Theoretical and practical approaches to the responsible management of artificial intelligence

5
2022

PhD Thesis

Per Rådberg Nagbøl
IT University of Copenhagen
Business IT

Supervisor
Oliver Krancher

Committee
Louise Harder Fischer
Ioanna Constantiou
Christian Janiesch

Company
Supervisor
Marius Hartmann
The Danish Business Authority

Table of Contents

Acknowledgments

I want to thank Oliver Müller for supporting me in getting this Ph.D. position, being my supervisor doing the first half of the Ph.D., and continuing the work with me throughout the Ph.D. project. I want to thank Oliver Krancher for taking over as my supervisor for the last half of the Ph.D. project. I feel very privileged to have had the two of you as supervisors. You have both been doing an excellent job and going beyond what I could have expected in supervising, collaborating, and teaching me the scientific craft. I want to thank all my colleagues at the IT University of Copenhagen.

I want to thank Carsten Ingerslev for supporting me in getting a Ph.D. position and continuing to support my research, trusting my projects, and allowing me unlimited freedom and flexibility to do my work. I want to thank my company supervisor Marius Hartmann for supporting my work in the Danish Business Authority, providing feedback on my scientific and practical work, and taking the time to advise and teach me how to work with IT in practice. I want to thank the rest of my colleagues at the Danish Business Authority for always being willing to participate in interviews, tests, workshops, and whatever I have required over the last couple of years. Your contribution has been invaluable to the thesis.

I want to thank my co-authors Aleksandre Asatiani, Pekka Malo, Esko Penttinen, Tapani Rinta-Kahila, and Antti Salovaara for a great collaboration, their understanding, and educational efforts in one of my first research projects.

I want to thank Christoph Müller-Bloch and Thomas Kude for hosting me as a guest Ph.D. Student at ESSEC and ensuring that I had a perfect stay. I want to thank the Information System research group for welcoming me and providing excellent feedback on my research presentation.

I want to thank Pedro Ferreira and Tiemo Thiess for being good friends and colleagues. You have always taken the time to support, teach, and advise me when needed.

I want to thank my friends and family for their support and understanding of my absence in their life the last couple of years.

I want to thank my mother, Lone Rådberg, and my father, Søren Nagbøl, for their endless love, support, and trust. I want to thank my wife, Sagal Rådberg Nagbøl, for being everything to me and continuously teaching me new perspectives and making me a better person.

Lastly, I want to dedicate this work to my late grandfather Tom Rådberg and my daughter Aya Rådberg Nagbøl.

Collaborative Ph.D.

This thesis is a product of a collaborative Ph.D. project between the IT University of Copenhagen and the Danish Business Authority (DBA) in an arrangement with an equal time distribution between the two organizations. The work at the IT University of Copenhagen has consisted of research and regular Ph.D. duties related to administration and teaching. The work at the DBA has been regular governmental work primarily focused on designing, building, implementing, using, and evaluating X-RAI.

Abstract English

Organizations increasingly use Artificial Intelligence (AI) to achieve their goals. However, the use of AI has led to negative side effects harming people. The work presented in this thesis focuses on harvesting the benefits of AI while preventing harm by presenting theoretical and practical approaches to the responsible management of AI. The thesis answers the research question: *How can organizations ensure responsible use of artificial intelligence?* Five papers contribute to answering this question. The first paper asks the research question: *How can an organization exploit inscrutable AI systems in a safe and socially responsible manner?* We answer this question with an exploratory case study in the Danish Business Authority. The paper provides two key contributions by introducing the concept of sociotechnical envelopment and how it enables organizations to manage the trade-off between predictive power and explainability in AI. The second paper asks the research question: *How can organizations reconcile the growing demands for explanations of how AI based algorithmic decisions are made with their desire to leverage AI to maximize business performance?* The paper is part of a double issue with the first paper, sharing a similar foundation but differentiating itself by targeting a practitioner's audience. The paper contributes by proposing a framework with six dimensions to explain the behavior of black-

box AI systems and four recommendations for explaining the behavior of black-box AI systems. The third paper asks the research question: *How do we ensure that machine learning (ML) models meet and maintain quality standards regarding interpretability and responsibility in a governmental setting?* We address this with the use of the action design research method. The paper introduces the action design research project in the Danish Business Authority and the first version of the design artifact X-RAI framework, including its four sub-frameworks. The fourth paper asks the research question: *How should procedures be designed to assess the risks associated with a new AI system?* The paper uses action design research, focuses on the first artifact of the X-RAI framework, the Artificial Intelligence Risk Assessment (AIRA) tool, and provides five design principles. The fifth paper asks the research question: *How to plan for successful evaluation of AI systems in production?* The paper uses action design research and focuses on the second artifact of the X-RAI framework, the Evaluation Plan. The paper finds five challenges in evaluating AI and prescribes five design principles to address them.

Abstract Danish

Organisationer benytter i stigende grad Kunstig Intelligens (KI) til at opnå deres mål. Brugen af KI har haft utilsigtede negative konsekvenser, der har påført mennesker skade. Arbejdet, der præsenteres i denne afhandling, fokuserer på at høste fordelene ved KI samtidig med at forhindre utilsigtede konsekvenser. Dette gøres ved at udvikle, formidle og implementere teoretiske og praktiske tilgange til en ansvarlig håndtering af KI. Afhandlingen besvarer forskningsspørgsmålet: *Hvordan kan organisationer sikre ansvarlig brug af kunstig intelligens?* Spørgsmålet bliver besvaret igennem afhandlingens fem forskningsartikler.

Den første artikel med titlen: ”Socioteknisk konvolutering af kunstig intelligens - en tilgang til organisatorisk implementering af uigennemskuelige kunstige intelligenssystemer” er publiceret i Journal of the Association for Information Systems (JAIS). Vi stiller forskningsspørgsmålet: *Hvordan kan en organisation udnytte uigennemskuelige KI-systemer på en sikker og socialt ansvarlig måde?* Forskningsspørgsmålet besvares gennem et undersøgende casestudie i Erhvervsstyrelsen. Artiklen har to hovedbidrag i form af en introduktion af begrebet socioteknisk konvolutering, og hvordan socioteknisk konvolutering gør det muligt for organisationen at balancere mellem præcision og forklarlighed i KI.

Den anden artikel med titlen: ”Udfordringer i forbindelse med at forklare Black-Box KI-systemers adfærd” er publiceret i MIS Quarterly Executive (MISQ-E). Vi stiller forskningsspørgsmålet: *Hvordan kan organisationer forene de stigende krav til forklarlighed i KI-baserede algoritmiske beslutningsprocesser med deres ønske om at udnytte KI til at maksimere forretningsværdi?* Artiklen er en del af en dobbelt udgivelse sammen med den første artikel, og de har af den grund et næsten identisk fundament. Den anden artikel adskiller sig ved at have en erhvervsorienteret målgruppe. Artiklen præsenterer et rammeværktøj med seks dimensioner til at forklare adfærden hos black box KI-systemer, samt fire anbefalinger til at forklare adfærden hos black box KI-systemer.

