, but clustering discovers more representative patterns that are easily interpreted and explained.

To summarize, while certain preprocessing or algorithm choices may not make major differences to UML data analysis regime outputs individually, the combinations thereof will. With different algorithms, different preprocessing choices become dominant. The entire UML data analysis regime, including its contextual factors, output analysis, and research setting, should all be considered concurrently and synchronously.

Computational Time Requirements

Preprocessing steps reduce data size and complexity (Denny & Spirling, 2017; Hardeniya et al., 2016). In our setting, data size refers to the BOW or TF-IDF document-term matrix dimensions, which is the number of tokens multiplied by the number of documents in the sample. Figure 2 demonstrates the uni-, bi-, and chunk tokenization dimensions. Dimensionality mainly concerns computational requirements.

The data dimensionality is clearly reflected in the results. The combinatorial nature of chunks and especially n-gram tokenizations require more computational resources. Vectorization, conversely, had no effect on time requirements, as the weights in the data matrices simply changed. Algorithm choice was again the major determinant of where the result ranked on the slow, moderate, or fast scale. A clear split emerged for topic modeling and clustering algorithms. Topic models ran faster than clustering algorithms. A detailed time requirement analysis results exploration is provided in the supplemental material.

Discussion

Regarding UML algorithm selection, we found that probabilistic topic modeling can discover nuanced and surprising patterns, but the interpretability and representativeness of the outcome is

often abstract and vague. For topic modeling, interpretability could be improved by degrading representativeness with vectorizer choice. However, clustering—a deterministic approach—discovered less surprising and more obvious patterns that were lucid (i.e., easily interpreted and explained) and representative. When comparing probabilistic and deterministic methodologies, the probabilistic methodology was always significantly lighter and faster computationally. Varying UML data analysis regimes creates trade-offs that should be accounted for—specifically whether they are desired or tolerable considering the goals of the analysis—when conducting data analysis. In contemporary research, such considerations are rare (in, e.g., Bellstam et al., 2021; Jeong et al., 2019; Kim & Chen, 2018; Westerlund et al., 2018; White et al., 2016).

It could be argued that topic modeling in our results behaved as expected, and that clustering is not optimized for unsupervised text analyses, unlike topic modeling. However, considering that contemporary UML data analysis rarely assesses representativeness or compares the created topics or clusters to their generative data, researchers may, with various preprocessing and algorithm choices, iterate for the one UML data analysis regime that yields the preferred topics or clusters without noting whether the data justifies the results. This is especially relevant to topic modeling, which created intricate interpretability patterns that were barely correlated with the data. Since this phenomenon was less prevalent with clustering, the question is *when* interpretable, nonrepresentative results are contextually justifiable. This also applies to computational time requirements. Methodological limits imposed by computational requirements or prioritizing the quick generation of results for the initial exploration of data require contextual justification.

Regarding preprocessing choices, we found that while a choice in any single preprocessing step may not cause major differences in output interpretability and representativeness, varying preprocessing choice combinations yield notably different outputs. One concrete example of preprocessing effects was that TF-IDF vectorization emphasizes rarer token discoveries with n-grams and chunks, a finding supported by previous research (Denny & Spirling, 2017). This emphasis on rarer tokens in topics and clusters can result in significantly poorer result representativeness and lead to ungeneralizable inferences and hypotheses based on only the rarest instances in the data. Contextual justifications for why analyzing only the rarest tokens would be both desirable and valid may exist, but these must be explicated for every research setting for transparency.

Our results further highlight how the combined effects of various preprocessing and algorithm choices can create issues in affirming the outputs', representativeness in relation to the data. If the UML data analysis regime output analysis is not elucidated beyond the study of the topic or cluster representations, issues regarding analysis reproducibility, transparency, and accountability emerge. At worst, this potentially allows the presentation of biased results (Covin & McMullen, 2019; Kirkman & Chen, 2011). To avoid transparency issues in UML research, preprocessing, and algorithm choices require rigid contextual justification due to their major impact and qualitative nature. As a framework for contextualizing UML data analysis regime choices, we offer the contextual justification principles in Table 7 to follow when conducting and reading UML-based research. Table 7 also offers illustrative answers to the questions posed in a research-setting scenario, in line with our previous example, in which making preprocessing choices to emphasize the rarest tokens would be justified.

The preprocessing and algorithm combination output analyses must be compliant with the research setting. The analysis's contextual justifications, preprocessing, and algorithm choices require disclosure to ensure comprehensive compatibility of the entire UML data analysis regime. We also include analysis considerations in our principles for reporting in Table 7. We argue that contextual justifications for analysis choices include descriptions of how the outputs were interpreted and whether the outputs and inferences based on the outputs were assessed for representativeness. Justifications for the compatibility and suitability of the combination of preprocessing, algorithm choice, and analysis choices also require elucidation. For example, topic modeling preprocessing

Table 7. Principles of Contextual Justifications in Reporting the Selection of Unsupervised Machine Learning Data Analysis Regimes.

| Phase | Question |
|------------------|--|
| Preprocessing | What preprocessing was done on the data before passing it on into algorithms? (e.g., “Common English stop words were removed and the documents were TF-IDF-vectorized.”) |
| | What preprocessing was not done on the data before passing it on into algorithms? (e.g., “Tokens were not stemmed or lemmatized.”) |
| | For each data preprocessing procedure, why was it justified over other options in light of the research goal? (e.g., “We are studying the evolution of jargon and changes in terminology, and since stop words remain common and consistent throughout, they are not considered meaningful to our purposes and were removed.”) |
| | Are the combined effects of the preprocessing choices suitable for the task and context at hand? (e.g., “Since we hope to find instances of terms used in previously unconventional ways, we wish to find all conjugations and forms of the terms and not lemmatize them, as well as emphasize the rarest forms with TF-IDF.”) |
| Algorithm choice | Were there limitations as to what preprocessing could not be considered? (e.g., “We wished to replicate a certain methodology with certain preprocessing, but some features were no longer supported by software.”) |
| | What unsupervised machine learning (UML) algorithm or algorithms were used? (e.g., “K-means clustering was used.”) |
| | What other possible UML algorithms were considered or trialed, and why were they not chosen? (e.g., “latent semantic indexing and latent Dirichlet allocation were trialed, but the task required representativeness that was not achieved with these algorithms.”) |
| | Why does the chosen UML algorithm suit the contextual situation? (e.g., “The short documents in the data cannot realistically cover multiple topics that would require topic modeling’s probabilistic qualities to catch.”) |
| Analysis | Why are the preprocessing choices in combination with the selected UML algorithms justified in the research setting? (e.g., “Clustering will group documents together that used similar uncommon terminology, and potentially allows for the identification of a group of documents that began a new branch of jargon.”) |
| | Were there limitations as to what UML algorithms could not be considered? (e.g., “Computational limits existed for the size of dataset that ruled out certain algorithms.”) |
| | How were the UML outputs analyzed to draw conclusions, i.e., how was output interpretability assessed? (e.g., “The documents in each cluster were analyzed for use of terms, time, and author as to whether the use of the same terminology was consequential or related to the same discussion.”) |
| | Were the outputs evaluated against the data that generated it, i.e., was representativeness assessed? Why is this contextually justifiable? (e.g., “The task requires validation of outputs against the data itself if substantive conclusions are to be drawn about the origins of terms, the appearance of the outputs is insufficient.”) |
| | Why is the chosen UML algorithm or algorithms in combination with the result analysis method justified in the research setting? |

(continued)

Table 7. (continued)

| Phase | Question |
|-------|---|
| | <p>(e.g., “Clustering groups documents with similar terminology together, which allows for straightforward comparative analysis of the actual documents within the cluster.”)</p> <p>Were there limitations as to what types on output analysis could not be considered?</p> <p>(e.g., “Only the clusters with terms of interest in the top tokens were assessed and other clusters were not scoured for whether they may have included documents with these terms.”)</p> |

choices can be made to improve output representativeness. However, if the results are then analyzed without investigating the outputs in relation to the data used to generate them, the justifications for the compatibility for the entire UML data analysis are inadequate. In the most lamentable case, one could assume sufficient result representativeness by only following suggestions from previous literature while failing to investigate the factual achieved outputs in relation to the data. To summarize, Table 7 provides principles for UML data analysis contextual justifications that can guide both those looking to employ rigorous UML methodology and those evaluating UML research.

Conclusions

Our results demonstrate trade-offs between UML outputs due to preprocessing and algorithm choices. Probabilistic topic modeling methods discovered intricate and interpretable patterns in outputs that were unsubstantiated by the factual data used. Contemporary UML reporting practices typically do not consider the alignment of the outputs with the data (i.e., representativeness). This absence, combined with oft-omitted preprocessing choices and algorithm considerations in UML research, creates research reproducibility, accountability, and transparency issues since others cannot validate, reproduce, or evaluate the methodology. These issues are especially pertinent in UML, since analysis inferences are always up for subjective interpretation.

We also found that contrary to topic modeling, clustering’s deterministic methodology creates outputs that are more aligned with the data but straightforward and less intricate. This may limit the possible use of clustering in UML contexts. In light of these results, providing solid logic to accompany the choices made in UML relating to the context and research setting becomes vital. Simply reporting preprocessing and algorithm choices without providing rigorous contextual justification for their suitability is insufficient. For example, researchers must explain why a probabilistic methodology suits their specific research setting better than a deterministic one.

To aid in disclosing and evaluating such justifications, we provided the principles in Table 7. Requiring such justifications limits potential misconduct (e.g., cherry-picking only the UML data analysis regimes that yielded the desired results) and mistakes. This increases research reproducibility, transparency, and accountability by preventing information omissions regarding the work put into titivating the final research results and their inferences.

No research is without limitations. It is possible that our preprocessing regime outputs varied drastically due to the particular programming libraries used and other specific choices that we made. The interpretability and representativeness of the regimes may be lower than ideal because our parameters were manipulated as little as possible, and better results in terms of interpretability and representativeness are certainly achievable using our data. This emphasizes our call for transparency regarding preprocessing methodologies, since achieving better results is likely to require greater preprocessing complexity, thereby creating even more accountability and reproducibility issues when these choices go unreported.

Furthermore, we omitted stop word removal and lemmatization contemplations from our analysis because they perform variably and require contextual consideration (Song et al., 2005; Toman et al., 2006). However, these were SML methodologies and thus were not wholly comparable, but we chose to prioritize more complex preprocessing choices. The decision to study only lemmatized documents was difficult to make, and we strongly suggest comparing lemmatization to no lemmatization in future research. However, since we aimed to study the differences between deterministic and probabilistic text clustering, and to protect the plausibility of the number of regimes studied, we decided not to compare lemmatization versus stemming versus neither. We prioritized studying n-grams and chunks over lemmatization because they address the issue of semantic information loss in BOW models (Fu et al., 2018; Sinoara et al., 2019; Zhao & Mao, 2018).

Moreover, finding generalizable results is becoming increasingly important (Church & Hestness, 2019), and that all the UML data analysis regimes in this study were run on the same dataset was a limitation. However, we consider that the demonstrated preprocessing and UML choice effects are clear, even with one dataset, particularly as random sampling was used and the data were not identical for all runs. The most obvious limitation of our research was the number of approaches tested, but it was impossible to study and compare all topic modeling and clustering methods exhaustively because there were too many. Nonetheless, the scope of the studied algorithms was clearly set, and inferences were not extended beyond the open-source versions studied. The surprising LDA behavior with an increasing number of topics possibly stemmed from the specific Gensim version becoming corrupted in this exact setting, but for our purposes, the number of topics for which no issues were raised sufficed to support our argument.

For a more comprehensive analysis, different algorithms, vectorizations, lemmatizations, and tokenizations should be studied from a larger variety of sources. In particular, algorithms that are better suited to high-dimensional and large datasets (such as CURE, DBCLASD, DBSCAN, STING, OPTICS; Xu & Tian, 2015), K-means variations that are more compatible with high dimensionality data or deep learning clustering (Ezugwu et al., 2022), and LDA versions without repetition issues in the topics created, should be studied. Interpretability and representativeness evaluations should be performed by more evaluators to increase the reliability of the results. Finally, using various datasets and further varied UML data analysis regimes in the future would also improve the reliability and generalizability of the conclusions. Our study lays the groundwork for further development of reporting practices in UML research to produce more reproducible, accountable, and transparent results in the field of organizational research.

Acknowledgements

We sincerely appreciate all valuable comments and suggestions from the reviewers and associate editor Steve Gove, which helped us to improve the quality of the manuscript. We also want to thank the research assistance of Santeri Heiskanen.


Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Tampere Chamber of Commerce and Academy of Finland, (grant number 279087).

ORCID iD

L. Valtonen  <https://orcid.org/0000-0003-4387-1239>

Supplemental Material

Supplemental material for this article is available online.

References

- Abney, S. (2007). *Semisupervised learning for computational linguistics*. CRC Press.
- Aggarwal, C. C., & Zhai, C. (2012). *Mining text data*. Springer Science + Business Media.
- Agarwal, A., Gans, J., & Goldfarb, A. (2020). How to win with machine learning. *Harvard Business Review*, 98(5), 126-133.
- Ashton, T., Evangelopoulos, N., Paswan, A., Prybutok, V. R., & Pavur, R. (2020). Assessing text mining algorithm outcomes. *Journal of Business Analytics*, 3(2), 107-121. <https://doi.org/10.1080/2573234X.2020.1785342>
- Banks, G. C., Woznyj, H. M., Wesslen, R. S., & Ross, R. L. (2018). A review of best practice recommendations for text analysis in R (and a user-friendly app). *Journal of Business and Psychology*, 33(4), 445-459. <https://doi.org/10.1007/s10869-017-9528-3>
- Bellstam, G., Bhagat, S., & Cookson, J. A. (2021). A text-based analysis of corporate innovation. *Management Science*, 67(7), 4004-4031. <https://doi.org/10.1287/mnsc.2020.3682>
- Blei, D. M., Carin, L., & Dunson, D. (2010). Probabilistic topic models. *IEEE Signal Processing Magazine*, 27(6), 55-65. <https://doi.org/10.1109/MSP.2010.938079>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(January), 993-1022.
- Braun, M. T., Kuljanin, G., & DeShon, R. P. (2018). Special considerations for the acquisition and wrangling of big data. *Organizational Research Methods*, 21(3), 633-659. <https://doi.org/10.1177/1094428117690235>
- Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, 9(2), 48-57. <https://doi.org/10.1109/MCI.2014.2307227>
- Cao, G., & Duan, Y. (2017). How do top- and bottom-performing companies differ in using business analytics? *Journal of Enterprise Information Management*, 30(6), 874-892. <https://doi.org/10.1108/JEIM-04-2016-0080>
- Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems 22 (NIPS 2009)* (pp. 288-296). Curran Associates Inc.
- Choudhury, P., Allen, R. T., & Endres, M. G. (2021). Machine learning for pattern discovery in management research. *Strategic Management Journal*, 42(1), 30-57. <https://doi.org/10.1002/smj.3215>
- Choudhury, P., Starr, E., & Agarwal, R. (2020). Machine learning and human capital complementarities: Experimental evidence on bias mitigation. *Strategic Management Journal*, 41(8), 1381-1411. <https://doi.org/10.1002/smj.3152>
- Church, K. W., & Hestness, J. (2019). A survey of 25 years of evaluation. *Natural Language Engineering*, 25(6), 753-767. <https://doi.org/10.1017/S1351324919000275>
- Covin, J. G., & McMullen, J. S. (2019). Programmatic research and the case for designing and publishing from rich, multifaceted datasets: Issues and recommendations. *Journal of Business Research*, 101, 40-46. <https://doi.org/10.1016/j.jbusres.2019.04.012>
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391-407. [https://doi.org/10.1002/\(SICI\)1097-4571\(199009\)41:6<391::AID-ASII>3.0.CO;2-9](https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASII>3.0.CO;2-9)

- Denny, M., & Spirling, A. (2017, September 27). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. SSRN. <https://doi.org/10.2139/ssrn.2849145>
- Erzurumlu, S. S., & Pachamanova, D. (2020). Topic modeling and technology forecasting for assessing the commercial viability of healthcare innovations. *Technological Forecasting and Social Change*, *156*, Article 120041. <https://doi.org/10.1016/j.techfore.2020.120041>
- Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, *110*, Article 104743. <https://doi.org/10.1016/j.engappai.2022.104743>
- Fokkens, A., van Erp, M., Postma, M., Pedersen, T., Vossen, P., & Freire, N. (2013). Offspring from reproduction problems: What replication failure teaches us. In H. Schuetze, P. Fung, & M. Poesio (Eds.), *Proceedings of the 51st annual meeting of the association for computational linguistics* (pp. 1691-1701). Association for Computational Linguistics.
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. Springer.
- Fu, M., Qu, H., Huang, L., & Lu, L. (2018). Bag of meta-words: A novel method to represent document for the sentiment classification. *Expert Systems with Applications*, *113*, 33-43. <https://doi.org/10.1016/j.eswa.2018.06.052>
- Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144. <https://doi.org/10.1016/j.ijinfomgt.2014.10.007>
- Haibe-Kains, B., Adam, G. A., Hosny, A., Khodakarami, F., Waldron, L., Wang, B., McIntosh, C., Goldenberg, A., Kundaje, A., Greene, C. S., & Broderick, T. (2020). Transparency and reproducibility in artificial intelligence. *Nature*, *586*(7829), E14-E16. <https://doi.org/10.1038/s41586-020-2766-y>
- Hannigan, T. R., Haans, R. F., Vakili, K., Tchalian, H., Glaser, V. L., Wang, M. S., Kaplan, S., & Jennings, P. D. (2019). Topic modeling in management research: Rendering new theory from textual data. *Academy of Management Annals*, *13*(2), 586-632. <https://doi.org/10.5465/annals.2017.0099>
- Hardeniya, N., Perkins, J., Chopra, D., Joshi, N., & Mathur, I. (2016). *Natural language processing: Python and NLTK*. Packt.
- Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2022). Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, *25*(1), 114-146. <https://doi.org/10.1177/1094428120971683>
- Huang, A. H., Lehavy, R., Zang, A. Y., & Zheng, R. (2018). Analyst information discovery and interpretation roles: A topic modeling approach. *Management Science*, *64*(6), 2833-2855. <https://doi.org/10.1287/mnsc.2017.2751>
- Jain, A. (2017). Weapons of math destruction: How big data increases inequality and threatens democracy. *Business Economics*, *52*(2), 123-125. <https://doi.org/10.1057/s11369-017-0027-3>
- Jeong, Y., Park, I., & Yoon, B. (2019). Identifying emerging research and business development (R&BD) areas based on topic modeling and visualization with intellectual property right data. *Technological Forecasting and Social Change*, *146*, 655-672. <https://doi.org/10.1016/j.techfore.2018.05.010>
- Kim, J. H., & Chen, W. (2018). Research topic analysis in engineering management using a latent Dirichlet allocation model. *Journal of Industrial Integration and Management*, *3*(4), Article 1850016. <https://doi.org/10.1142/S2424862218500161>
- Kirkman, B. L., & Chen, G. (2011). Maximizing your data or data slicing? Recommendations for managing multiple submissions from the same dataset. *Management and Organization Review*, *7*(3), 433-446. <https://doi.org/10.1111/j.1740-8784.2011.00228.x>
- Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018a). Text mining in organizational research. *Organizational Research Methods*, *21*(3), 733-765. <https://doi.org/10.1177/1094428117722619>
- Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018b). Text classification for organizational researchers: A tutorial. *Organizational Research Methods*, *21*(3), 766-799. <https://doi.org/10.1177/1094428117719322>

- Kuang, L., Yang, L. T., Chen, J., Hao, F., & Luo, C. (2015). A holistic approach for distributed dimensionality reduction of big data. *IEEE Transactions on Cloud Computing*, 6(2), 506-518. <https://doi.org/10.1109/TCC.2015.2449855>
- Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans. In M. G. Shafto, & P. Lang (Eds.), *Proceedings of the 19th annual conference of the cognitive science society* (pp. 412-417). Lawrence Erlbaum Associates, Inc.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444. <https://doi.org/10.1038/nature14539>
- Lee, H., & Kang, P. (2018). Identifying core topics in technology and innovation management studies: A topic model approach. *The Journal of Technology Transfer*, 43(5), 1291-1317. <https://doi.org/10.1007/s10961-017-9561-4>
- Lee, I., & Shin, Y. J. (2020). Machine learning for enterprises: Applications, algorithm selection, and challenges. *Business Horizons*, 63(2), 157-170. <https://doi.org/10.1016/j.bushor.2019.10.005>
- Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, 28(4), 587-604. <https://doi.org/10.1111/j.1468-2958.2002.tb00826.x>
- Manning, C. D., Schütze, H., & Raghavan, P. (2008). *Introduction to information retrieval*. Cambridge University Press.
- McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: The management revolution. *Harvard Business Review*, 90(10), 60-68.
- Mohr, J. W., & Bogdanov, P. (2013). Introduction—topic models: What they are and why they matter. *Poetics*, 41(6), 545-569. <https://doi.org/10.1016/j.poetic.2013.10.001>
- Muslea, I., Minton, S., & Knoblock, C. A. (2006). Active learning with multiple views. *Journal of Artificial Intelligence Research*, 27, 203-233. <https://doi.org/10.1613/jair.2005>
- Robinson, T. J., Giles, R. C., & Rajapakshage, R. U. (2020). Discussion of “experiences with big data: Accounts from a data scientist’s perspective.” *Quality Engineering*, 32(4), 543-549. <https://doi.org/10.1080/08982112.2020.1758333>
- Rosso, C. (2018). The human bias in the AI machine. *Psychology Today*. Retrieved June 13, 2002, from <https://www.psychologytoday.com/us/blog/the-future-brain/201802/the-human-bias-in-the-ai-machine>
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513-523. [https://doi.org/10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0)
- Schmiedel, T., Müller, O., & vom Brocke, J. (2019). Topic modeling as a strategy of inquiry in organizational research: A tutorial with an application example on organizational culture. *Organizational Research Methods*, 22(4), 941-968. <https://doi.org/10.1177/1094428118773858>
- Sinoara, R. A., Camacho-Collados, J., Rossi, R. G., Navigli, R., & Rezende, S. O. (2019). Knowledge-enhanced document embeddings for text classification. *Knowledge-Based Systems*, 163, 955-971. <https://doi.org/10.1016/j.knosys.2018.10.026>
- Song, F., Liu, S., & Yang, J. (2005). A comparative study on text representation schemes in text categorization. *Pattern Analysis and Applications*, 8(1-2), 199-209. <https://doi.org/10.1007/s10044-005-0256-3>
- Srivastava, A. N., & Sahami, M. (2009). *Text mining: Classification, clustering, and applications*. CRC Press.
- Talafidaryani, M. (2021). A text mining-based review of the literature on dynamic capabilities perspective in information systems research. *Management Research Review*, 44(2), 236-267. <https://doi.org/10.1108/MRR-03-2020-0139>
- Toman, M., Tesar, R., & Jezek, K. (2006). Influence of word normalization on text classification. *Proceedings of InSciT*, 4, 354-358.
- Tonidandel, S., King, E. B., & 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>

- Westerlund, M., Leminen, S., & Rajahonka, M. (2018). A topic modelling analysis of living labs research. *Technology Innovation Management Review*, 8(7), 40-51. <https://doi.org/10.22215/timreview/1170>
- White, G. O., Guldiken, O., Hemphill, T. A., He, W., & Khoobdeh, M. S. (2016). Trends in international strategic management research from 2000 to 2013: Text mining and bibliometric analyses. *Management International Review*, 56(1), 35-65. <https://doi.org/10.1007/s11575-015-0260-9>
- Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2), 165-193. <https://doi.org/10.1007/s40745-015-0040-1>
- Zhang, Y., Jin, R., & Zhou, Z. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4), 43-52. <https://doi.org/10.1007/s13042-010-0001-0>
- Zhang, Y., & Shaw, J. D. (2012). Publishing in AMJ—part 5: Crafting the methods and result. *Academy of Management Journal*, 55(1), 8-12. <https://doi.org/10.5465/amj.2012.4001>
- Zhao, R., & Mao, K. (2018). Fuzzy bag-of-words model for document representation. *IEEE Transactions on Fuzzy Systems*, 26(2), 794-804. <https://doi.org/10.1109/TFUZZ.2017.2690222>
- Zhong, N., & Schweidel, D. A. (2020). Capturing changes in social media content: A multiple latent changepoint topic model. *Marketing Science*, 39(4), 827-846. <https://doi.org/10.1287/mksc.2019.1212>
- Ziegler, C. (2012). *Mining for strategic competitive intelligence: Foundations and applications*. Springer Science + Business Media.

Author Biographies

L. Valtonen is a PhD candidate at Tampere University. Their research interests concern the social embeddedness of technology, especially the impacts of beliefs and bias regarding technology on both decision-making in organizational settings and the social and environmental sustainability of technology.

Saku J. Mäkinen is a professor of industrial engineering and management at University of Turku, Finland. His broad research interests consider value creation from various perspectives in organizational settings.

Johanna Kirjavainen is a post-doctoral researcher at Tampere University. Her research interests focus on product strategies and strategic foresight.

PUBLICATION

II

Supervised machine learning in detecting patterns in competitive actions

Valtonen, L., Mäkinen, S. J., and Kirjavainen, J.,

In *Proceedings of the 2021 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 442–446)

Publication reprinted with the permission of the copyright holders.

Supervised Machine Learning in Detecting Patterns in Competitive Actions

L. Valtonen¹, S.J. Mäkinen¹, J. Kirjavainen¹

¹Industrial Engineering and Management, Tampere University, Tampere, Finland
(laura.valtonen@tuni.fi)

Abstract - This paper explores possibilities to investigate how patterns of competitive actions could be detected with supervised machine learning (SML) methods. Competitive dynamics and the resource-based view are used as theoretical frameworks for classifying competitive actions. These in turn represent the dynamics of industry evolution from competitive actions point of view. We find promising ways to furthering our understanding of detectable patterns in competitive dynamics and industry evolution. Our results show that standard SML methods can be used in pattern recognition but reporting the methods used in detail are of paramount importance in facilitating peer-review and scientific replication and producing credible results.

Keywords - Competitive actions, supervised machine learning, competitive dynamics

I. INTRODUCTION

In this paper we explore how an industry evolves and specifically we investigate the actors and their activities in an industry. Industry functions as a system of actors in which in order to survive and thrive in an industry, these firms as actors seek to gain and maintain competitive advantage [1]. Following the competitive dynamics view, firms use competitive actions as tools for building competitive advantage [2, 3] in order to gain rents above industry average. The resource-based view (RBV), on the other hand, views specific types of resources and capabilities as building blocks of sustained competitive advantage [4] similarly used to gain above industry average rents. Hence, the resources and capabilities of the firm can be seen to enable and be the result of firm actions [5, 6], and we may decipher RBV well compatible with competitive actions and dynamics views.

Following the above logic, we are looking at signals companies give of their actions to markets as news bits following traditional competitive dynamics research traditions (e.g. [7]). We gather the news items from global news source, namely news from Lexis-Nexis. Studying this type of complex systems that create a large amount of data - news bits in this case - is more approachable with methods that can deal with data vastness, namely machine learning in addition or in place of content analysis conducted by human coders. In this research we explore the possibility to map with supervised machine learning (SML) methods competitive actions regarding different resource categories [8]. We also pay attention to how to deal with known issues related to supervised machine learning regarding its black-box nature, since in making decisions, regarding

competitive actions for instance, the interpretability of the data may be prioritized [9].

Traditionally, structured content analysis has been used to distinguish action patterns [10]. However, business intelligence and scientific research investigating industry evolution, competitive dynamics and events in the marketplace need to utilize machine learning, at least in some form, as the amount of data has surpassed the capabilities of human coding and overall human labor becomes too costly and time consuming as the amount of data increases.

In sum, our goal in this paper is to investigate whether patterns in competitive actions can be detected with human coding and supervised machine learning (SML). From the news we code events that are classified with resource-based view categories as these represent the competitive actions companies are signaling with content analysis first by humans and then supervised machine learning and finally, we compare these results. Then we map these outcomes and seek patterns and whether both or either one is able to detect patterns from the data. We find that agreement rate between SML and human coding is heavily dependent on the algorithm and other choices made in SML, so care needs to be taken with reporting these methodological issues as well as the results. Finally, we find new, interesting opportunities for classifying, with new topical outlays based on SML that are not built-in to our resource-based view categorizations.

Hence, we conclude that SML holds promises for furthering our understanding of industry evolution and competitive dynamics, our results also pave way for combining the traditions of resource-based view and competitive dynamics research streams as called for in recent research [6].

II. METHODOLOGY

Our dataset was collected from the LexisNexis with the keyword “Statkraft” over the years 1987 – 2020. Statkraft is energy company owned by Norwegian government and it is third largest energy company in the Nordics. Research assistants refined the resulting dataset to include news texts in English which included Statkraft as an active actor. This left 2198 news texts for further analysis. The research assistants recognized events and coded these actions into one of the RBV categories [8]. The coding scheme was tested and interpretations made coherent via discussion with the authors while going over

a data sample. The assistants extracted from each news text the time of publication, time of performed action if given, the title of the news piece, and the sentence depicting the competitive action. These descriptive sentences and their assigned RBV-categories were then used as input in following text mining and analysis. After coding, the documents were filtered to even out the amount of data in each category: RBV-categories with over 250 documents assigned to them were truncated to contain only 250 documents. The resulting dataset contained 1387 items.