Den tredje artikel med titlen: X-RAI- et rammeværktøj til transparent, ansvarlig, og præcis brug af machine learning i den offentlige sektor er publiceret i EGOV-CeDEM-ePart. Vi stiller forskningsspørgsmålet: *Hvordan sikrer vi, at machine learning-modeller (ML) efterlever og opretholder kvalitetsstandarder vedrørende fortolkningsevne og ansvar i en statslig ramme?* Spørgsmålet er undersøgt ved brug af action design research. Artiklen introducerer action design research forskningsprojektet i Erhvervsstyrelsen og den første version af design artefakten X-RAI-rammeværktøjet og dets fire underrammeværktøjer. X-RAI er et akronym for Transparent (X-RAY – engelsk for røntgen), Ansvarlig (R for Responsible) og Forklarlig (X - eXplainable) Kunstig Intelligens (AI for Artificial Intelligence).

Den fjerde artikel med titlen: ”Design af et risikovurderingsværktøj til kunstige intelligenssystemer” er publiceret i International Conference on Design Science Research in Information Systems and Technology, Springer. Vi stiller forskningsspørgsmålet: *Hvordan skal procedurer designes for at kunne vurdere de risici, der er forbundet med et nyt KI-system?* Artiklen anvender action design research metoden til at besvare forskningsspørgsmålet. Fokusset er på den første artefakt i X-RAI-rammeværktøjet ved navn Kunstig Intelligens Risiko Vurderings redskabet (AIRA for Artificial Intelligence Risk Assessment). Artiklens forskningsbidrag er de fem designprincipper Multiperspektivisk ekspertvurdering, struktureret intuition, forventede konsekvenser, vurderer performance på mere end præcision, samt konvolutering af sorte bokse.

Den femte artikel med titlen: Udfordringer og praksisser i evalueringen af KI-systemer i den offentlige sektor er sendt til fagfællebedømmelse til en kommende forskningshåndbog om

offentlig forvaltning og kunstig intelligens. Vi stiller forskningsspørgsmålet *Hvordan planlægger man en vellykket evaluering af KI-systemer i produktion?* Artiklen benytter action design research og fokuserer på artefakt nummer to - Evalueringsplanen i X-RAI-rammeværktøjet. Artiklen identificerer fem typer udfordringer i forhold til evaluering af KI og præsenterer fem designprincipper til at adressere dem.

Introduction

Artificial intelligence (AI) offers many potential applications such as increasing efficiency, detecting cancer, and supporting decision making. Nevertheless, despite good intentions, AI has led to unfortunate outcomes with severe consequences for those affected. Academia and media have noted several examples of how AI systems and algorithms have harmed people through facial recognition in policing (Hill, 2020), commercial facial detection software (Buolamwini and Gebru, 2018), crime prediction (Angwin *et al.*, 2016), skin cancer detection (Lashbrook, 2018), Google searches (Allen, 2016), and online advertisements (Sweeney, 2013). These are not single cases as the National Institute of Standards and Technology in the United States of America describes: *"There is no shortage of examples where bias in some aspect of AI technology and its use has caused harm and negatively impacted lives, such as in hiring [6 sources], health care [10 sources], and criminal justice [13 sources]"* (Schwartz *et al.*, 2022, p. 1). It is imaginable that it is only the tip of the iceberg, and increases in public interests, AI systems, and tools and approaches to detect bias and discrimination will reveal more cases in the future.

Problematic biases that negatively influence individuals, organizations, and society can lead to reduced public trust in AI (Schwartz *et al.*, 2022). The importance of public trust in AI to the Danish government is evident in the National Strategy for Artificial Intelligence. Initiative 1.5 ("Transparent use of artificial intelligence by the public sector") emphasizes the need for citizens and businesses to have confidence in public authorities' use of AI to avoid weakening their general confidence in public authorities. Central to the initiative was a pilot project focused on developing and testing methods to help public authorities fulfill statutory requirements for reasonable, responsible, and transparent AI use. The pilot project was intended to contribute to developing common guidelines and methods for using AI in the Danish public sector (Government, 2019). The Danish Business Authority (DBA) was selected as the case organization. The term "responsible" is defined as accountable, liable,

and capable of distinguishing between right and wrong (Merriam-Webster.com, 2022); harmful AI is not considered responsible AI in contexts where it is wrong to cause harm.

Before this thesis, there was scientific interest in the potential of AI in information systems. Despite substantial IS risk management research, literature on AI system risk management was lacking (Moeini and Rivard, 2019). The body of literature has grown over the course of this thesis, and literature from affiliated fields has emerged and provided different perspectives on AI, such as the interpretability of machine learning (Lipton, 2018; Du, Liu and Hu, 2019), approaches to audit for bias detection in facial recognition tools and datasets (Buolamwini and Gebru, 2018), bias and fairness (Suresh and Guttag, 2020), dataset documentation (Gebru *et al.*, 2020), and machine learning model documentation (Mitchell *et al.*, 2019). To my knowledge, there remains a dearth of literature describing how to holistically manage and maintain AI systems from a longitudinal perspective and addressing continuous evaluation after the AI systems have gone live.

It is against this background that I ask the following research question for this thesis: *How can organizations ensure responsible use of artificial intelligence?*

I answer the question through two types of research designs. The first type is qualitative behavioral research analyzed with grounded theory; the second is design science research, with action design research as the chosen method. The behavioral research contributes to advancing the understanding of envelopment and provides specific measures to explain the behavior of black box AI and enhancing responsible AI conduct. The action design research approach allows addressing AI from theoretical and technological perspectives. The artifact contribution allows practitioners to use, redesign, and improve the X-RAI to benefit us all.

When starting the design of the artifact, it was difficult to find relevant theory on responsible AI within the field of information systems to ingrain in the artifact. Instead, we relied on reference disciplines such as computer science to incorporate the interpretability of machine learning (Lipton, 2017) into the artifact. The work with the behavioral research was initiated simultaneously and provided another theoretical perspective to ingrain in the design artifact AIRA according to the methodological guidelines of action design research (Sein *et al.*, 2011). The behavioral papers describe best practices of responsible AI conduct in the DBA. These experiences are then ingrained into the artifacts to maintain equivalent or higher

standards for future AI systems. In other words, it was necessary to develop the theoretical foundation to design the artifact for the organization. The first three papers of this thesis inform the last two papers.

Table 1 Thesis overview

#	Title	Authors	Research Question	Type	Publication status
1	Sociotechnical Envelopment of Artificial Intelligence: An Approach to Organizational Deployment of Inscrutable Artificial Intelligence Systems	Aleksandre Asatiani, Pekka Malo, Per Rådberg Nagbøl, Esko Penttinen, Tapani Rinta-Kahila, and Antti Salovaara	*How can an organization exploit inscrutable AI systems in a safe and socially responsible manner?*	Exploratory Case Study	Published in Journal of the Association for Information Systems, Volume 22, Issue 2, Special Section: Artificial Intelligence in IS Research
2	Challenges of Explaining the Behavior of Black-Box AI Systems	Aleksandre Asatiani, Pekka Malo, Per Rådberg Nagbøl, Esko Penttinen, Tapani Rinta-Kahila, and Antti Salovaara	*How can organizations reconcile the growing demands for explanations of how AI based algorithmic decisions are made with their desire to leverage AI to maximize business performance?*	Exploratory Case Study	Published in MIS Quarterly Executive, Volume 19, Issue 4.
3	X-RAI: A Framework for the Transparent, Responsible, and Accurate Use of Machine Learning in the Public Sector	Per Rådberg Nagbøl and Oliver Müller	*How do we ensure that machine learning (ML) models meet and maintain quality standards regarding interpretability and responsibility in a governmental setting?*	Action Design Research	Published in EGOV-CeDEM-ePart