The extracted descriptive sentences were stripped of extra whitespace and processed with Python’s Spacy [11] library: Punctuation, numerals, URLs, and email addresses were removed. The effects of stopword removal were studied: The SML algorithms were ran on two datasets, one cleansed from stopwords and the other not. The datasets were then vectorized with both bag-of-words (BOW) [12,13] and term frequency inverse document frequency (TF-IDF) [14-16] tokenizers from Python’s scikit-learn library [17]. These transformed descriptive sentences along with their designated RBV-categories were studied with scikit-learn classification algorithms. To be studied, algorithms had to be ready-to-use on scikit-learn, well known, and suitable for classification into multiple categories with a small text dataset. Solver parameters were iterated to test what is most suitable for our dataset size. Depth and layer parameters were given small values to keep initial tests computationally light, but some iteration was performed to see how the classifier behaves regarding changes in the parameter. The set of studied algorithms is presented in Table 1. The classifiers were trained on 70% and 85% of the datasets and tested on 30% and 15%. From

scikit-learns metrics the 30% and 15% test sets were assessed with accuracy, confusion matrices and classification results metrics [18]. Out of these splits, the 85% for training and 15% for testing was chosen for further analysis since no major differences emerged between the two. Based on the accuracies the three best performing algorithms with were selected for further scrutiny: Multi-layer Perceptron (MLP) classification [19], Logistic Regression [20], and Random Forest [21]. The accuracies of these classifiers with the different pre-processing combinations (stopwords removed or not, TF-IDF or BOW vectorization) and normalized confusion matrices [18] were extracted. The most representative tokens of each RBV-category were extracted for each pre-processing combination.

Finally, after the above procedures we constructed event maps of competitive actions. In these maps we compared in the timeline the actions that human coders identified with the actions the best performing SML classifier identified correctly. With this exploratory analysis, we derive conclusions on how well SML can detect patterns in the stream of actions.

III. RESULTS

Table 1 displays the accuracies of each studied SML classifier. Out of the classifiers that consistently reach accuracies over 70%, only logistic regression and MLP performance do not appear to vary dramatically depending on pre-processing – unlike SVC and Random Forest, which display one pre-processing result to be distinctly better or worse than the others.

Table 1. Accuracies of explored SML classifiers with different pre-processing approaches

| Algorithm | With stopwords | | Without stopwords | |
|---|----------------|--------|-------------------|--------|
| | BOW | TF-IDF | BOW | TF-IDF |
| DecisionTreeClassifier(max_depth=5) | 0.44 | 0.39 | 0.44 | 0.39 |
| DecisionTreeClassifier(max_depth=10) | 0.6 | 0.54 | 0.54 | 0.56 |
| DecisionTreeClassifier(max_depth=15) | 0.62 | 0.57 | 0.53 | 0.53 |
| LogisticRegression() | 0.74 | 0.74 | 0.73 | 0.73 |
| LogisticRegression(solver='liblinear') | 0.74 | 0.75 | 0.73 | 0.74 |
| LogisticRegression(solver='sag') | 0.74 | 0.74 | 0.73 | 0.73 |
| LogisticRegression(solver='saga') | 0.76 | 0.74 | 0.73 | 0.73 |
| LinearDiscriminantAnalysis() | 0.54 | 0.53 | 0.57 | 0.57 |
| KNeighborsClassifier() | 0.6 | 0.71 | 0.49 | 0.69 |
| KNeighborsClassifier(n_neighbors=10) | 0.55 | 0.72 | 0.5 | 0.7 |
| KNeighborsClassifier(n_neighbors=15) | 0.49 | 0.71 | 0.48 | 0.7 |
| GaussianNB() | 0.62 | 0.61 | 0.61 | 0.61 |
| SVC(decision_function_shape='ovo') | 0.7 | 0.74 | 0.69 | 0.72 |
| MLPClassifier(alpha=1e-05, hidden_layer_sizes=(50, 10)) | 0.72 | 0.71 | 0.71 | 0.72 |
| MLPClassifier(alpha=1e-05, hidden_layer_sizes=(50, 10), solver='lbfgs') | 0.66 | 0.66 | 0.67 | 0.71 |
| MLPClassifier(alpha=1e-05) | 0.72 | 0.73 | 0.71 | 0.72 |
| MLPClassifier(alpha=1e-05, solver='lbfgs') | 0.69 | 0.71 | 0.71 | 0.7 |
| MLPClassifier(alpha=1e-05, hidden_layer_sizes=(500, 100)) | 0.72 | 0.71 | 0.73 | 0.74 |
| MLPClassifier(alpha=1e-05, hidden_layer_sizes=(500, 100), solver='lbfgs') | 0.71 | 0.72 | 0.71 | 0.7 |
| RandomForestClassifier(max_depth=5) | 0.67 | 0.66 | 0.71 | 0.64 |
| RandomForestClassifier(max_depth=10) | 0.73 | 0.71 | 0.72 | 0.71 |
| RandomForestClassifier(max_depth=15) | 0.76 | 0.71 | 0.77 | 0.76 |
| AdaBoostClassifier() | 0.44 | 0.43 | 0.44 | 0.44 |

Table 2. Accuracies of the previously best performing classifiers ran a second time on the same data

| ACCURACIES | Stopwords removed | | Stopwords NOT removed | |
|---|-------------------|------|-----------------------|------|
| | TF-IDF | BOW | TF-IDF | BOW |
| LogisticRegression(solver='lbfgs') | 0,74 | 0,75 | 0,72 | 0,76 |
| MLPClassifier(alpha=1e-05, random_state=42, hidden_layer_sizes=(25,)) | 0,74 | 0,76 | 0,75 | 0,77 |
| RandomForestClassifier(max_depth=15) | 0,65 | 0,67 | 0,68 | 0,69 |
| RandomForestClassifier(max_depth=20) | 0,69 | 0,71 | 0,73 | 0,72 |

Out of the best performing classifiers in Table 1, four were selected for further exploration: Table 2 demonstrates the behavior in which the results can vary drastically in comparison to another run (Table 1) and due to the black-box nature of most SML classifiers, it cannot be always determined what causes these differences. Here the likely cause is that in the splitting of the data into the training and testing sets, the division and distribution of data per RBV-category is better suited per classifier in differing manners.

In Table 2, the classifiers mainly achieve slightly better accuracies when stopwords are not removed. However, the top tokens per RBV-category become confusing with stopwords. As the accuracies are in the same range, we decided to study the results with stopwords removed.

Figure 1 shows the top tokens and the words in them. These are qualitatively surprisingly descriptive for their respective RBV-categories. The confusion matrices produced by each of the methods reflected the accuracies in the sense that MLP had a slight advantage over the other two classifiers: MLP confusion matrices, especially with BOW vectorization, classified all RBV-categories well whereas with the other classifiers, some categories clearly bled into other categories in the matrices, i.e. some RBV-categories were more difficult to classify for Linear Regression and Random Forest than others.

Hence, we chose MLP results to create action maps from for comparison against the original human coded set.

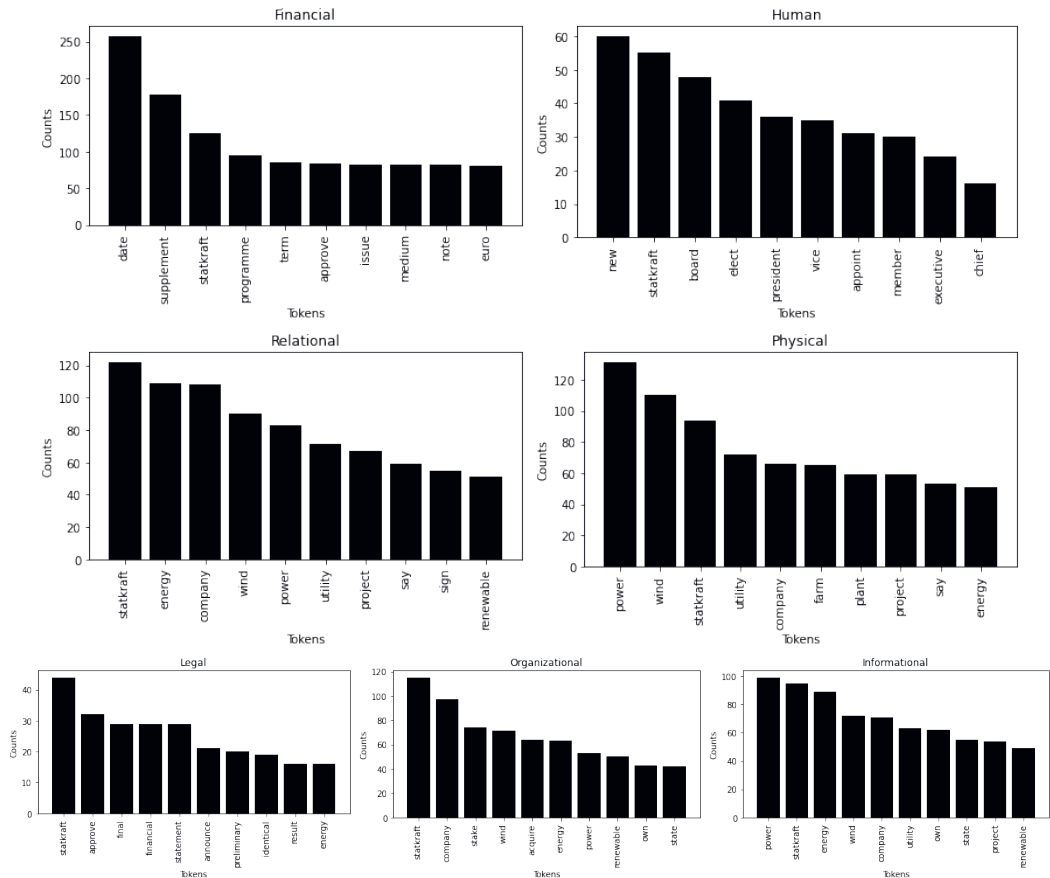


Fig. 1. Top representative tokens per RBV-category

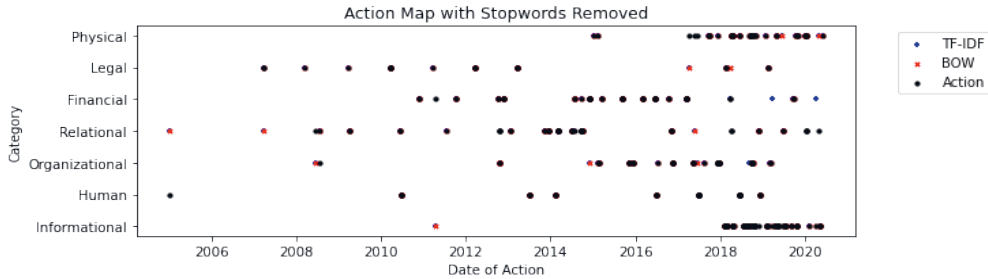


Fig. 2. MLP BOW and TF-IDF classifications against Human coded RBV-categories

As we can see from the Fig. 2, both BOW and TF-IDF classify actions in a very similar fashion as the human coders. The inter-rater agreement rate between human coding and BOW is 27 false out of 319, namely 91.5% and for TF-IDF is 29 false out of 319, namely 91%. This result as such is rather surprisingly high and lends credence to a possibility that pattern recognition can be done with SML tools.

We can also detect some distinct patterns in Fig. 2, e.g. Physical actions are abundant after considerable amount of Relational actions and Financial actions. This is due to the nature of the industry and clearly can be used for detecting industry dynamics and competitive action patterns. Surprisingly, Informational actions are also abundant only in later years in our data.

In Fig. 3 and 4 we can see in detail the false categorizations for both vectorizers. Both vectorizers classify more events to Physical, Relational, and Organizational categories than human coders. Noticeable detail is also that there are 8 occurrences where human coders, BOW and TF-IDF differ from one another in their

classifications. For example, “Statkraft sells minority interest in UK onshore portfolio to reinvest in new renewable energy” has been classified by humans and BOW as financial and by TF-IDF as Relational. This type of confusion is understandable as even human coders would be hard pressed to decipher which aspect of the news is more pressing or the core, the relationship or the financial aspect of the action.

Furthermore, despite the 8 occurrences (about 30% of the false assessments) where vectorizers assessment differ from one another, there are differences between humans and vectorizers in that vectorizers agree but differently from a human coder. For example, “E.ON and Norwegian group Statkraft sign deal” has been categorized as Relational by human coders whereas both vectorizers treat this as Organizational arrangement, and both interpretations are valid with the event description. Furthermore, we can see that vectorizers identify additional pattern of activity for Physical actions that humans were not able to detect.

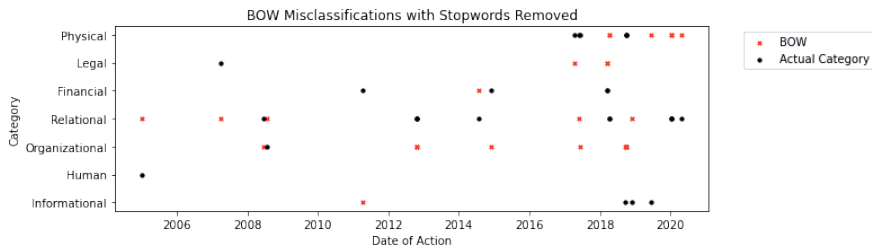


Fig. 3. MLP BOW classification miscategorizations against actual human coded RBV-category

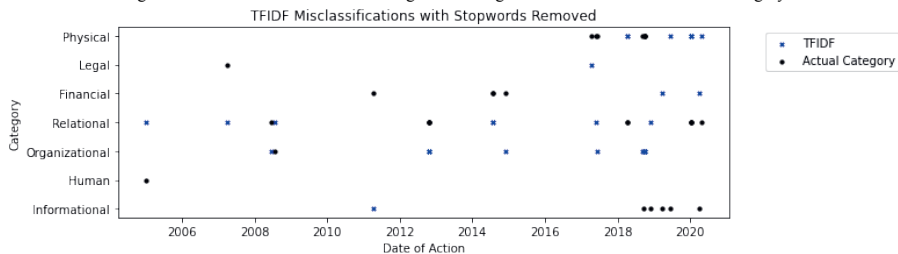


Fig. 4. MLP TF-IDF classification miscategorizations against actual human coded RBV-category

Most common mistake (7 occurrences) was classifying a partnership action as a physical action. Confusing organizational actions with relational and physical were second most common mistakes (5 occurrences both). Other cross-classifications were not as prevalent. When looking at the documents that were misclassified, it was noted that often the MLP classification was more sensible than the human coded one. For instance, the action “*Statkraft has acquired the Irish and UK wind development businesses of the Element Power Group*” was coded as physical, but classified as organizational, which is certainly more correct. Moreover, when mistakes were made in classification, the documents were ambiguous enough to understand exactly why the misclassification happened, for instance: the human coded informational text “*Statkraft AS has made the Offering Circular dated 26 March 2020 for the EUR 6,000,000,000 Euro Medium Term Note Programme available for viewing*” was classified as a financial document by TF-IDF vectorization and it sure enough concerns finance. The physical document “*Statkraft has closed a first power purchase agreement in France under the country’s new support mechanism for renewable energy, the Norwegian utility said March 27*” was classified as a legal document, which seems a fair assessment when the document concerns agreements with a national state.

IV. CONCLUSION

The interpretability of the mistakes highlights the possibilities of SML to assist in mapping competitive actions. Moreover, since even the mis-categorizations make a compelling case about why the classifiers should be right instead of the human, classifications results can be used in iteratively improving the original coding set as a coder re-evaluates possible conflicts for what category is truly more descriptive. However, using machine learning has notable issues and risks with bias and transparency. When retrieving patterns of competitive actions with SML, the results need both transparent reporting and an actual human being to validate and interpret the results – at least to some degree. This may impose some restrictions of what type of machine learning can be used.

ACKNOWLEDGMENT

We are grateful for the effort of our research assistants H. Pukkala, R. Harjula, E. Periviita, M. Kukkula, A. Korin, M. Rautio. This work has been partially supported by Procem++ and funded by Business Finland.

REFERENCES

- [1] M. E. Porter, *The Competitive Advantage: Creating and Sustaining Superior Performance*, Free Press, NY, 1985
- [2] Bettis, Richard A., and David Weeks. "Financial returns and strategic interaction: The case of instant photography." *Strategic Management Journal* 8(6), 1987, pp. 549-563.
- [3] MacMillan, I., McCaffery, M. L., and Van Wijk, G. "Competitors' responses to easily imitated new products—

Exploring commercial banking product introductions”, *Strategic Management Journal*, 6(1), 1985, pp. 75-86.

- [4] Barney, J. "Firm resources and sustained competitive advantage." *Journal of management* 17(1), 1991, pp. 99-120.
- [5] Kraaijenbrink, J., J.-C. Spender, and A. J. Groen, "The resource-based view: A review and assessment of its critiques", *Journal of management*, 36(1), 2010, pp. 349-372.
- [6] Chen, Ming-Jer, John G. Michel, and Wenchen Lin, "Worlds apart? Connecting competitive dynamics and the resource-based view of the firm", *Journal of Management*, 01492063211000422, 2021
- [7] Chen, M.- J., Miller, D., "Competitive dynamics: Themes, trends, and a prospective research platform", *Academy of Management Annals*, 6, 2012, pp. 135-210.
- [8] Morgan, R. M., Hunt, S. D., "Relationship-Based Competitive Advantage: The Role of Relationship, Marketing in Marketing Strategy", *Journal of Business Research*, 46(3), 1999, pp. 281-290.
- [9] Lee, I., and Yong Jae Shin, "Machine learning for enterprises: Applications, algorithm selection, and challenges", *Business Horizons*, 63(2), 2020, pp. 157-170.
- [10] Smith, K. G., Ferrier, W. J., & Ndofor, H. (2001). Competitive dynamics research: Critique and future directions. *Handbook of strategic management*, 315, 361.
- [11] Honnibal, M., Montani, L., Van Landeghem, S., and Boyd, A., spaCy: Industrial-strength Natural Language Processing in Python, Zenodo, 2020, <https://doi.org/10.5281/zenodo.1212303>
- [12] scikit-learn documentation. (n.d.-a). https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html#sklearn.feature_extraction.text.CountVectorizer
- [13] Zhang, Y., Jin, R., & Zhou, Z.-H., "Understanding bag-of-words model: a statistical framework", *International Journal of Machine Learning and Cybernetics*, 1(1-4), 2010, pp. 43-52.
- [14] Salton, G., & Buckley, C., "Term-weighting approaches in automatic text retrieval", *Information Processing & Management*, 24(5), 1988, pp. 513-523.
- [15] Schütze, H., Manning, C. D., and Raghavan, P., *Introduction to information retrieval* (Vol. 39), Cambridge University Press Cambridge, 2008
- [16] scikit-learn documentation. (n.d.-b). https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html?highlight=vectorizer#sklearn.feature_extraction.text.TfidfVectorizer%7D
- [17] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E., "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research*, 12, 2011, pp. 2825-2830.
- [18] scikit-learn documentation. (n.d.-c). <https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics>
- [19] scikit-learn documentation. (n.d.-d). https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html?highlight=mlpclassifier#sklearn.neural_network.MLPClassifier
- [20] scikit-learn documentation. (n.d.-e). https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html?highlight=logisticregression#sklearn.linear_model.LogisticRegression
- [21] scikit-learn documentation. (n.d.-f). <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

PUBLICATION

III

Human-in-the-loop: Explainable or accurate artificial intelligence by exploiting human bias?

Valtonen, L., and Mäkinen, S. J.,

In Proceedings of the 2022 IEEE 28th International Conference on Engineering, Technology and Innovation (ICE/ITMC) & 31st International Association For Management of Technology (IAMOT) Joint Conference (pp. 1–8)

Publication reprinted with the permission of the copyright holders.

Human-in-the-loop: Explainable or accurate artificial intelligence by exploiting human bias?

Laura Valtonen
Faculty of Management and Business
Tampere University
Tampere, Finland
laura.valtonen@tuni.fi

Saku J. Mäkinen
Faculty of Technology
University of Turku
Turku, Finland
saku.makinen@utu.fi

Abstract—Artificial intelligence (AI) is a major contributor in industry 4.0 and there exists a strong push for AI adoption across fields for both research and practice. However, AI has quite well elaborated risks for both business and general society. Hence, paying attention to avoiding hurried adoption of counter-productive practices is important. For both managerial and general social issues, the same solution is sometimes proposed: human-in-the-loop (HITL). However, HITL literature is contradictory: HITL is proposed to promote fairness, accountability, and transparency of AI, which are sometimes assumed to come at the cost of AI accuracy. Yet, HITL is also considered a way to improve accuracy. To make sense of the convoluted literature, we begin to explore qualitatively how explainability is constructed in a HITL process, and how method accuracy is affected as its function. To do this, we study qualitatively and quantitatively a multi-class classification task with multiple machine learning algorithms. We find that HITL can increase both accuracy and explainability, but not without deliberate effort to do so. The effort required to achieve both increased accuracy and explainability, requires an iterative HITL in which accuracy improvements are not continuous, but disrupted by unique and varying human biases shedding additional perspectives on the task at hand.

Keywords—*human-in-the-loop, industry 4.0, artificial intelligence, accuracy, explainability*

I. INTRODUCTION

The “fourth industrial revolution” approaches us with artificial intelligence (AI) as a major contributing innovation force [1,2], presenting possible unprecedented side-effects and externalities (e.g., [3]). Naturally, in the quest to utilize emergent technological developments, there are attempts to acclimate AI into various fields, especially in industry 4.0, because it has been noted that organizations that leverage available data in quantities typical to AI and machine learning (ML) are considered to triumph over those that do not [4-6]. However, AI has well established risks for both business and society at large. For society, such risks include issues regarding social responsibility, and for management and organizations, these issues include managerial overreliance on algorithms [7,8] and loss of unique human knowledge within organizations [9]. Interestingly, for both managerial and general social issues, the same solution is sometimes proposed: human-in-the-loop (HITL) [7,10-11].

HITL refers to AI approaches that keep human input in the algorithmic decision-making process at some level.

Sometimes, this is referred to as “augmentation” of decision making with AI, whereas a lack of HITL is defined as “automation” of decision making [12]. In practice, HITL often translates to iteratively revising the outputs of an AI and updating the corresponding inputs with corrections made by people. However, the literature on HITL is contradictory: HITL is proposed to promote fairness [11], accountability, and transparency [10], which again are assumed to come at the cost of AI accuracy [13-16]. Regardless, HITL is also considered a way to improve accuracy (see e.g., [17-19]). Moreover, HITL approaches are also promoted due to the opportunity of humans learning from the machines instead of only the other way around [20-23]. In all such cases, HITL is framed as a solution to a problem, when in practice, HITL or “augmentation” is not a choice, but simply how AI implementation plays out in reality [12,24].

In addition to the above, explainability has been lifted forward in order to increase trust, fairness, accountability, and transparency towards solutions given by explainable AI (XAI) [25, 26] but empirical work that evaluates and quantifies explainability from users’ points of view remains scarce [25, 26]. Most approaches remain on a conceptual level [27]. The scarcity of empirical works exploring explainability among interpretability, accountability and transparency persists for HITL, except for some highly specific and applied technical studies like that of [18] who use a single deep neural network classifier for feature selection. Moreover, only few of the already small number of works, 5% [25], that evaluate explainability empirically consider non-tabular data such as text, which is the most abundant form of organizational data [28, 29].

We begin to address this gap of empirical research regarding the effects of HITL on AI explainability and accuracy. Hence, our research question is “How does HITL influence the relationship between explainability and accuracy of AI classification?” In addition to qualitative assessments on how explainability is constructed in the HITL process, to provide perspective into the contradictions regarding HITL impact on accuracy, we quantitatively study the variance of accuracies of the algorithms.

Our classification uses the resource-based view and competitive dynamics as frameworks to construct categories of events to be found in the data. The data consists of news texts, which are a quintessential form of organizational data in the field of competitive intelligence. Our comparative setting covers multiple ML classifier algorithms on a multi-class text dataset and assesses the impact of using a HITL

approach to achieve explainability and accuracy. In order to have a meaningful data set, we use the well-known case of Kodak to study the effects on implementing HITL methodology into using AI in competitive intelligence tasks. Further, our intention is to shed light on the discrepancies between accuracy and explainability as a part of HITL implementations.

In agreement with previous literature, we find that a previously agreed upon classification comes under scrutiny as humans in the loop refine their understanding on what is the exact task at hand. This leads to a more comprehensive understanding of the classification task, including strengths and points of improvement. In addition, and contrary to expectations, our results point towards that the understanding is a result of the richness of human opinion, or human bias, due to which the coders disagree on some points even after diligent discussions over the coding scheme. Our results show that the discrepancy in thought is reflected in the accuracies of the algorithms as well: the human-in-the-loop one (HITL-1) improves the average accuracy of the studied algorithms, whereas the second iterative round of interference from a second human-in-the-loop two (HITL-2) degrades average accuracy while increasing explainability. Hence, we may state that instead of focusing solely on improving the accuracy, the harnessing of the human bias inherent in classification tasks is recommended to raise new insights and point of discussion for comprehensive AI explainability.

II. THEORETICAL BACKGROUND

A. Bias and Black Boxes

Contemporary AI typically refers to employing supervised machine learning (SML) algorithms [30,31]. In SML, the data inserted into the algorithm is “labeled”. The algorithm then finds the best way it can to figure out how to map a datapoint to a correct “label” based on the examples it was given for learning. This type of machine learning algorithm is called a classifier. Biased labels used in SML are a typical example of how bias creeps into AI implementations [32]. Some simple examples of how labels can become biased are a biased human labeler or poorly representative data. Biased data is especially dangerous, because AI is typically sold on and optimized for accuracy [33]. Accuracy here refers to the percentage of data a machine learning algorithm can classify into a correct category according to a pre-labeled dataset. The accuracy metric of a method demands that for a higher score, discriminatory past data should be classified so: Conforming to biased data is rewarded.

However, bias can exist in the used data, human judgement, or in the algorithms themselves [14], and bias is only one issue with AI. Another associated risk with AI is the black box nature of algorithms, or, in other words, algorithm opaqueness [15]. A black box AI is a “full opaqueness” AI implementation, in which there is no transparency into the decision process – data goes in, and a result comes out, and no one truly understands how that decision was made. This raises questions of accountability – who is responsible for the decisions no-one understands?

B. Overreliance

Bias and opacity become more complex issues when combined with overreliance on AI: Humans tend to become biased to agree with technological information. Some examples of this include strategic managerial choice [7],

purchasing recommender systems [34], retirement savings [35], and more (e.g., [9,36]). Overreliance caused due to loss of unique human knowledge [9], and consequent capability for exploration, combined with by humans becoming biased to agree with already biased AI no one can decipher, will surely emphasize and create wicked problems.

C. Fairness, Accountability and Transparency

Naturally, there exists concern over the lack of accountability and transparency of such methods [37-40]. To combat such issues, there is a stream of research called explainable AI (XAI) that contributes to achieving AI fairness, accountability, and transparency [8]. Fairness refers to a lack of discriminatory bias in AI [41]. Transparency is the degree to which AI can be understood, explained, and observed [41]. Accountability refers to the possibility to audit the results of AI and assign responsibility of the consequences to an actual entity [41].

D. Human-in-the-loop

Keeping humans augmenting and auditing the AI (HITL) is proposed to increase fairness, accountability, and transparency of AI [10,11], as well as to combat issues of managerial overreliance [7]. However, HITL refers to a variety of modes in which a human can audit an algorithm. If observing AI predictions incline people to agree with them over their own initial views [9], and the convergence of a variety of opinions can be detrimental to the performance of a group, as was the case with viewing public information and stock results [42], does a HITL becoming biased in the loop to agree with the algorithm truly meet the requirements of fairness, accountability, and transparency?

III. METHODOLOGY

To begin addressing the “drought” of empirical work that evaluates and quantifies said explainability beyond a conceptual level [25-27], we explore how explainability is constructed in a HITL process, and how algorithm accuracy is affected as its function. To do this, we study qualitatively and quantitatively a multi-class classification task with multiple SML algorithms, instead of a bi-class task, to coax out a larger variety of possible points of bias. Here, we use a typical HITL approach in which humans revised the AI classifications and made corrections as necessary and the corrected classifications were replaced in the original AI training data for further AI trainings. The methodology described as follows is depicted in Fig. 1.