4	Designing a Risk Assessment Tool for Artificial Intelligence Systems	Per Rådberg Nagbøl, Oliver Müller, and Oliver Krancher	*How should procedures be designed to assess the risks associated with a new AI system?*	Action Design Research	Published in *International Conference on Design Science Research in Information Systems and Technology*. Springer, Cham Paper Awarded: Vinton G. Cerf Award For The Best Student Authored Paper Of The Conference
5	Challenges and Practices in the Evaluation of AI Systems in the Public Sector	Per Rådberg Nagbøl, Oliver Krancher, and Oliver Müller	*How to plan for successful evaluation of AI systems in production?*	Action Design Research	Submitted for review to the forthcoming Research Handbook on Public Management and Artificial Intelligence

Paper 1 (see table 1) presents a sociotechnical envelopment-based approach to responsibly implementing inscrutable AI systems in an organizational context. My co-authors and I asked the following research question: *How can an organization exploit inscrutable AI systems in a safe and socially responsible manner?* We investigate this in the context of the DBA's envelopment theory addressing an AI system's boundary, training data, input, output, and function (Robbins, 2020) from a sociotechnical perspective balancing attention toward both instrumental and humanistic outcomes (Sarker *et al.*, 2019).

Paper 2 (see table 1) addresses challenges related to explaining the behavior of black box AI systems. While this is similar to paper 1, it is differentiated by targeting a practitioner audience to answer the research question: *How can organizations reconcile the growing demands for explanations of how AI based algorithmic decisions are made with their desire to leverage AI to maximize business performance?* We are answering this question by

defining the six elements of a hypothetical AI agent: the model, goals, training data, input data, output data, and environment. After that, we propose a framework that explains the behavior of black box AI systems with six dimensions that correspond to the six elements. We hereafter describe how the DBA has addressed these dimensions before concluding the paper with four recommendations (based on the case study) for explaining the behavior of black box AI systems.

While papers 1 and 2 are based on behavioral research, paper 3 (see table 1) is a research-in-progress paper aimed at developing an artifact. The paper describes the early design, theoretical foundation, and vision for the X-RAI framework, along with four sub-frameworks. The paper asks the research question: *How do we ensure that machine learning (ML) models meet and maintain quality standards regarding interpretability and responsibility in a governmental setting?* The paper addresses this question by using an action design research approach with theory on the interpretability of machine learning (Lipton, 2017) ingrained (Sein *et al.*, 2011). The X-RAI framework, the designed artifact, is an acronym for transparency (X-ray), responsible (R), and explainable (X) AI, with four sub-frameworks: the Model Impact and Clarification Framework, Evaluation Plan Framework, Evaluation Support Framework, and Retraining Execution Framework.

While paper 3 presents an overview of the X-RAI framework and the overall project, paper 4 (see table 1) focuses on the first X-RAI sub-framework, the risk assessment of artificial intelligence systems before going live. To support organizations in assessing and mitigating risk while harvesting AI benefits, we ask the research question: *How should procedures be designed to assess the risks associated with a new AI system?* We used the action design research method (Sein *et al.*, 2011) to answer the question by building, intervening, and evaluating the artifact Artificial Intelligence Risk Assessment (AIRA) tool in the DBA. We drew on risk management literature (Moeini and Rivard, 2019) and AI-specific literature addressing interpretability (Lipton, 2017), envelopment (Robbins, 2020; Asatiani *et al.*, 2021), and model documentation (Mitchell *et al.*, 2019; Gebru *et al.*, 2020).

Paper 5 (see table 1) focuses on the second artifact of the X-RAI framework described in paper 3 and supplements paper 4. The paper focuses on challenges and solutions related to planning and conducting post-go-live evaluations for AI systems. The paper asks the research question, *How does one plan for successful evaluation of AI systems in production?* We

answered this question with action design research to build, intervene, and evaluate the designed artifact: the Evaluation Plan in the DBA context combined with follow-up stakeholder interviews. We drew on literature describing how to involve domain and AI experts (Doshi-Velez and Kim, 2017; Lebovitz, Levina and Lifshitz-Assaf, 2021; Lou and Wu, 2021; Nagbøl, Müller and Krancher, 2021; van den Broek, Sergeeva and Huysman, 2021), representation theory (Recker *et al.*, 2019), self-determination theory (Ryan and Deci, 2000), and control theory (Eisenhardt, 1985).

Background

This part of the thesis contains a short introduction to the theoretical and technological foundation I have relied on in my work with responsible AI. The chapter starts with a brief introduction to the definitions of AI used throughout the thesis and then introduces key technological and theoretical perspectives that inform this thesis's work with responsible AI.

Definitions of AI

It would not be provocative to say AI has no universally agreed-upon definition. In their MISQ editorial "Managing Artificial Intelligence," Berente et al. describe *"AI as the frontier of computational advancements that references human intelligence in addressing ever more complex decision-making problems"* (Berente *et al.*, 2021, p. 5). This definition suggests that AI is a dynamic concept, where the semantic reference point remains static while the referenced meaning is subject to continuous negotiation and change. The means we use to manage AI must follow these changes.

The papers in the thesis rely on different definitions that share common elements. Paper 1 (Asatiani *et al.*, 2021) relies on Kaplan and Haenlein's definition of AI as a "system's ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals" (Kaplan and Haenlein, 2019, p. 17).

Paper 2 (Asatiani *et al.*, 2020) bases its definition on by Russel and Norvig's (2010) interpretation of AI: *"an intelligent agent, whether human or machine, pursues goals by processing data and interacting with other agents in the environment"* (Asatiani *et al.*, 2020, p. 261).

Paper 4 (Nagbøl, Müller and Krancher, 2021) uses the European Commission definition: "systems that display intelligent behavior by analyzing their environment and taking actions—with some degree of autonomy—to achieve specific goals" (Commission, 2018). The definitions are similar in that they require the AI system to analyze its context to achieve its goals.

Responsible AI

The thesis relies on an in-practice contextual interpretation of responsible. The Merriam-Webster dictionary defines responsible*: ": liable to be called on to answer," ": liable to be called to account as the primary cause, motive, or agent,"* "*: being the cause or explanation,"* "*: liable to legal review or in case of fault to penalties*," *": able to answer for one's conduct and obligations: TRUSTWORTHY," ": able to choose for oneself between right and wrong*," *": marked by or involving responsibility or accountability,"* and *":* politically answerable"* (Merriam-Webster.com, 2022). These definitions ascribe different attributes to being responsible.

Responsible AI assigns accountability for actions caused by the AI system. Those accountable can answer for the AI's conduct and obligations, choosing between right and wrong applications of AI, answering politically for the AI system, and being held responsible for the actions. The AI system can, to a context-dependent degree, choose between right and wrong (in the accuracy of predictions) and is accessible to reviews.

The definitions support the rationalities behind this work since they aim not to decide what is right or wrong and who should be responsible but to provide the perspective and tools to make an informed decision by involving stakeholders with relevant expertise for the given context, decisions, and AI systems. Harmful AI is not responsible AI in contexts where it is wrong to cause harm, and it must be possible to assign the responsibility for the harm.

This thesis builds on the assumption that the meaning of responsible AI will be subject to continuous change and will be negotiated and renegotiated over time. The responsible of today might be irresponsible tomorrow, while the responsible for context x might be irresponsible for context y and vice versa. Specifying what and who is responsible concerning

AI would limit the generalizability of the work in context and time. The definition in this thesis does not define what it means to be responsible in a legal sense.

Technical foundations

The Canadian Algorithmic Impact Assessment tool (Secretariat, Treasury Board of Canada, 2020) was analyzed (in 2019) to clarify if the tool could be adapted to a Danish context. The tool was available online on GitHub as open source (Government of Canada, 2020), allowing for an analysis of code and functionality. The tool is a questionnaire with weights added to the answers, allowing the calculation of the raw impact score and the mitigation score into the current score, which provides placement in one of four categories of the Canadian directive on automated decision-making. This allows the user to self-assess their compliance with the requirements (Secretariat, Treasury Board of Canada, 2019). It was decided not to adapt the Canadian Algorithmic Impact Assessment tool to a Danish context but instead let us be inspired in developing the early version of the AIRA tool (Nagbøl, Müller and Krancher, 2021).

In recent year, the European Commission published a draft of the forthcoming Artificial Intelligence Act (AIA). The AIA divides AI into three categories: “normal” risk, high-risk, and prohibited AI. An AI system classified as high-risk is subject to a range of requirements such as those described in chapter 2: article 9 Risk management system, article 10 Data and data governance, article 11 Technical documentation, article 12 Record keeper, article 13 Transparency and provision of information to users, article 14 Human oversight, and article 15 Accuracy, robustness and cybersecurity (European Commission, 2021).

The AIA will have a significant impact on the use and management of AI in European organizations. It is, therefore, important that the tools and approaches developed to support the use and management of AI within the European Union, at minimum, do not conflict with the legislation. It would increase the benefit of approaches like X-RAI to support legal compliance with the AIA with inspiration from the Canadian work in developing the algorithmic Impact Assessment tool to support compliance with the directive on automated decision-making.

Theoretical foundations

The theoretical foundation for the thesis consists of a wide range of complementary theories that individually contribute different perspectives on AI.

Risk management and AI

The topic of risk management in information systems is well researched, providing a rich body of literature to inform our work with AI. Risk is a context-dependent word, but *"[t]he most common definition of risk in software projects is in terms of exposure to specific factors that present a threat to achieving the expected outcomes of a project."* (Bannerman, 2008, p. 2119). Bannerman (2008) warns that research's conceptualization of risk is narrower than the nature of practical problems requires. Moeini and Rivard (2019) present a comprehensive literature review, dividing risk management into two bodies of knowledge. When it comes to risk assessment, the normative body of knowledge assumes that deliberate analysis usually outperforms intuition, while the experiential body of knowledge assumes that intuition usually outperforms deliberate analysis. The two bodies of knowledge are bridgeable when understood as complementary instead of diverging, allowing managers to apply the best approach for their situation (Moeini and Rivard, 2019). Research has found that software practitioners identify more risks when using a checklist than they would without one, but they are also likely to identify risks that are not present. So, while the checklist supports risk identification, it simultaneously presents the practitioners with risks they would not have seen without the checklist and that are not present in the given scenario. The use of checklists might cause the practitioners to approach risk identification with less thought (Keil *et al.*, 2008).

Risk management is a process dividable into two categories: risk assessment, with the subcategories of risk identification, risk analysis, and risk prioritization, and risk control, with the subcategories of risk management planning, risk resolution, and risk monitoring (Boehm, 1991). Boehm points out that risk management involves a lot of human judgment. *"Good people, with good skills and good judgment, are what make projects work. Risk management can provide you with some of the skills, an emphasis on getting good people, and a good conceptual framework for sharpening your judgement"* (Boehm, 1991, p. 41).

Bannerman has found that risk management literature does not meet the needs of the practice, and risk management in practice lacks the knowledge and prescriptions from the literature (Bannerman, 2008). The learning capacity and data use in AI systems introduces new needs and risks for risk management, requiring new identification and mitigative approaches to opaqueness (Lipton, 2017), countermeasures to the lack of dataset and model documentation (Mitchell *et al.*, 2019; Gebru *et al.*, 2020), operational space (Robbins, 2020), and awareness of bias, fairness, and unintended consequences (Suresh and Guttag, 2020) of AI. This thesis relies on reference literature from other disciplines to address these issues.

This thesis relies on research by Bannerman (2008) that defines risk in software projects, acknowledging that the research conceptualization may be too narrow to fit the requirements of practical problems while arguing that AI software introduces new ones. The balancing act between intuition and deliberate analysis (Moeini and Rivard, 2019) structures the design approach.

Bias, fairness, and unintended consequences

Suresh and Guttag (2020) provide a framework for understanding bias and unintended consequences in machine learning. They describe how six different kinds of bias – historical, representation, measurement, aggregation, evaluation, and deployment bias – become sources of harm (Suresh and Guttag, 2020). They introduce different approaches to fairness and conclude that it is not one-size-fits-all. They suggest prioritizing knowledge-based application-appropriate solutions informed by stakeholder engagement instead of relying on general concepts defining fairness (Suresh and Guttag, 2020). Understanding bias and fairness helps identify wrongful conduct and its causes.

Interpretability

AI systems can become black-boxed, leaving users in the dark about the inner workings between input and output. Interpretability has no agreed-upon definition but covers ideas seeking to explain the workings of machine learning models, including black-boxed systems (Lipton, 2017). Common metrics for supervised machine learning combine predictions with ground truth to produce a score. Evaluation metrics such as Receiver Operating Characteristic Area Under the Curve (ROC AUC) (Spackman, 1989; Fawcett, 2006) and accuracy provide low assurance of acceptable behavior related to discrimination based on race, and the

demands for fairness lead to demands for interpretable models (Lipton, 2018). The conceptual and technological foundation of interpretability of machine learning has been divided by scholars into two subcategories, specifically, transparency and post hoc interpretability (Lipton, 2017) or intrinsic interpretability and post-hoc interpretability (Du, Liu and Hu, 2019). Lipton addresses transparency according to three different levels of simulatability for the entire model requiring human computability, decomposability for components such as input, parameters, and calculation, restricting the use of opaque components like overly engineered or anonymous features, and algorithmic transparency for the training/learning algorithm (2017). Post hoc interpretability consists of approaches such as text explanations, visualizations, explanation by example, and local explanations to derive insights from trained models without describing the exact workings of the model (Lipton, 2017). While global interpretability or explanation provides an overall insight into the system's function, provides local interpretability an explanation for an individual prediction or decision (Weller, 2019). Techniques such as Local Interpretable Model-agnostic Explanations (LIME) for local explanations (Ribeiro, Singh and Guestrin, 2016) and frameworks like SHAP (SHapley Additive exPlanations) for feature importance (Lundberg and Lee, 2017) provide valuable insights.

Stakeholders and tasks must be considered when designing and evaluating the AI system and its explanations. Doshi-Velez and Kim (2017) suggest a three-level taxonomy of interpretability evaluation in machine learning consisting of application-grounded model evaluation according to the task. Hence, a machine learning model supporting doctors in diagnosing patients should be compared to doctors diagnosing patients. The baseline is the domain expert's explanation. Human-grounded metrics refers to humans doing simplified tasks. Laypeople are useable in this approach instead of domain experts to test the quality of the explanation without necessarily having a specified goal. Finally, functionally grounded evaluation uses formal definitions of interpretability instead of humans as a proxy for the explanation quality. The focus here can be on improving performance with interpretable models such as decision trees, testing immature methods, or unethical human subject experiments (Doshi-Velez and Kim, 2017).