A. Data

The dataset for the task was collected using the keyword “Kodak” from the LexisNexis over the years 2002 to 2018 and refined to include news texts in English. Texts that depicted an action by an organization were retained. This yielded altogether 2295 news texts. The events described in the texts were coded by research assistants into one of the following categories based on what type of resource was concerned: informational, human, organizational, relational, financial, legal, or physical [43]. The assistants for the task were carefully selected second- and third-year industrial engineering and management students and were given training for the task after which the coding scheme was tested with the assistants on a small sample, and any remaining conflicts regarding the coding task were discussed and solved

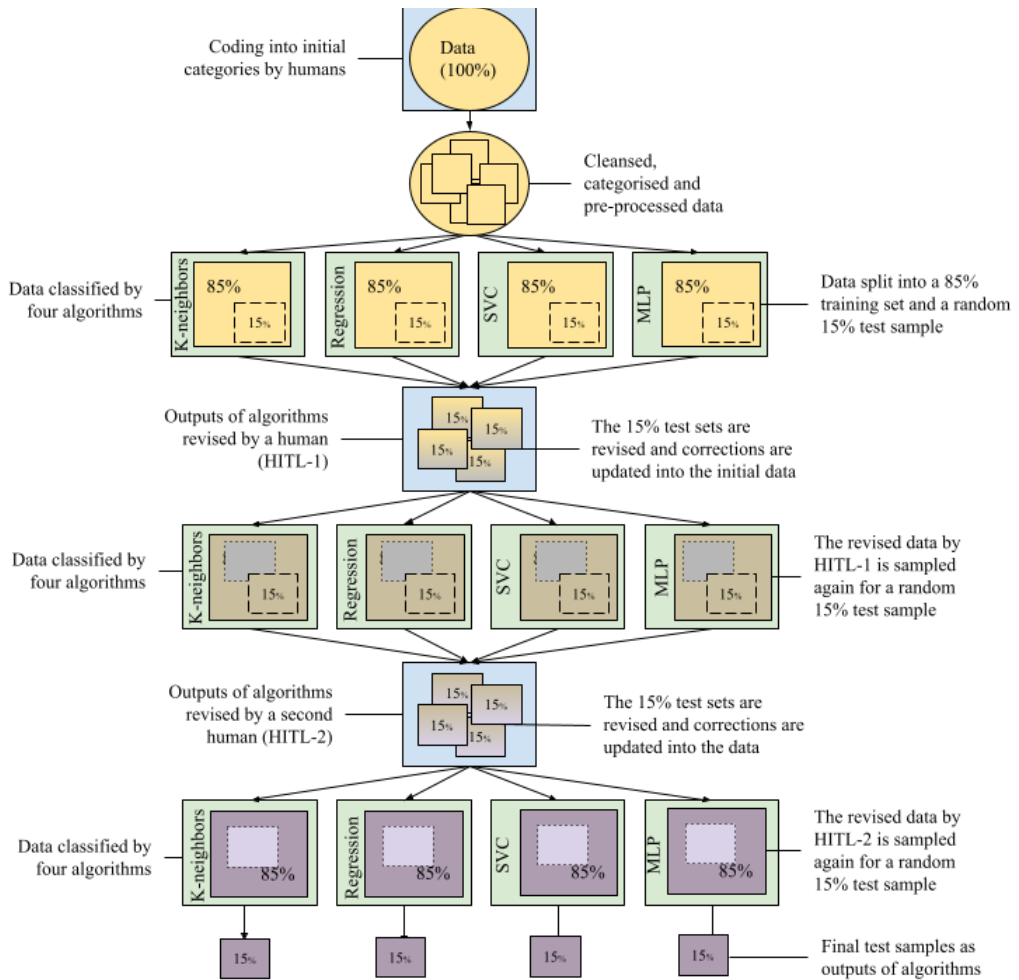


Fig. 1. The used methodological process.

in discussion with the authors. The assistants extracted the sentences describing the action, which were then used as data for SML with the assigned categories as labels.

B. Pre-processing

The texts were filtered so that resource categories with over 250 documents assigned to them were truncated to contain only 250 documents, leaving altogether 1553 items. The extracted action sentences were stripped of extra whitespace and processed with Python’s spaCy [44] library: Stopwords, punctuation, numerals, URLs, and email addresses were removed with spaCy’s inbuilt methods. The datasets were then vectorized with the frequency inverse document frequency (TF-IDF) [45-47] vectorizer from Python’s scikit-learn library [48].

C. Algorithm Choice

The pre-processed sentences were ran through a variety of ready scikit-learn machine learning algorithms that were suitable for multi-class classification with a small text dataset. Parameters were iterated to test which were the most suitable for our dataset size. The set of studied algorithms contained 21 different classifiers configurations, which were trained on 70% and 85% of the datasets and tested on 30% and 15% and assessed for accuracy and confusion matrices [49]. The 85% and 15% split was chosen for further analyses since no major differences emerged between the two splits. Out of these tests, the four classifiers that yielded the best performances in terms of accuracy over multiple runs and

samples, were chosen for further analysis. They were: Multilayer Perceptron (MLP) classification [50], Logistic Regression [51], SVC [52], and K-neighbors classification [53]. We focus on accuracy as a performance measure because it is often the selling point for algorithms [33].

Four random data samples were extracted to be ran through the four algorithms with arbitrary criteria for randomness in selection, which were as described as follows. The chosen parameter configurations were at least in some samples the best out of all. Then, for K-neighbors results the first sample in which K-neighbors with 10 as parameter for the number of neighbors was the best performing K-neighbors configuration was chosen as a sample, for Logistic Regression results the first sample in which Logistic Regression with standard parameters was the best performing Logistic Regression configuration was chosen as a sample, for SVC results the first sample in which SVC with the solver parameter set to “ovo” was the best performing SVC configuration was chosen as a sample, and for MLP results the first sample in which MLP with the following parameters was the best performing MLP configuration: $\alpha=1e-05$, $hidden_layer_sizes=(500, 100)$, $solver="lbfgs"$ was chosen as a sample.

D. Human-in-the-loop

The outputs of the algorithms (15% of the whole data as classified by the algorithm) were then re-coded by HITL-1 and the resulting new human revised labels were substituted

into the original dataset. The resulting dataset is referred to as “post HITL-1”. We use the 15% split to simulate a situation in which it would be implausible for a human coder to go fully through the size of a typical machine learning dataset. The same algorithms were ran with the post HITL-1 data, and the same sampling was used, accuracies were again recorded. Then, using the post HITL-1 data, the test sets were again re-coded by HITL-2 and again the new labels substituted into the dataset. The resulting dataset is referred to as “post HITL-2”. Both times, the humans were coding blind to the algorithm results. Again, accuracies were recorded.

E. Effects on Accuracy

Comparisons were made for the average performance before and after each human in the loop for each algorithm for all samples. A confidence value was recorded, and it was assessed whether HITL-1 or HITL-2 re-labelling significantly affected the accuracy of the algorithms. Confusion matrices were recorded for the predictions per the classifiers as well as the post HITL-1 and post HITL-2 datasets. During the coding, the humans kept a record of their own thoughts on the labelling process, and a brief log of the conversations that happened to unify the coding scheme before the first loop, after a small test set.

Afterwards, HITL-1 and HITL-2 were presented with the cases in which their classifications had differed, and both reclassified these instances independently. The results of this last round of classifications were discussed and remaining disagreements (altogether 75) were agreed upon in conversation. This conversation was documented, and notes were made on the perceived causes of difference, and the conclusions on how they should be resolved.

IV. RESULTS

Table 1 presents the accuracies before and after each HITL processing round and the accuracy differences according to sample. The overall average accuracy of the classifiers increases significantly post HITL-1, and decreases significantly post HITL-2, compared both to the initial pre HITL-1 labels and post HITL-1 classifications. Table 2 presents the accuracies before and after each human was introduced in the loop and their differences according to used classifier. No specific classifier can be seen as responsible for the phenomenon for the average accuracy increase post HITL-1, whereas regarding the decrease of accuracy post HITL-2, it can be said MLP classification was not significantly affected.

In most samples, there exists a trend in which HITL-1 corrections have focused on increasing the amount of classification into category zero (informational resources), and from zero to others. This is expected, since in the discussions leading up to HITL-1 processing, the informational category was formally structured via discussions to include the news in which the sentence describing the action includes a verb of information sharing. This was due to specifications on what constitutes an action, even if the general theme of the news piece might differ.

No other category was scrutinized as thoroughly pre HITL-1, but some attention was paid to differences between classes one and two. For instance, “Kodak announces plans to cut 200 jobs”, is an informational action, “Kodak plans to cut 200 jobs” is an organizational (class 2) action of planning, whereas “Kodak cuts 200 jobs” is an action concerning human resources (class 1).

Here, clearer disagreements appear. These are the cases that contribute to the decreases in accuracy post HITL-2.

TABLE I. ACCURACIES OF STUDIED CLASSIFIERS BEFORE AND AFTER EACH HUMAN RE-LABELED A SAMPLE OF THE

| | Classifiers | Accuracy | | | Difference | | |
|----------------------|---------------------|----------|-------|-------|------------|----------|----------|
| | | H0 | H1 | H2 | H1-H0 | H2-H0 | H2-H1 |
| Sample 1 | Logistic Regression | 0.645 | 0.645 | 0.640 | 0.000 | -0.005 | -0.005 |
| | SVC | 0.632 | 0.619 | 0.618 | -0.013 | -0.014 | -0.001 |
| | MLP | 0.628 | 0.589 | 0.601 | -0.039 | -0.027 | 0.012 |
| | K-neighbors | 0.619 | 0.589 | 0.583 | -0.030 | -0.036 | -0.005 |
| Sample 2 | SVC | 0.667 | 0.714 | 0.605 | 0.048 | -0.061 | -0.109 |
| | Logistic Regression | 0.654 | 0.675 | 0.614 | 0.022 | -0.040 | -0.061 |
| | MLP | 0.584 | 0.636 | 0.605 | 0.052 | 0.021 | -0.031 |
| | K-neighbors | 0.610 | 0.645 | 0.539 | 0.035 | -0.071 | -0.106 |
| Sample 3 | Logistic Regression | 0.628 | 0.701 | 0.610 | 0.074 | -0.018 | -0.092 |
| | MLP | 0.606 | 0.684 | 0.596 | 0.078 | -0.010 | -0.087 |
| | K-neighbors | 0.589 | 0.688 | 0.557 | 0.100 | -0.032 | -0.131 |
| | SVC | 0.589 | 0.710 | 0.627 | 0.121 | 0.038 | -0.083 |
| Sample 4 | Logistic Regression | 0.688 | 0.680 | 0.610 | -0.009 | -0.079 | -0.070 |
| | SVC | 0.662 | 0.662 | 0.614 | 0.000 | -0.048 | -0.048 |
| | K-neighbors | 0.658 | 0.658 | 0.548 | 0.000 | -0.110 | -0.110 |
| | MLP | 0.636 | 0.623 | 0.623 | -0.013 | -0.014 | -0.001 |
| Mean | | 0.631 | 0.657 | 0.600 | 0.027 | -0.031 | -0.058 |
| St.Dev. ^a | | 0.030 | 0.038 | 0.028 | 0.046 | 0.036 | 0.046 |
| 95% CI ^b | | | | | 0.004 | (-0.049) | (-0.080) |
| | | | | | -0.049 | (-0.014) | (-0.035) |

H0: INITIAL LABELS, H1: DATASET WITH 15% RE-LABELED BY HUMAN, H2: DATASET WITH 15% RE-LABELED BY A SECOND HUMAN

^a Standard deviation

^b Confidence interval

Most disagreements happen between classes 4 and 0: financial and informational. What HITL-2 has coded as class 2 has been more various things according to HITL-1: Many previously informational, physical, and financial news have been coded as organizational by HITL-2. Many originally relational news had been coded into informational or physical, and previously organizational news as physical by HITL-2. In discussing the re-classifications done by both HITLs on points of disagreements, the trends of Fig. 3 persisted and clear differences of points of view were the most common reason. Out of the whole set of disagreements 59 were agreed upon without discussion on a second look, but 75 were still disagreed upon. The most common points of disagreement were clear: HITL-2 did not consider extra clauses in the texts, whereas HITL-1 included the clauses and classified them based on action appearing in the clause, altogether 10 were such cases E.g. “Eastman Kodak Co. will lay off 66 employees from its Kettering operations by April 19, the company said in a notice filed with the state” was classified as a human resource action by HITL-2 and an informational action by HITL-1. Understanding that the vectorization of text to numerical does not acknowledge what is clause and what not, as a “bag of words”, the ignorance of clauses is likely a major confusing cause for the classifications post HITL-2.

Another 10 differences were caused by news that HITL-2 treated as organizational planning, but which were commonly classified as either informational or financial actions by HITL-1. E.g., “The group plans to shut its photo film finishing plant at Annesley in Nottingham with the loss of 350 jobs.” and “Eastman Kodak Co. plans to save about \$223 million by slashing healthcare benefits to about 16,030 retired employees and their dependents. Kodak asked to terminate

certain benefits effective May 1.” respectively. This was resolved by restricting planning as organizational, since no informational action can be inferred.

HITL-2 classified market situation descriptions as financial and HITL-1 as informational. HITL-1 classified the posting of financial reports as a financial resource, whereas HITL-2 considered the action “to post” as informational. These explain the great discrepancy between classes 0 and 4. Another difference appeared on some news that included the announcements by Kodak or other organizations. HITL-2 had coded all as informational, whereas HITL-1 classified 7 according to the contents of the announcement. The differences of classification depicting market descriptions were realized to be relational – to competitors – whereas previously HITL-1 had mostly coded them as informational and HITL-2 as financial. This was only possible due to the emergent disagreement via the HITL process.

Moreover, cases using verbs such as “expect”, “forecast”, and “estimate” were classified differently by both HITLs and with slight variance. However, in discussion it was realized that these are all informational resources that do not necessarily concern sharing information but does focus on the processing of information. Hence, the informational class criterion was refined post HITL-2.

Within the HITL process, via discussions, a formalization logic for the classification became clearer and more structured throughout. This resulted in news being classified according to the formal logic, at times against what would appear to most sensible at first intuition. E.g., “Kodak forecast a loss of \$200 million to \$400 million from continuing operations in 2009” seems to concern financial resources, but becomes in fact an information resource.