The literature on the interpretability of machine learning aids the understanding of the workings of AI systems. A deeper understanding of how the AI system works supports our

capability to satisfy organizational requirements and generally identify risk and discover harm.

Dataset and model documentation

Data is vital for AI systems' performance. When deciding if an AI system is usable, it is important to understand both the data it is trained on and its attributes. Datasheets for datasets draw inspiration from the electronic industry, where datasheets containing relevant information accompany the components. The purposes of datasheets for datasets are to document datasets, improve communication between the two key stakeholder groups in the form of dataset creators and consumers, and enhance responsible conduct. Datasheets for datasets aim to enhance dataset creators' reflection on creating, distributing, and maintaining datasets and support the consumers in making informed decisions. Datasheets for datasets contain the categories of motivation, composition, collection process, preprocessing/cleaning/labeling, uses distribution, and maintenance (Gebru *et al.*, 2020). Model cards for model reporting complement datasheets for datasets by providing documentation accompanying machine learning models covering model details, intended use, factors, metrics, evaluation data, training data, quantitative analyses, ethical considerations, and caveats and recommendations. The purpose is to be able to compare models beyond traditional evaluation metrics. Stakeholders include machine learning and AI practitioners, model developers, software developers, policymakers, organizations, machine-learning-knowledgeable individuals, and impacted individuals (Mitchell *et al.*, 2019).

Envelopment

Envelopment is an approach to advance responsible use of artificial intelligence (Robbins, 2020). The term envelopment originates from a physical context: *"In robotics, an envelope is the three-dimensional space that defines the boundaries that a robot can reach"* (Floridi, 2011a, p. 228). The envelope supports the machine in achieving its purpose in a confined and safe environment. For example, you can choose a dishwasher over a humanoid washing dishes (Floridi, 2011b; Robbins, 2020). Robbins conceptually transfers envelopment from the physical to the virtual AI context by defining the five properties of *"training data, inputs, functions, output, and boundaries"* (Robbins, 2020, p. 1) necessary to constrain the AI system and allow it to fulfill its purpose while preventing harm (2020). Envelopment provides a responsible supplement to "opening" up the black-box of AI allowing the use of

beneficial black box AI systems that contribute to the good of society (Robbins, 2020). Envelopment is not a risk-free approach and has moved from being a stand-alone phenomenon such as dishwashers or placed in industrial buildings to our everyday life (Floridi, 2011b). Floridi indicates that we have, without realizing it, enveloped the world for decades, potentially shaping it physically and conceptually to an extent where humans must adjust to fit (Floridi, 2011a). This leads to the question of whether the world should be shaped for humans, machines or something else.

Utilizing stakeholder expertise

The use of AI systems introduces new forms of engagement and collaboration in organizations on strategic and practical levels. Strategic: Articulating an organization's AI orientation differs from conventional IT orientation by increased board involvement in forming the orientation (Li *et al.*, 2021). Practical: Working with AI in knowledge work is a sociotechnical endeavor requiring both domain and AI expertise. Van den Broek et al. (2021) describe the collaboration between domain experts and ML developers as important for developing an AI hiring system, where the human-ML hybrid practice emerged from combining expertise in an interdependent relationship. The ML developers delivered unknown insights from the data to the domain experts in a symbiotic practice while the domain experts defined, evaluated, and complemented the input and output of the machine to benefit the ML developers (van den Broek, Sergeeva and Huysman, 2021). Lou and Wu made a similar discovery about the importance of combining domain and AI expertise in an iterative and ongoing collaboration to develop and use AI tools for drug development (Lou and Wu, 2021). Lebovitz et al. (2021) provide a different perspective on the relationship between domain experts and AI systems by pointing to the tension between how AI systems are evaluated based on the "know-what" (ROC AUC and ground truth) measures while domain experts evaluate their work according to "know-how". Lebovitz et al. advise against considering the ground truth objective when the underlying knowledge is tied to uncertainty; they suggest that a human should be responsible for making the final decision in cases of high uncertainty. Areas with established knowledge claims should rely on practical performance standards and domain expert know-how when choosing quality measures for training and validation (2021). The term "Borg" originates from the term "cyborg" and describes human behavior in which AI use has led to increased performance but decreased individual unique knowledge for the humans. The loss of individual unique human knowledge results in Borg (cyborg-like) behavior because the human begins to mirror the decisions of other humans and

AIs, potentially leading to a scenario where an algorithm's quality defines the quality of the work, hence causing a lack of originality (Fügener *et al.*, 2021).

Methodology

This section describes the research paradigm and summarizes the thesis's methodologies. Papers 1 and 2 contain behavioral research based on an exploratory case study. They share a similar foundation since they were part of the double issue of the Journal of the Association of Information Systems (Benbya, Pachidi, and Jarvenpaa, 2021) and the MISQ Executive (Benbya, Davenport and Pachidi, 2020). Papers 3, 4, and 5 resulted from the same action design research (Sein *et al.*, 2011) project but focused on different artifact elements.

Research paradigm

This thesis takes a pragmatist stance. According to Almeder (2014), pragmatism is a philosophical movement offering different solutions to epistemological and logical problems of the natural sciences, where *"[p]ragmatists believe that the rational justification of scientific beliefs ultimately depends on whether the method generating the beliefs is the best available for advancing our cognitive goals of explanation and precise prediction"* (Almeder, 2014, p. 103). It has been argued that design science adheres to the pragmatic belief in consequences and effects as essential components of meaning and truth through its contributions to the application environment (Hevner, 2007). Purao et al. advocate for the complementarity and benefit of combining action research with design research (science) (2010). They argue for the possibility of placing both design research and action research within pragmatism based on similarities in "…*the ontology to which both research approaches subscribe assumes that the phenomenon of interest does not remain static through the application of the research process"* (Purao, Rossi and Sein, 2010, pp. 189–190), *"… the epistemology that both research approaches subscribe to assumes a mode of knowing that involves intervening to effect change and reflecting on this intervention"* (Purao, Rossi and Sein, 2010, p. 190), and *"… the axiology that both subscribe to is evident in the manner in which both value the relevance of the research problem and emphasis on practical utility and theoretical knowledge simultaneously"* (Purao, Rossi and Sein, 2010, p. 190).

Behavioral research

Paper 1 (Asatiani *et al.*, 2021), "Sociotechnical Envelopment of Artificial Intelligence: An Approach to Organizational Deployment of Inscrutable Artificial Intelligence Systems", and Paper 2 (Asatiani *et al.*, 2020), "Challenges of Explaining the Behavior of Black-Box AI Systems", share a nearly identical empirical foundation from being a part of the special double issue on AI in the Journal of the Association for Information Systems (Benbya, Pachidi and Jarvenpaa, 2021) and the MIS Quarterly Executive (Benbya, Davenport and Pachidi, 2020) about disseminating the scientific outcome to both academia and practitioners.

We based the empirical foundation on an exploratory case study borrowing data analysis methods from grounded theory in the DBA, with interviews and observations as the primary data source. The data collection and analysis were conducted in a four-stage iterative process with overlapping phases, with the earlier stages informing the following stages. The first phase was explorative, aiming at establishing collaboration and gaining insight into the current and future ML projects and visions from both a casework and data science perspective. The second phase focused on achieving a thorough understanding of the various ML projects from the involved actors of the DBA. We interviewed all ML Lab members and two caseworkers in the third phase. The fourth and final phase focused on validating interpretations and gaining deeper insights into technical infrastructure. We conducted semi-structured interviews from August 2018 to October 2020 (see paper 1, page 333, table 1) and in January 2020 (see paper 2, page 276, table), transcribed them into 167,006 and 153,195 words, and supplemented with participant observations and document analysis. Alongside task descriptions and meeting notes, the observations were documented in a handwritten field diary dating back to 2017. The document analysis was conducted on documentation and user stories from a project management system and material on the Git repository. Conversations on a personal email at the organization informed the work when needed. Furthermore, we conducted an assessment exercise where the informants of the ML lab filled out an input-ML-model-out framework.

The analysis approach was abductive in nature (by first being inductive) and was later informed by a theoretical lens functioning that emerged as an appropriate sensitizing device (Tavory and Timmermans, 2014; Sarker *et al.*, 2018). We analyzed the interviews in three stages using coding and analysis techniques from a less procedure-oriented version of grounded theory (Charmaz, 2006; Belgrave and Seide, 2019). The three coding stages produced concepts (first-order constructs), themes (second-order constructs), and aggregated

dimensions. The first stage was open coding, in which the codes were entirely grounded in the data. The second stage focused on emerging themes, and the third stage was theoretical coding, with envelopment (Robbins, 2020) as a sensitizing lens.

Action design research

Action design research (ADR) combines action research and design research (Sein *et al.*, 2011). They are both proactive approaches to research in which learning occurs through intervention and problem solving (Purao, Rossi and Sein, 2010). This presents a methodological fit for this collaborative Ph.D. project in which I, as a researcher, spent my time approximately equally divided between the IT University of Copenhagen and the DBA. The intervention consisted of designing the X-RAI artifacts (see figure 1) and participating in everyday work life in the DBA.

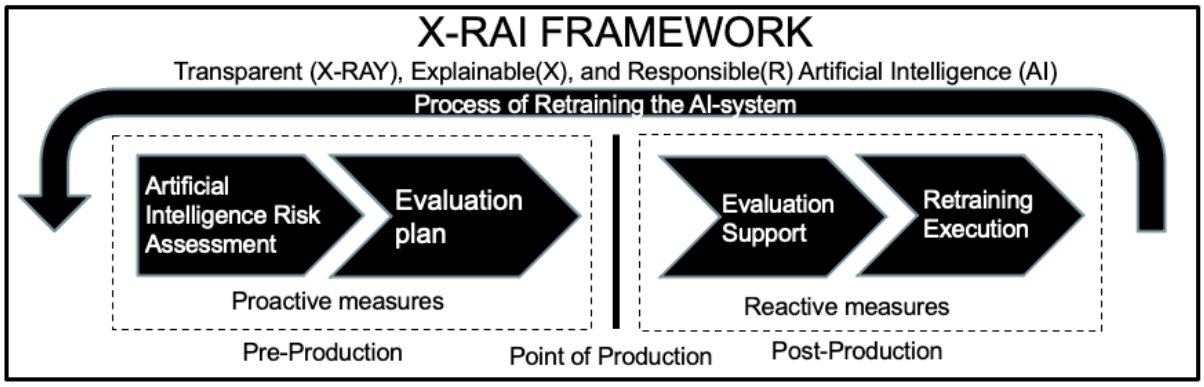


Figure 1 X-RAI Framework, with the four subframeworks adapted and revised from Nagbøl and Müller (2020).

ADR is a four-stage approach with iterations between the first three stages before finalizing in the fourth stage (Sein *et al.*, 2011), as shown in figure 2.

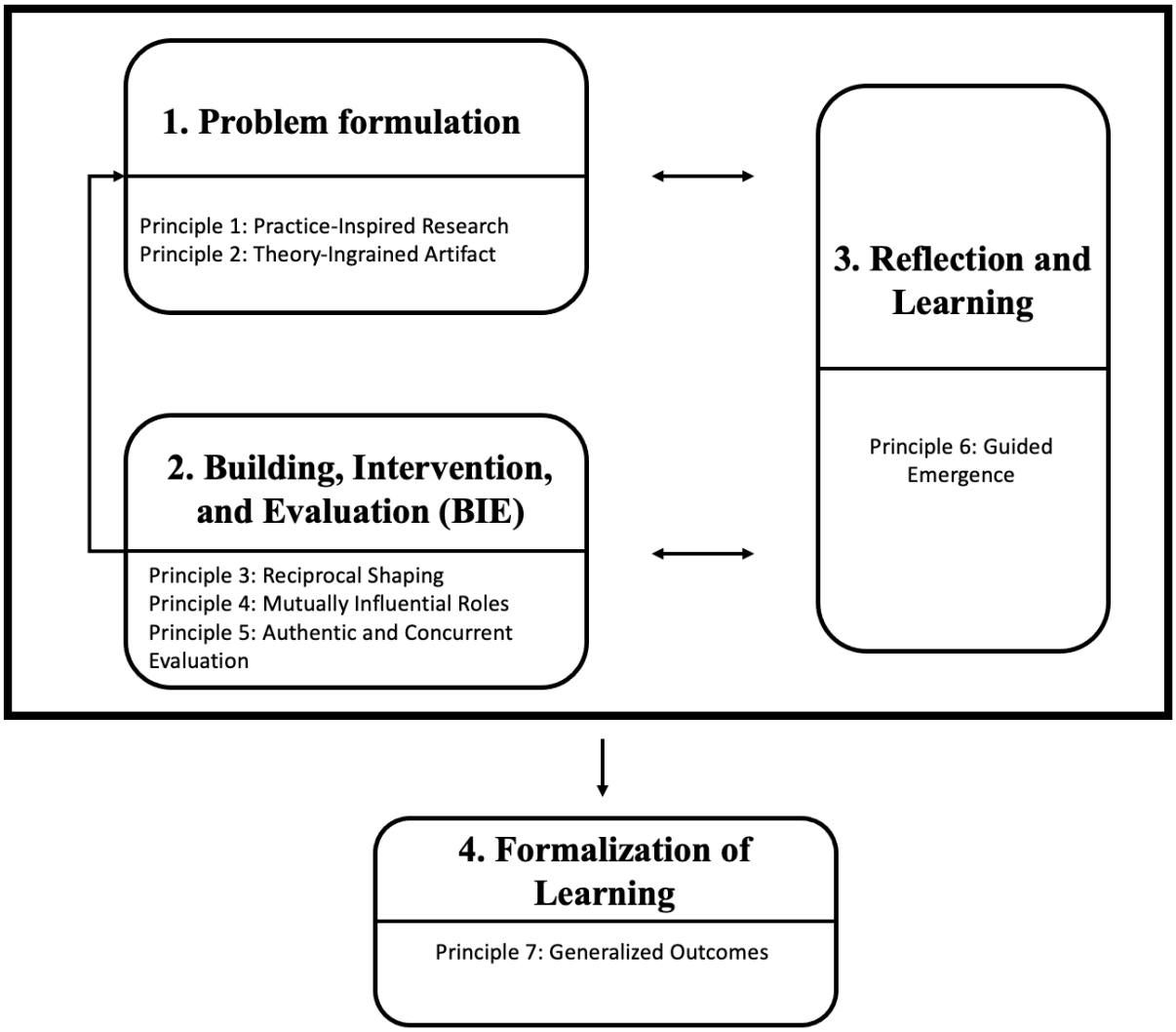


Figure 2 ADR Method: Stages and Principles adapted from Sein et al. (2011)