TABLE II. THE ACCURACIES OF THE STUDIED CLASSIFIERS BEFORE AND AFTER EACH HUMAN RE-LABELLED A SAMPLE OF THE DATASET ACCORDING TO USED CLASSIFIER

| | Classifier | Difference | | | Classifier | Difference | | |
|----------------------|---------------------|------------|----------|----------|-------------|------------|----------|----------|
| | | H1 - H0 | H2 - H0 | H2 - H1 | | H1 - H0 | H2 - H0 | H2 - H1 |
| Sample 1 | Logistic Regression | 0.000 | -0.005 | -0.005 | MLP | -0.039 | -0.027 | 0.012 |
| Sample 2 | | 0.022 | -0.040 | -0.061 | | 0.052 | 0.021 | -0.031 |
| Sample 3 | | 0.074 | -0.018 | -0.092 | | 0.078 | -0.010 | -0.087 |
| Sample 4 | | -0.009 | -0.079 | -0.070 | | -0.013 | -0.014 | -0.001 |
| Mean | | 0.022 | -0.035 | -0.057 | | 0.019 | -0.007 | -0.027 |
| St.Dev. ^c | | 0.032 | 0.028 | 0.032 | | 0.047 | 0.017 | 0.038 |
| 95% CI ^d | | (-0.010) | (-0.063) | (-0.088) | | (-0.027) | (-0.024) | (-0.064) |
| | | -0.053 | (-0.008) | (-0.025) | | -0.066 | -0.010 | -0.011 |
| | Classifier | Difference | | | Classifier | Difference | | |
| | | H1 - H0 | H2 - H0 | H2 - H1 | | H1 - H0 | H2 - H0 | H2 - H1 |
| Sample 1 | SVC | -0.013 | -0.014 | -0.001 | K-neighbors | -0.030 | -0.036 | -0.005 |
| Sample 2 | | 0.048 | -0.061 | -0.109 | | 0.035 | -0.071 | -0.106 |
| Sample 3 | | 0.121 | 0.038 | -0.083 | | 0.100 | -0.032 | -0.131 |
| Sample 4 | | 0.000 | -0.048 | -0.048 | | 0.000 | -0.110 | -0.110 |
| Mean | | 0.039 | -0.021 | -0.060 | | 0.026 | -0.062 | -0.088 |
| St.Dev. ^c | | 0.053 | 0.039 | 0.041 | | 0.048 | 0.031 | 0.049 |
| 95% CI ^d | | (-0.013) | (-0.059) | (-0.100) | | (-0.021) | (-0.093) | (-0.136) |
| | | -0.090 | -0.017 | (-0.020) | | -0.073 | (-0.031) | (-0.040) |

H0: INITIAL LABELS, H1: DATASET WITH 15% RE-LABELLED BY HUMAN, H2: DATASET WITH 15% RE-LABELLED BY A SECOND HUMAN

^c. Standard deviation

^d. Confidence interval

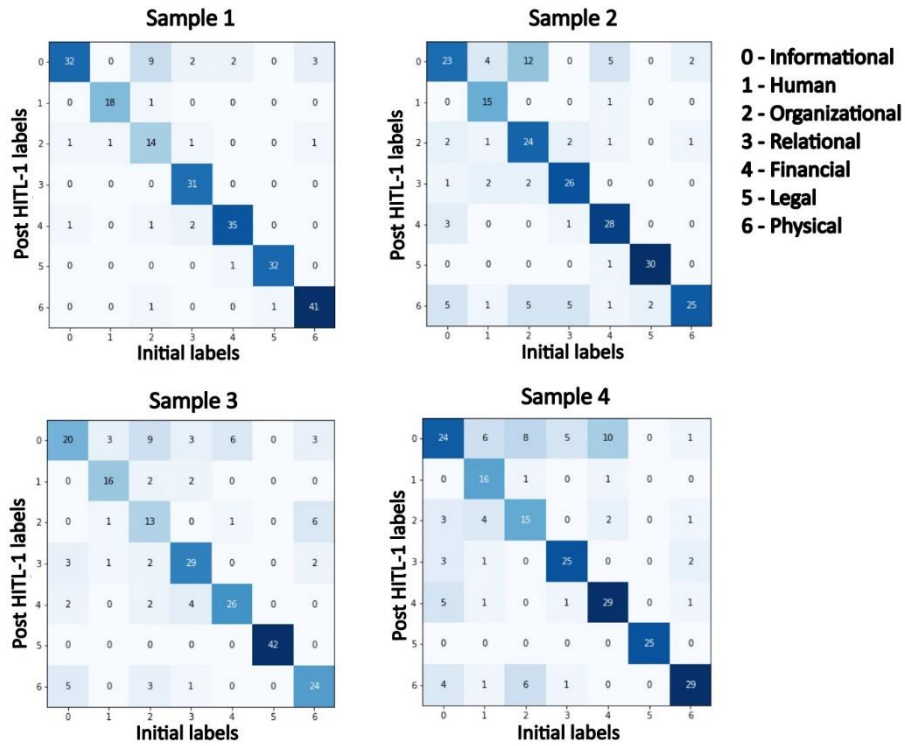


Fig. 2. Differences in classifications by human-in-the-loop one and the initial labels.

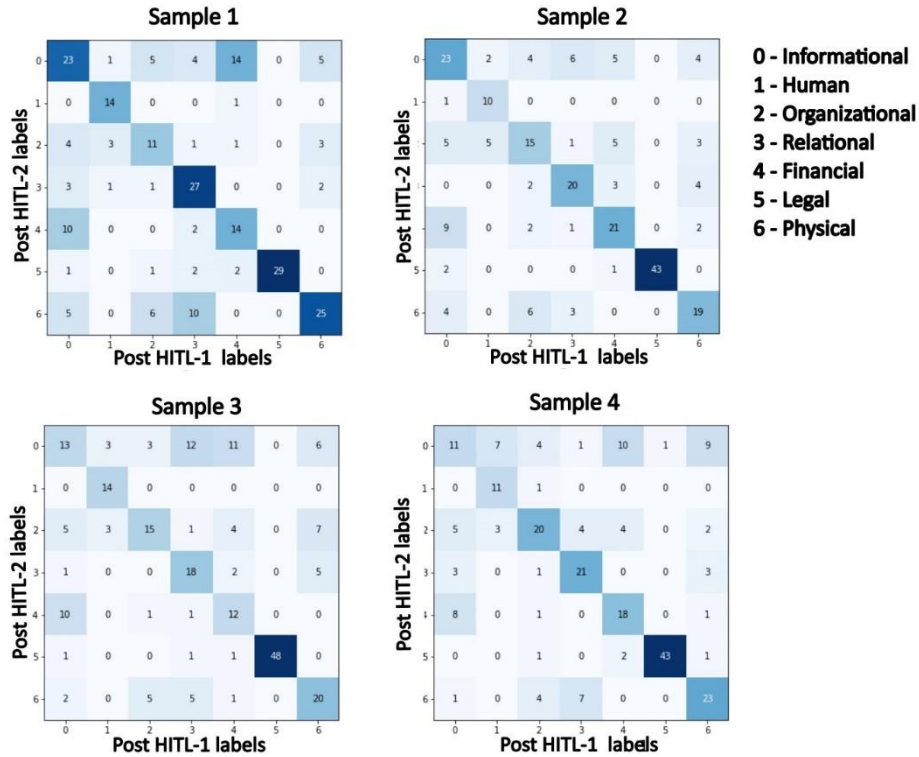


Fig. 3. Differences in classifications by human-in-the-loop two and the labels augmented with the corrections by human-in-the-loop one.

This hybrid learning and refining was the main cause which enabled the humans in the loop to explain the classification decisions by both themselves and the algorithms trained on these decisions. Due to the refinement of the coding and increased understanding, explainability, and consequent transparency and accountability were more attainable. However, this would not have been the case if the process did not allow for discrepancies and focused only on

achieving the best accuracy possible for the task as fast as possible with or without a human in the loop.

V. DISCUSSION AND CONCLUSIONS

We find, in accordance with previous studies [33], that humans begin to reflect on what they know, why and how in a critical manner. This thought process then translates into the AI process, which is the key enabler for increased explainability of the AI: the humans in the loop become more

able to explain the decisions made for the labelling of the data in training. Hence, the concepts of hybrid learning and XAI were equivalent in our process. However, the increase in explainability did come with trade-offs. The taxonomy of labelling the resource categories became more abstract, but this enables understanding and explaining it, and observing whether the AI follows the same logic, or whether machine-human discrepancies need to be dealt with similarly to human-human discrepancies.

Some discrepancies were present throughout the refinement of the coding scheme – during the first and last discussions, and as a conclusion, the whole taxonomy would need to be rethought: Missing classes were identified that should be added later, and a hierarchical labelling system was deemed better than the original, and the technical realities regarding clauses became clearer in the coding instructions.

Altogether, without the intentional exploitation of different types of human bias, at the cost of accuracy, explainability would have become a trade-off for accuracy. Hence, we argue based on our results that in the emergent HITL processes, increasing average accuracy of a varied methodology is possible, but focusing on maximizing accuracy performance for a certain task instead of contesting the classification via implicit existing human biases, may lead to less-than-optimal explainability of the AI process. HITL can increase both accuracy and explainability, but it is no magic tool that will automatically increase both. Instead, how HITL is implemented in the analysis process becomes important to ensure that neither is neglected. Hence, in pushing for AI adoption, HITL discussion and recommendations need clarification and further research, since the present situation of incongruities of HITL promises and fairness trade-offs does not give a clear picture of HITL. HITL becoming biased in the loop to optimize for accuracy does not ensure fairness and simultaneously, HITL can decrease accuracy – at least in the short term.

ACKNOWLEDGMENT

We thank the three anonymous reviewers for the constructive feedback and guidance. We also thank H. Pukkala for research assistance. The usual disclaimer applies.

REFERENCES

[1] K. Schwab, *The fourth industrial revolution*. Currency, 2017.

[2] D. Chalmers, N. G. MacKenzie, and S. Carter, "Artificial intelligence and entrepreneurship: implications for venture creation in the fourth industrial revolution." *Entrepreneurship Theory and Practice*, vol. 45, no. 5, pp. 1028–1053, 2021.

[3] S. Indhul, "Towards a Constructor Theory Conception for Wicked Social Externalities: Delineating the Limits and Possibilities of Impactful Pathways to a Better World" in *Transcendent Development: The Ethics of Universal Dignity*, vol. 25, A. Thakathi, Eds. Bingley: Emerald Publishing Limited, 2022, pp. 43–52.

[4] G. Cao, and D. Yanqing, "How do top-and bottom-performing companies differ in using business analytics?," *Journal of Enterprise Information Management*, vol. 30, no. 6, pp. 874–892, 2017.

[5] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. J. Patil, and D. Barton, "Big data: the management revolution." *Harvard business review*, vol. 90, no.10, pp. 60-88, 2012.

[6] O. E. Olabode, N. Boso, M. Hultman, and C. N. Leonidou, "Big data analytics capability and market performance: The roles of disruptive business models and competitive intensity", *Journal of Business Research*, vol. 139, pp. 1218–1230, 2022.

[7] C. Keding, and P. Meissner, "Managerial overreliance on AI-augmented decision-making processes: How the use of AI-based advisory

systems shapes choice behavior in R&D investment decisions", *Technological Forecasting and Social Change*, vol. 171, pp. 120970, 2021.

[8] D. Shin, "User perceptions of algorithmic decisions in the personalized AI system: perceptual evaluation of fairness, accountability, transparency, and explainability." *Journal of Broadcasting & Electronic Media*, vol. 64, no. 4, pp. 541-565, 2020.

[9] A. Fügener, J. Grahl, A. Gupta, and W. Ketter, "Will Humans-in-the-Loop Become Borgs? Merits and Pitfalls of Working with AI", *Management Information Systems Quarterly (MISQ)*, Vol. 45, 2021.

[10] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt, "It's reducing a human being to a percentage' perceptions of justice in algorithmic decisions." *Proceedings of the 2018 Chi conference on human factors in computing systems*, 2018.

[11] M. H. Teodorescu, L. Morse Y. Awwad, and G. C. Kane, "Failures of fairness in automation require a deeper understanding of human-ML augmentation", *MIS Quarterly*, vol. 45, no. 3, 2021.

[12] S. Raisch, and S. Krakowski, "Artificial intelligence and management: The automation–augmentation paradox", *Academy of Management Review*, vol. 46, no. 1, pp. 30 (6): pp. 874–892, 2021.

[13] J. Dodge, Q. V. Liao, Y. Zhang, R. K. Bellamy, and C. Dugan, "Explaining models: an empirical study of how explanations impact fairness judgment", *Proceedings of the 24th international conference on intelligent user interfaces*, 2019.

[14] G. M. Johnson, "Algorithmic bias: on the implicit biases of social technology", *Synthese*, vol. 198, no.10, pp. 9941-9961, 2021.

[15] A. Rai, "Explainable AI: From black box to glass box", *Journal of the Academy of Marketing Science*, vol. 48, no. 1, pp. 137-141, 2020.

[16] X. Zeng, F. Song, Z. Li, K. Chusap, C. Liu, "Human-in-the-loop model explanation via verbatim boundary identification in generated neighborhoods". *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, Springer, Cham, 2021.

[17] S. Asthana, S. Kwatra, C. T. Wolf, P. Chowdhary, and T. Nakamura. "Human-in-the-Loop Business Modelling for Emergent External Factors." *2020 IEEE International Conference on Big Data (Big Data)*, IEEE, 2020.

[18] A. Correia, and F. Lecue, "Human-in-the-loop feature selection." *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, No. 01, 2019.

[19] I. Rahwan, "Society-in-the-loop: programming the algorithmic social contract", *Ethics and Information Technology*, vol. 20, no. 1, pp. 5-14, 2018.

[20] D. Dellermann, A. Calma, N. Lipusch, T. Weber, S. Weigel, and P. Ebel, "The future of human-AI collaboration: a taxonomy of design knowledge for hybrid intelligence systems", *arXiv preprint arXiv:2105.03354*, 2021.

[21] J. Ostheimer, S. Chowdhury, and S. Iqbal, "An alliance of humans and machines for machine learning: Hybrid intelligent systems and their design principles," *Technology in Society*. vol. 66, p. 101647, 2021.

[22] S. S. Sundar, "Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAI)," *Journal of Computer-Mediated Communication*, vol. 25, no. 1, pp. 74-88, 2020.

[23] B. M. Abdel-Karim, N. Pfeuffer, G. Rohde, and O. Hinz "How and what can humans learn from being in the loop?" *KI-Künstliche Intelligenz*, vol. 34, no. 2, pp. 199-207, 2020.

[24] T. Grønsund, Tor, and M. Aanestad, "Augmenting the algorithm: Emerging human-in-the-loop work configurations," *The Journal of Strategic Information Systems*, vol. 29, no. 2, p. 101614, (2020).

[25] N. S. Hernandez, S. Ayo, and D. Panagiotakopoulos, "An Explainable Artificial Intelligence (xAI) Framework for Improving Trust in Automated ATM Tools," *2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC)*, pp. 1–10, 2021.

[26] A. Adadi, and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)," *IEEE access*, vol. 6, pp. 52138-52160, 2018.

[27] N. Kordzadeh, and M. Ghasemaghaei. "Algorithmic bias: review, synthesis, and future research directions," *European Journal of Information Systems*, pp. 1-22, (2021).