Stage 1, problem formulation, starts by engaging with a practical problem and scoping the project. Principle 1, practice-inspired research, focuses on turning practical, non-unique problems into knowledge creation. Treating problems as representations of problems existing elsewhere allows for the generation of knowledge and the development of solutions applicable for solving a similar class of problems (Sein *et al.*, 2011). The identification of potential problems related to the use of AI systems in the DBA started the development of an early version of X-RAI with only three sub-frameworks to solve problems. Principle 2, theory-ingrained artifact, accentuates that the theory must be ingrained into artifact through at least one of the three overlapping approaches: structuring the problems, identifying solutions, and guiding the design before the organizational exposure (Sein *et al.*, 2011). We did not identify one theory that could address all the identified problems related to planning evaluations, evaluating, and retraining the AI systems; instead, we decided to rely on an

ensemble of different theories. The fourth framework was identified through technological inspiration from the Canadian government to address an AI system's suitability to a given context (Nagbøl and Müller, 2020; Nagbøl, Müller and Krancher, 2021).

Stage 2, building, intervention, and evaluation (BIE), continues the first stage's work in an iterative process building the IT artifact, intervening in the organization, and continuously evaluating both problem and artifact, leading to the artifact's design. Principle 3, reciprocal shaping, focuses on the mutual influence of the IT artifact and organizational context. Principle 4, mutual influential roles, emphasizes the necessity of mutual learning between the participants in the design project. Principle 5, authentic and concurrent evaluation, integrates decisions regarding the design, shaping, and reshaping of the artifact and organizational intervention into the authentic and ongoing evaluation (Sein *et al.*, 2011).

Paper 3 (Nagbøl and Müller, 2020) was a research-in-progress paper introducing the action design research project in the DBA and describing the early status, design, theoretical foundation, and visions of the X-RAI artifacts. The first artifact, the Model Impact Clarification Framework, was applied and tested on four AI systems three times in its first version and once in its second version. The second artifact, the Evaluation Plan Framework, was applied and tested on eight AI systems in three incrementally different versions. The third artifact, the Evaluation Support Framework, was applied and tested five times on three different AI systems in three different versions. Finally, the fourth artifact, the Retraining Framework, was applied and tested twice on two different AI systems in two incrementally changed versions.

Paper 4 went in-depth on the first artifact, the Model Impact Clarification framework (now AIRA; (Nagbøl, Müller and Krancher, 2021)). We designed the AIRA tool to assess the risk associated with a new AI system. Between April 2019 and March 2021, we developed the AIRA tool in three iterations (see page 332, table 1 from (Nagbøl, Müller and Krancher, 2021)): building, evaluating, and testing. The first author spent this period every other workday at the organization providing a solid empirical foundation using transcripts, field notes, documents, and artifacts produced by engaging in everyday interactions and meetings with employees at the DBA. Around 30 meetings, including 12 one-on-one sessions with the team leader of the ML lab, shaped the design of AIRA.

Paper 5 went more in-depth with the second artifact, the Evaluation Plan. The evaluations from ADR research were supplemented with field notes, IT system documentation, and seven qualitative semi-structured interviews. The interviews were then coded into challenges and solutions. The solutions were further developed into design principles (Nagbøl, Krancher and Müller, Submitted).

Stage 3: Reflection and learning continue simultaneously with stages 1 and 2. The research perspective goes beyond solving a problem by supporting conceptual change, from solving the problem to applying the learnings to a class of problems. The identification of knowledge contributions occurs through reflections on the problem scope, selected theory, and the emerging ensemble artifact. Principle 6: Guided emergence acknowledges that the shaping of the designed artifact will not be limited to the scientist but will likewise occur through organizational use, perspective, participants, authentic outcomes, and concurrent evaluation (Sein *et al.*, 2011).

Stage 4: The formalization of learning facilitates the development of the ADR projects' situated learnings into general solutions. The research makes a conceptual move from providing a solution for a single problem to doing so for a whole class of field problems. The formalization of learning occurs by describing the artifact's achievements and the organization outcomes in design principles and refinements to the ingrained theories (Sein *et al.*, 2011). Principle 7: Generalized outcomes focus on moving from "specific-and-unique" to "generic-and-abstract", and the solution and problem are both generalizable. Sein et al. *"...suggest three levels for this conceptual move: (1) generalization of the problem instance, (2) generalization of the solution instance, and (3) derivation of design principles from the design research outcomes"* (Sein *et al.*, 2011, p. 44). Paper 3 did not formulate any design principles as a research-in-progress paper to introduce the design artifact X-RAI. The design principles of papers 4 and 5 were articulated according to the guidelines of Gregor et al. (2020).

The X-RAI artifact

This part of the thesis contains the X-RAI framework with the four sub-frameworks (see figure 3). X-RAI is here in a framework format. The intention is to make a tool edition available on GitHub in the future.

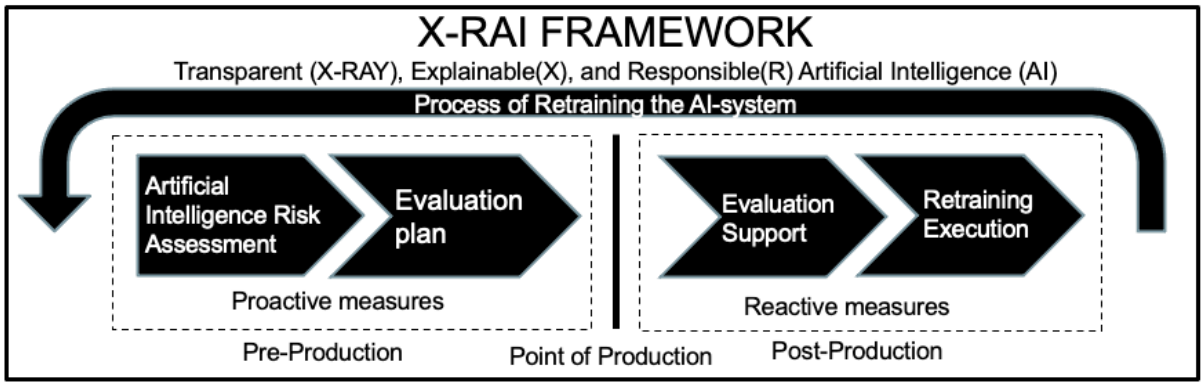


Figure 3 X-RAI Framework, adapted and revised from Nagbøl and Müller (2020).

Artificial Intelligence Risk Assessment (AIRA) framework

This part of the thesis contains the AIRA framework. AIRA consists of three parts in the form of the business part (see table 2), the data scientist part (see table 3), and the legal part (see table 4). For description of AIRA, see Nagbøl et al. (2021), for the early version, see Nagbøl and Müller (2020). For AIRA´s five design principles, see Nagbøl et al. (2021).