- [28] A. Gandomi, Amir, and H. Murtaza, "Beyond the hype: Big data concepts, methods, and analytics," *International journal of information management*, vol. 35, no. 2, pp. 137-144, (2015).
- [29] T. J. Robinson, R. C. Giles, and R. U. Rajapakshage, "Discussion of "Experiences with big data: Accounts from a data scientist's perspective"," *Quality Engineering*, vol. 32, no. 4, pp. 543-549, (2020).
- [30] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. (521), vol. (7553), pp. 436-444, 2015.
- [31] T. Rouleau, "What are the types of machine learning," Sama, Retrieved from <https://www.sama.com/blog/types-of-machine-learning>, 2020.
- [32] A. F. Cooper, and E. Abrams, "Emergent Unfairness in Algorithmic Fairness-Accuracy Trade-Off Research," *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021.
- [33] S. Lebovitz, N. Levina, and H. Lifshitz-Assaf, "Is AI ground truth really "true"? The dangers of training and evaluating AI tools based on experts' know-what," *Management Information Systems Quarterly*, in-press, 2021.
- [34] G. Adomavicius, J. C. Bockstedt, S. P. Curley, and Zhang, "Reducing recommender system biases: An investigation of rating display designs," *MIS Quarterly*, vol. 43, no. 4, pp. 1321-1341, 2019.
- [35] J. Gunaratne, L. Zalmanson, and O. Nov, "The persuasive power of algorithmic and crowdsourced advice," *Journal of Management Information Systems*, vol. 35, no. 4, pp. 1092-1120, 2018.
- [36] J. M. Logg, J. A. Minson, and D. A. Moore, "Algorithm appreciation: People prefer algorithmic to human judgment," *Organizational Behavior and Human Decision Processes*, vol. 151, pp. 90-103, 2019.
- [37] A. Agrawal, J. Gans, and A. Goldfarb, "How to win with machine learning." *Harvard Business Review*, 2020.
- [38] A. Jain, "Weapons of math destruction: how big data increases inequality and threatens democracy," pp. 123-125, (2017).
- [39] C. Rosso, Cami. "The human bias in the AI machine," *Psychology Today*, Retrieved from <https://www.psychologytoday.com/us/blog/the-future-brain/201802/the-human-bias-in-the-ai-machine>, 2018.
- [40] S. Tonidandel, E. B. King, and J. M. Cortina, "Big data methods: Leveraging modern data analytic techniques to build organizational science," *Organizational Research Methods*, vol. 21, no. 3, pp. 525-547, (2018).
- [41] D. Shin, and Y. J. Park, "Role of fairness, accountability, and transparency in algorithmic affordance," *Computers in Human Behavior*, vol. 98, pp. 277-284, 2019.
- [42] Z. Da, and X. Huang, "Harnessing the wisdom of crowds," *Management Science*, vol. 66, no. 5, pp. 1847-1867, 2020.
- [43] R. M. Morgan, and S. Hunt, "Relationship-based competitive advantage: the role of relationship marketing in marketing strategy," *Journal of Business Research*, vol. 46, no. 3, pp. 281-290, 1999.
- [44] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength Natural Language Processing in Python", Zenodo, 2020, <https://doi.org/10.5281/zenodo.1212303>.
- [45] G. Salton, and C. Buckley, "Term-weighting approaches in automatic text retrieval", *Information Processing & Management*, vol. 24, no.5, pp. 513-523, 1988.
- [46] H. Schütze, C. D. Manning, and P. Raghavan, Introduction to information retrieval (Vol. 39), Cambridge University Press Cambridge, 2008
- [47] scikit-learn documentation. (n.d.-a). https://scikitlearn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html?highlight=vectorizer#sklearn.feature_extraction.text.TfidfVectorizer%7D
- [48] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.
- [49] scikit-learn documentation. (n.d.-b). <https://scikitlearn.org/stable/modules/classes.html#module-sklearn.metrics>
- [50] scikit-learn documentation. (n.d.-c). https://scikitlearn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html?highlight=mlpclassifier#sklearn.neural_network.MLPClassifier
- [51] scikit-learn documentation. (n.d.-d). https://scikitlearn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html?highlight=logisticregression#sklearn.linear_model.LogisticRegression
- [52] scikit-learn documentation. (n.d.-e). <https://scikitlearn.org/stable/modules/generated/sklearn.svm.SVC.html>
- [53] scikit-learn documentation (n.d.-f). <https://scikitlearn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>

PUBLICATION

IV

**Exploring the relationships between artificial intelligence transparency,
sources of bias, and types of rationality**

Valtonen, L., and Mäkinen, S. J.,

*In Proceedings of the 2022 IEEE International Conference on Industrial Engineering and
Engineering Management (IEEM) (pp. 1296–1300)*

Publication reprinted with the permission of the copyright holders.

Exploring the relationships between artificial intelligence transparency, sources of bias, and types of rationality

L. Valtonen¹, S.J. Mäkinen²

¹ Industrial Engineering and Management, Tampere University, Tampere, Finland

² Department of Mechanical and Material Engineering, University of Turku, Finland
(laura.valtonen@tuni.fi)

Abstract - Artificial intelligence (AI) is permeating one human endeavor after another. However, there is increasing concern regarding the use of AI: potential biases it contains, as well as mis-judged AI use. This study continues the recent investigations into the biases and issues that are potentially introduced into human decision-making with AI. We experimentally set-up a decision-making classification task and observe human classifiers when they are guided in their decision-making either by AI or other humans. We find that over-reliance or authoritative stigmatization is present when AI is concerned and that with human guidance discursive explanatory decision-making is present. We conclude that while AI is seen as authoritative even in a low stake decision-making setting, it does not suppress choice, but combined with a lack of transparency, AI suppresses visibility into rationality creation by the decision maker. Based on the emergent explorative relationships between types of rationality, AI transparency and authoritativeness, we provide future research avenues based on our findings.

Keywords – Artificial intelligence, decision-making, bias, rationality

I. INTRODUCTION

Artificial intelligence (AI) is often portrayed as a major transforming force for industry (e.g., [1,2]). A clear reason for AI utilization is its competitive promise; Organizations that exploit data quantities typical for AI, or big data analysis, have been more successful than those that do not [3-5]. However, along increasing AI importance, the need for research and dialogue between AI and management and organizational scholars is rising [6], since poor management of the idiosyncrasies associated with the use of big data can lead to failures to meet expectations [7].

Expectations fail even in cases in which the utilized AI approaches are of high accuracy and based on expert data (e.g., [8]). One suggested reason for such shortcomings is that AI is often developed and tested and optimized for accuracy in laboratory settings that do not translate into organizational realities [6]. Another possibility for failed expectations is simply that the expectations were too high to begin with [9]; The *technology effect* refers to the presence of excessive optimism regarding unfamiliar technology being equated with success. This biased thinking permeates experts and novices alike, as well as parties both pitching and investing in technology [10-12]. In reference to algorithms specifically, the term “algorithm appreciation” is used [13], and is observed in AI-

augmented decision making, which can lead to failures regarding insufficient risk accounts and uncertainty considerations [14]. In addition to success, people equate novel technology with mystery, alienness, and complexity [12]. The mystery of novel technology is posed as a reason to forego critical analysis and attempts and understanding the technology in question [12,15].

The technology effect is attributed to constant exposure to technology success stories [10,11]. Such constant availability of the connection between technology and success can create biased thinking towards implicit associations between the narrated relationship [16]. In addition to overrepresented narratives of technology successes [11,12], suggested reasons for misplaced optimism and overreliance on AI include narratives of AI as an objective, unbiased, value-free [14,17-20] “supercarrier of formal rationality” [9,18].

Formal rationality is a bureaucratic means-end rationality, which relies on calculations based on universals and abstracts. It is contrasted with substantive rationality, which is based on values and allows for a plurality of rationalization of action in accordance with the actor’s values. [21] However, the conceptualization of AI as a formal rationality is problematic; When talking about artificial intelligence and its contemporary successes, usually a supervised machine learning (SML) model is being applied [22,23]. SML algorithms require an originally human labelled dataset, based on which a classifier algorithm is built to map a certain input to its given label. The classifier is evaluated on its performance, or often “accuracy”, which refers to how many input items it could correctly label in a certain test dataset.

The SML data can contain biased inputs or labels or both: the data can be incomplete [24] or reflect social bias in the labels (e.g., [25-27]). Moreover, outside of data, bias can creep into AI processes from the algorithms themselves [28], or the interpretations of results [29]. Hence, if AI is conceptualized as a bias-free source of objective information, it transforms what is inherently substantively rational label data created by value-laden, subjective humans into a means-end calculation optimized for a certain outcome. This suppresses substantive rationality and represents it as formal rationality. This project of rationality metamorphosis via AI could lead to the “end of choice” [18] in which we hide possible pluralities of rationalization via AI and risk substantive rationality [18] and the maintenance of unique human knowledge in organizations [30].

In the face of such obstacles and acknowledged issues with AI, the perseverance of optimism regarding this novel information technology [10-11,14] can be explained by information avoidance [9]. Information avoidance refers to active avoidance of information that people are aware that exist and would be free to access. This may be done to “bind one’s own hands while facing an inner conflict” and lead to even the abandonment of responsibility of actions and decisions. Forms of information avoidance include physical avoidance, inattention, biased interpretation of information, forgetting, and self-handicapping. Reasons for information avoidance include disappointment and regret aversion, dissonance avoidance, and optimism maintenance. [31,32]

Dissonance refers to an unpleasant mental state caused by an individual holding contradictory cognitions. The resolution of this feeling is achieved by changing cognitions to a desired logical state [33,34]. The impact of cognitive dissonance is recognized widely to impact management [35]. Dissonance emerges as new information is attained that contradicts previously held cognitions [34], thus, methods of information avoidance can help resolve or mitigate the emergent dissonance. For instance, failures of novel technology can create dissonance with cognitions primed with technology being inherently associated with success. This dissonance may prompt information avoidance regarding issues with the novel technology.

Another relevant cognitive bias is confirmation bias: Confirmation bias emerges when people apply information avoidance in an information search to find confirmatory results of their prior cognitions [29]. This bias also affects collections of people and may lead to what is called “groupthink”: Delusions of optimism due to information avoidance and willful interpretation of information are contagious in groups. In hierarchies these delusion “trickle down” from leaders [36]. Assuming that such leaders may be experts, it is interesting to note that in regarding the technology effect, experts are more prone to polarized beliefs and using their information and intelligence to enforce their prior beliefs [31,37].

We know “what”: Managers overrely on AI. We have literature suggestions as to “why”: the technology effect [10,11], algorithm appreciation [13], and dominant narratives of technology success and superior rationality [12,18]. We even have suggestions for “why” for the “whys”: information avoidance [9,31]. However, the discussion between the posed relationships between all concepts above remains hypothetical: There is scarce if any empirical research on *the reasons why* people come to overrely on AI. We begin to address the question “how does decision-making *reasoning* differ in an AI-guided setting in comparison to a human-guided setting?” We study the differences that contribute to dissimilar, possibly overly optimistic, views on AI guidance in comparison to human guidance.

In our setting people perform a simple decision-making task of categorizing news to predefined categories. After initial categorization, respondents are divided into two groups as the categories are used in decision-making;

First group is advised on validating categories by AI and the other group by other people. Both groups are told the source of validation information. We find that for the group guided by their peers, people elaborated on their thinking process rationale, while for the AI guided group, people defended their decisions. Our research directs attention especially to the relationship between AI transparency and creation of decision-making rationales.

II. METHODOLOGY

To address the question “how does decision-making reasoning differ in an AI-guided setting in comparison to a human-guided setting?” we split six participants in a simple, no accountability, low risk decision-making situation into two groups of three: Those who see the recommendation for a decision given by a human agency (human group), and those who see the recommendation given by an SML-based AI (AI group). All participants perform a labelling task individually to form a subjective frame of reference on how the task should be performed, after which they complete the task again with a subset of the data along which they are provided the recommendations. The dataset and labelling were chosen as to hold inherent ambiguity to support various substantive rationalities to emerge on how the task should be done.

A. Data and Labelling

The dataset was collected using the keyword “European Organization for Nuclear Research” (CERN) from LexisNexis. The search was refined to include newswires, press releases, newspapers, and trade press news in English through the years 2016-2019. From the resulting news the sentences including the search term were extracted along with the sentence before and after. This yielded 1687 three sentence text documents.

The set was split into subsets for labelling so that every text is labelled by three separate people into one of the following categories based on CERN’s mission statement: (1) technology, (2) scientific knowledge, or (3) human capital. CERN’s mission statement “Our mission is to: (1) provide a unique range of particle accelerator facilities that enable research at the forefront of human knowledge, (2) perform world-class research in fundamental physics, (3) unite people from all over the world to push the frontiers of science and technology, for the benefit of all” [38] was given to the labelers, who were asked to label each news text according to the part of the mission they see most relevant to the text. The “human recommendation” labels were chosen as the democratic majority out of the three labels provided for each news text.

B. Artificial Intelligence Recommendations

The labelled texts were cleansed by removing extra whitespace and stopwords, punctuation, numerals, URLs, and email addresses were removed with the spaCy library for Python [39]. The corpus of text documents was vectorized for SML with the frequency inverse document

frequency (TF-IDF) [40-42] as well as the bag-of-words [43] vectorizers from Python's scikit-learn library [44]. Any rows with not-a-number values were dropped, which yielded altogether 1414 documents for further analysis.

A variety of multi-class classification supervised machine learning algorithms that were suitable for a small text dataset were selected out of ready scikit-learn machine learning algorithms and tested with varied parameters to find the best performing one according to performance on the accuracy and confusion matrix with all 70/30, 85/15, and 90/10 training-test data splits [45]. The best performing algorithm over all runs was the Multilayer Perceptron (MLP) classifier [46] with the following parameters: `alpha=1e-05, hidden_layer_sizes=(50, 10), solver="lbfgs"`. The classifications by the TF-IDF vectorization were superior to those of bag-of-words regarding the confusion matrix and thus those classifications were chosen as the "AI recommendations."

C. Interviews

The participants were surveyed for their attitudes towards AI [47] and split so that each group had a member who saw AI as in an overall positive light, an overall negative light, and a polarized light in the sense that they saw both great risks and possibilities in AI. Both groups had one participant who had experience in algorithm development and one participant who had no information technology background.

Each participant was called in for a structured interview regarding the labeling task they had performed without any knowledge as to what the interview would entail. Upon arrival to the interview the participants were informed that the interview would be recorded and given an instruction to categorize, within an hour, a set of 140 news pieces in a similar manner as previously and explain for each news piece why they selected a certain label. The difference between the groups was that the recommendation was given as either "by the best AI algorithm (from various compared ones) performing the same task" or "the best categorizations from various people performing the same task". The interviews were then analyzed in terms of common and differing themes, and the qualitative results are described below.

III. RESULTS AND DISCUSSION

Overall, all participants employed a ruling-out reasoning logic on several points of both disagreement and agreement. They weighed several label options and chose one through the negation of others, explaining why it is indeed not the other category. Some examples of this type of reasoning were answers such as: "Human capital, because does not directly concern technology facilities, and I don't really see this as science either," and "Here, there is no human mentioned, so scientific knowledge." Though not explicitly mentioned in the latter, the weighing between two categories is present. Every interviewee also employed this reasoning strategy through affirmation instead of negation: They weighed several options and chose a label

by affirming some attribute of the news text. Some examples of this type of logic were answers such as: "Well, this could be human capital or technology, but they have purchased technology from elsewhere in particular, so it is human capital," and "Even though the topic is very human-centered, technology, because the facilities are referenced." Thus, no overall indication emerged to suggest that the interviewees were averse to acknowledging the variety of choice they had in the labelling process per text.

However, one interviewee in the AI group expressed that some choices were made against their will, with phrases like "Well I would want to put scientific knowledge, but". No indications were given as to why they could not have chosen according to the expressed want, but this interviewee was one of the two interviewees who assumed that the decision for the recommended label had further or better information as its basis. Out of these two interviewees, the one in the human group used the wording "is apparently" and the one in the AI group used "appears to be". Despite the similar semantic, the difference was in that the interviewee in the AI group assumed that the algorithm had been able to access the whole news text instead of the three-sentence snippet, whereas the interviewee in the human group assumed that the other labelers had more knowledge regarding the subject matter in the text. Neither knew about the functionality of the applied algorithm or the human labelers, and thus, both assumptions were spontaneous.

The lack of transparency into the process of the recommendation generation was what enabled these assumptions to take place. Had the interviewee in the AI group known what data the algorithm had access to, such unfounded assumptions of its capabilities would be less understandable. Similarly, had the interviewee in the human group known the background of the creators of the recommended labels, the assumed difference in substantive knowledge could be traced back to some concrete evidence or information, while now the pessimistic self-assessments are spontaneous and without any referable comparative framework.

All interviewees expressed uncertainty and highlighted their subjectivity with phrases like "seems like", "probably", "maybe", "I feel like", and even "well, I have to guess." Interviewees in both groups exhibited rhetoric aimed at creating rapport with phrases like "I can see that, but" In such cases, however, two interviewees in the AI group used a defensive rhetoric that was not apparent in the interviews of the human group. Both interviewees, in disagreeing with the recommendation, said they were "sticking with" the label they assigned to the text. In similar situations in the human group, the language used did not describe a defensive action, as if their opinion was being challenged in the situation. Instead, they used phrases like "this is rather" and "maybe, this is still." The interviewee in the AI group with knowledge in algorithm development was rhetorically more alike the human group in this aspect.

Both interviewees who used this rhetoric of "sticking with" spontaneously acknowledged discomfort due to

seeing the AI recommendations. One said they were “annoyed to see the AI response/answer”, while the other interviewee said at the end of the interview that they tried their best not to look at the recommendations, because they thought they may become biased, despite both having very polarized views of AI: The other scored a high positive AI attitude on the preliminary survey, and the other a very negative one in comparison. Despite this difference, the technology effect appeared in both seeming to regard the algorithm recommendation as something authoritative that posed a threat to their subjective opinion. Moreover, one of these interviewees complemented the view of AI as an authority figure by using rhetoric like “I’m leaning on [the AI recommended label], because there is really no sense in this” when they expressed uncertainty regarding the text and came to choose the recommended label.

Within the human group interviewees, excluding the one outstandingly conformist interviewee who chose differently from the recommended label half the time less than the two others, a type of choice justification appeared that was not present in the AI group. Despite every interviewee explaining their earlier methodology to perform the labelling task to a certain degree, when these two interviewees elaborated on their label choice and referred to their own initial conceptualization about the labeling task, they used the original mission statement as a justification. In other words, they referred to instructional documents to base their own subjective framework on, thus performing a similar type of conflation between formal and substantive rationalities as described as a function of AI previously [18]. For example, both “I put all these benefitting organizations into human capital, because no other mission statements describe”, and “Patents I put into human capital, because they are not directly concerned with the other mission statements” were offered by one interviewee. The other interviewee elaborated with examples such as “Yeah, here, from the definition of the human capital, the generation of common good is probably fulfilled, so I’m putting this into human capital” and “From the definition, the creation of common good, so I’m putting it there”. Thus, this dialogue between the “definition of the mission statement”, formal rationality, and “I”, substantive rationality, appeared throughout the interviews.

With these results, we may begin to approach empirically why people may come to overrely on AI. The differences in reasoning and rationality in our research between groups decision-making with AI and human recommendations suggest that formal rationality may be seen as inherently already present in the AI assisted decision-making situation, whereas formal rationality was both referenced and actively created within the human-assisted decision-making context. Such spontaneous assumptions of formal rationality to exist for an AI are present in common AI narratives [9, 14, 18-20].

Moreover, the lack of transparency into the creation of the AI recommendation was both an enabling factor for people to make that assumption of rationality, and a deterrent to potentially scrutinizing it. In our setting, the presence of AI suppressed a discussion and synthesis

between substantive and formal rationalities: A vacuum regarding information on the decision rationale process development was created. If AI leads to less information shared regarding the decision-making rationale and its development, AI overreliance can be both a result of and a cause of information scarcity – be it serendipitous or self-induced.

V. CONCLUSION

The emergent defensive rhetoric of “sticking with” and the dialectic creation of rationality within the human group but not the AI group reflects a situation in which the AI is perceived with a different type of authority: It is not negotiated with and there is little reward in explaining your thinking to something that is incapable of seeing your logic or point. Due to the set-up having been with no stakes for the participants, the authority yielded to the AI is notable. In the future, further research into the relationship between uncertainty and AI authority could be furthered by adding stakes into the setting. Moreover, further research requires larger sample sizes, since the small number of interviewees sets clear limitations to this study.

With this initial explorative study, we have been able to tease out details of the relationships between AI transparency, the technology effect, and types of rationality for further empirical research. Our research directs attention especially to the relationship between AI transparency and creation of decision-making rationales. Even if free choice was equally present with AI assistance as human assistance, the visibility into the process for making that choice dissolved with AI. An opaque AI does not offer its own rationale for discussion or scrutiny, and here its presence suppressed elaborations of the way people merged their substantive and given formal rationales for decision-making, furthering the lack of transparency and information in the research setting.

ACKNOWLEDGMENT

We thank the reviewers for the constructive feedback and guidance. The usual disclaimer applies.

REFERENCES

- [1] K. Schwab, *The fourth industrial revolution*. Currency, 2017.
- [2] D. Chalmers, N. G. MacKenzie, and S. Carter, “Artificial intelligence and entrepreneurship: implications for venture creation in the fourth industrial revolution,” *Entrepreneurship Theory and Practice*, vol. 45, no. 5, pp. 1028–1053, 2021.
- [3] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. J. Patil, and D. Barton, “Big data: the management revolution,” *Harvard business review*, vol. 90, no. 10, pp. 60–88, 2012.
- [4] G. Cao, and D. Yanqing, “How do top-and bottom-performing companies differ in using business analytics?”, *Journal of Enterprise Information Management*, vol. 30, no. 6, pp. 874–892, 2017.
- [5] O. E. Olabode, N. Boso, M. Hultman, and C. N. Leonidou, “Big data analytics capability and market performance: The roles of disruptive business models and competitive intensity”, *Journal of Business Research*, vol. 139, pp. 1218–1230, 2022.

- [6] S. Raisch, and S. Krakowski, "Artificial intelligence and management: The automation–augmentation paradox," *Academy of Management Review*, vol. 46, no. 1, pp. 874–892, 2021.
- [7] J. Gao, A. Koronios, and S. Selle. "Towards a process view on critical success factors in big data analytics projects," in *Proc. 21st Americas Conference on Information Systems, AMCIS (2015)*, Puerto Rico, pp. 1-14.
- [8] S. Lebovitz, N. Levina, H. Lifshitz-Assaf, "Is AI ground truth really "true"? The dangers of training and evaluating AI tools based on experts' know-what," *Management Information Systems Quarterly*, vol. 45, no. 3b, pp. 1501-1525, 2021.
- [9] M. Zaitsava, E. Marku, and M.C. Di Guardo, "Is data-driven decision-making driven only by data? When cognition meets data." *European Management Journal*, in press, 2022.
- [10] B. B. Clark, C. Robert, and S.A. Hampton, "The technology effect: how perceptions of technology drive excessive optimism," *Journal of Business and Psychology*, vol. 31, no.1, pp.87-102, 2016.
- [11] T. C. Dunne, B. B. Clark, J.P. Berns, and W.C. McDowell, "The technology bias in entrepreneur-investor negotiations," *Journal of Business Research*, vol. 105, pp. 258-269, 2016.
- [12] K.D. Elsbach, and I. Stigliani, "New information technology and implicit bias," *Academy of Management Perspectives*, vol. 33, no. 2, pp.185-206, 2019.
- [13] J.M. Logg, J.A. Minson, and D.A. Moore, "Algorithm appreciation: People prefer algorithmic to human judgment," *Organizational Behavior and Human Decision Processes*, vol. 151, pp.90-103, 2019.
- [14] C. Keding, and P. Meissner, "Managerial overreliance on AI-augmented decision-making processes: How the use of AI-based advisory systems shapes choice behavior in R&D investment decisions," *Technological Forecasting and Social Change*, vol. 171, pp. 120970, 2021.
- [15] A. Vishwanath, and K.H. LaVail, "The role of attributional judgments when adopted computing technology fails: a comparison of Microsoft Windows PC user perceptions of Windows and Macs," *Behaviour & Information Technology*, vol. 32, no. 11, pp.1155-1167, 2013.
- [16] A. Tversky, and D. Kahneman, "Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty," *Science*, vol. 185, no. 4157, pp.1124-1131, 1974.
- [17] S. S. Sundar, "The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility," in *Digital Media, Youth, and Credibility*, M.J. Metzger, A.J. Flanagin, Eds., Cambridge: MIT Press, 2008, pp. 73–100.
- [18] D. Lindebaum, M. Vesa, and F. Den Hond, "Insights from "The Machine Stops" to Better Understand Rational Assumptions in Algorithmic Decision Making and its Implications for Organizations," *Academy of Management Review*, vol. 45, no. 1, pp.247-263, 2020.
- [19] K. Parry, M. Cohen, and S. Bhattacharya, "Rise of the machines: A critical consideration of automated leadership decision making in organizations," *Group & Organization Management*, vol. 41, no. 5, pp.571-594, 2016.
- [20] K. Brunsson, and N. Brunsson, *Decisions: The complexities of individual and organizational decision-making*. Cheltenham, UK: Edward Elgar, 2017.
- [21] S. Kalberg, "Max Weber's types of rationality: Cornerstones for the analysis of rationalization processes in history," *American journal of sociology*, vol. 85, no. 5, pp.1145-1179, 1980.
- [22] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. (521), vol. (7553), pp. 436-444, 2015.
- [23] T. Rouleau, "What are the types of machine learning," Sama, Retrieved from <https://www.sama.com/blog/types-of-machine-learning>, 2020.
- [24] P. Choudhury, E. Starr, and R. Agarwal, "Machine learning and human capital complementarities: Experimental evidence on bias mitigation", *Strategic Management Journal*, vol. 41, no. 8, pp.1381-1411, 2020.
- [25] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt, "It's reducing a human being to a percentage' perceptions of justice in algorithmic decisions," *Proceedings of the 2018 Chi conference on human factors in computing systems*, pp. 1-14.
- [26] J. Dodge, Q. V. Liao, Y. Zhang, R. K. Bellamy, and C. Dugan, "Explaining models: an empirical study of how explanations impact fairness judgment", *Proceedings of the 24th international conference on intelligent user interfaces, IUI2019*, pp. 275-285.
- [27] M. H. Teodorescu, L. Morse Y. Awwad, and G. C. Kane, "Failures of fairness in automation require a deeper understanding of human-ML augmentation", *Management Information Systems Quarterly*, vol. 45, no. 3, 2021.
- [28] G. M. Johnson, "Algorithmic bias: on the implicit biases of social technology", *Synthese*, vol. 198, no.10, pp. 9941-9961, 2021.
- [29] J. Lallement, S. Dejean, F. Euzéby, and C. Martinez, "The interaction between reputation and information search: Evidence of information avoidance and confirmation bias," *Journal of Retailing and Consumer Services*, vol. 53, p.101787, 2020.
- [30] A. Fügener, J. Grahl, A. Gupta, and W. Ketter, "Will Humans-in-the-Loop Become Borgs? Merits and Pitfalls of Working with AI," *Management Information Systems Quarterly*, vol. 45, no. 3, 2021.
- [31] R. Golman, D. Hagmann, and G. Loewenstein, "Information avoidance," *Journal of Economic Literature*, vol. 55, no. 1, pp.96-135, 2017.
- [32] J. Dana, R.A. Weber, and J.X. Kuang, "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness," *Economic Theory*, vol. 33, no. 1, pp. 67-80, 2007.
- [33] L. Festinger, *A theory of cognitive dissonance (Vol. 2)*. Stanford, CA: Stanford university press, 1957.
- [34] H. Bahnmler, "The Intersections between Self-Deception and Inconsistency: An Examination of Bad Faith and Cognitive Dissonance Hannah Bahnmler," *Stance: An International Undergraduate Philosophy Journal*, vol. 8, pp.71-80, 2015.
- [35] A.S. Hinojosa, W.L. Gardner, H.J. Walker, C. Coglisier, and D. Gullifor, "A review of cognitive dissonance theory in management research: Opportunities for further development," *Journal of Management*, vol. 43, no. 1, pp. 170-199, 2017.
- [36] R. Bénabou, "Groupthink: Collective delusions in organizations and markets," *Review of Economic Studies*, vol. 80, no. 2, pp. 429-462, 2013.
- [37] D.M. Kahan, E. Peters, M. Wittlin, P. Slovic, L.L. Ouellette, D. Braman, and G. Mandel, "The polarizing impact of science literacy and numeracy on perceived climate change risks," *Nature Climate Change*, vol. 2, no. 10, pp. 732-735, 2012.
- [38] CERN, "Our Mission," <https://home.cern/about/who-we-are/our-mission#:~:text=Our%20mission%20is%20to%3A,for%20the%20benefit%20of%20all> (accessed May 31, 2022).
- [39] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength Natural Language Processing in Python," *Zenodo*, 2020, <https://doi.org/10.5281/zenodo.1212303>.
- [40] G. Salton, and C. Buckley, "Term-weighting approaches in automatic text retrieval", *Information Processing & Management*, vol. 24, no.5, pp. 513–523, 1988.
- [41] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval (Vol. 39)*. Cambridge: Cambridge University Press, 2008.
- [42] scikit-learn documentation. (n.d.-a). https://scikitlearn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html?highlight=vectorizer#sklearn.feature_extraction.text.TfidfVectorizer%7D
- [43] scikit-learn documentation. (n.d.-b). https://scikitlearn.org/stable/tutorial/text_analytics/working_with_text_data.html#tokenizing-text-with-scikit-learn
- [44] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [45] scikit-learn documentation. (n.d.-c). <https://scikitlearn.org/stable/modules/classes.html#module-sklearn.metrics>
- [46] scikit-learn documentation. (n.d.-d). https://scikitlearn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html?highlight=mlpclassifier#sklearn.neural_network.MLPClassifier
- [47] A. Schepman, and P. Rodway, "Initial validation of the general attitudes towards Artificial Intelligence Scale," *Computers in human behavior reports*, vol. 1, p.100014, 2020.

PUBLICATION

V

**Artificial intelligence in the quest for the end of choice: Black boxes as
Sartrean bad faith [Manuscript under revision]**

Valtonen, L.,

Academy of Management Review. Manuscript under revision

Publication reprinted with the permission of the copyright holders.