Table 2 AIRA: business part

Q#	Question	Answer
bus.1.0	What is the business need that the model should support?	
bus.1.1	What is the use-case/user story for the model?	
bus.1.2	What are the expected effects of the model?	
bus.1.3	Who are the anticipated users of the model?	
bus.1.4	What are the premises for the data application?	
bus.1.5	Which IT system should use the model?	
bus.1.6	Who is responsible for the quality of the model's decisions?	
bus.1.7	What is the user story for the model's explainability?	

bus.1.8	Are there any comments on the above?	
bus.2.0	What is the consequence of a True Positive classification?	
bus.2.1	Is there a person (human-in-the-loop) who sees the model's True Positive classification?	
bus.2.2	What is the consequence of a False Positive classification?	
bus.2.3	Is there a person (human-in-the-loop) who sees the model's False Positive classification?	
bus.2.4	What is the consequence of a True Negative classification?	
bus.2.5	Is there a person (human-in-the-loop) who sees the model's True Negative classification?	
bus.2.6	What is the consequence of a False Negative classification?	
bus.2.7	Is there a person (human-in-the-loop) who sees the model's False Negative classification?	
bus.2.8	Is it decided how classifications without a human-in-the-loop can be systematically quality assured?	
bus.2.9	Is it possible for the user to instantly validate or reject the truthfulness of the model's classification (output)?	
bus.3.0	Are there comments about the user's possibility of instantly validating or rejecting the truthfulness of the model's classification (output)?	

Table 3 AIRA: Data Scientist Part

Q#	Question	Answer
dsc.1.0	What is the name of the model?	
dsc.1.1	What version of the model is being described?	
dsc.1.2	What is the purpose of the model?	
dsc.1.3	What is the output of the model?	
dsc.1.4	what is the purpose of the algorithms?	
dsc.1.5	What algorithms are used?	
dsc.1.6	Which libraries are used?	
dsc.1.7	What output do the algorithms deliver?	
dsc.1.8	Is unsupervised machine learning used?	
dsc.1.9		
dsc.1.9.1	Specify which performance metrics the model is optimized for	
dsc.1.9.2	Performance in numbers:	
dsc.1.9.3	Comments on performance optimization (eg, for multiclass classifications)	
dsc.2.2		

dsc.2.2.1	Write the numbers of the used performance metrics for the model.	
dsc.2.2.2	Comments on performance optimization (eg, for multiclass classifications)	
dsc.2.4		
dsc.2.4.1	What is the threshold set at	
dsc.2.4.2	True positive *(count/%)*	
dsc.2.4.3	False Positive *(count/%)*	
dsc.2.4.4	False Negative *(count/%)*	
dsc.2.4.5	True Negative *(count/%)*	
dsc.2.5	How does the algorithm discover rules?	
dsc.2.6	Do you understand the discovered rules? If yes, describe:	
dsc.2.7	What methods (if any) do we use to understand how features influence the model's predictions/classifications (Global Explanations)? And what insight is derived from it?	
dsc.2.8	What methods (if any) do we use to explain how the model has arrived at a single prediction/classification (local explanation)? And what insight is derived from it?	
dsc.2.9	Who is the insight/explanation in/of the models working directed towards?"	
dsc.3.0	Are there different stakeholders who need insight/explanation?	
dsc.3.1	Specify stakeholders:	
dsc.3.2	Do these stakeholders have different needs? If yes, specify the needs:	
dsc.3.3	Are external data sources used? If yes, which:	
dsc.3.4	Are internal data sources used? If yes, which:	
dsc.3.5	Which file formats are used?	
dsc.3.6	What is the number of observations?	
dsc.3.7	What is the number of features in the model?	
dsc.3.8		
dsc.3.8.1	How is the data distribution in the training data?	
dsc.3.8.2	*Positive class:*	
dsc.3.8.3	*Negative class:*	
dsc.3.9	Are there any comments on the above?	
dsc.4.0	Describe briefly the process related to the preprocessing, cleaning, and labeling of data.	
dsc.4.1	Which methodological approaches were applied?	
dsc.4.2	Who was involved in the process, and what was their role?	
dsc.4.3	What was the rationality behind the approach?	
dsc.4.4	Are there discrepancies between training and production data that can affect classifications? If yes describe	

dsc.4.5	Is there knowledge of potential future mismatches between training and production data that may affect classifications? If yes describe	
dsc.4.6	# Repeat question dsc.4.6.1, dsc.4.6.2, dsc.4.6.3, and dsc.4.6.4 must be repeated for every feature in the model.	
dsc.4.6.1	What is the name of the feature?	
dsc.4.6.2	What Variable type is the feature	
dsc.4.6.3	Which Data, origin/ model, does the feature have?	
dsc.4.6.4	Describe what the feature covers	
dsc.4.7	# Personal, sensitive, and protected data category	
dsc.4.7.1	Does the model process contact information? If yes, describe:	
dsc.4.7.1.1	Is contact information included in the dataset? If yes, describe:	
dsc.4.7.1.2	Is contact information included as a feature? If so, which feature:	
dsc.4.7.1.3	Is contact information included as the target? If yes, describe:	
dsc.4.7.1.4	Is contact information indirectly included in the dataset through a proxy? If yes, describe:	
dsc.4.7.1.5	Has a negative or positive bias been observed in relation to contact information? If yes, describe:	
dsc.4.7.1.6	Are there any comments to the questions above?	
dsc.4.7.2	Does the model process employment information? If yes, describe:	
dsc.4.7.2.1	Is employment information included in the dataset? If yes, describe:	
dsc.4.7.2.2	Is employment information included as a feature? If so, which feature:	
dsc.4.7.2.3	Is employment information included as the target? If yes, describe:	
dsc.4.7.2.4	Is employment information indirectly included in the dataset through a proxy? If yes, describe:	
dsc.4.7.2.5	Has a negative or positive bias been observed in relation to employment information? If yes, describe:	
dsc.4.7.2.6	Are there any comments to the questions above?	
dsc.4.7.3	Does the model process information on ethnicity? If yes, describe:	
dsc.4.7.3.1	Is information on ethnicity included in the dataset? If yes, describe:	
dsc.4.7.3.2	Is information on ethnicity included as a feature? If so, which feature:	

dsc.4.7.3.3	Is ethnicity information included as the target? If yes, describe:	
dsc.4.7.3.4	Is ethnicity indirectly included in the dataset through a proxy? If yes, describe:	
dsc.4.7.3.5	Has a negative or positive bias been observed in relation to ethnicity? If yes, describe:	
dsc.4.7.3.6	Are there any comments to the questions above?	
dsc.4.7.4	Does the model process information on a person's political, religious, or philosophical beliefs? If yes, describe:	
dsc.4.7.4.1	Is information on a person's political, religious, or philosophical beliefs included in the dataset? If yes, describe:	
dsc.4.7.4.2	Is a person's political, religious, or philosophical belief? Included as a feature? If so, which feature:	
dsc.4.7.4.3	Is a person's political, religious, or philosophical belief included as the target? If yes, describe:	
dsc.4.7.4.4	is a person's political, religious, or philosophical belief? Indirectly included in the dataset through a proxy? If yes, describe:	
dsc.4.7.4.5	Has a negative or positive bias been observed in relation to a person's political, religious, or philosophical beliefs? If yes, describe:	
dsc.4.7.4.6	Are there any comments to the questions above?	
dsc.4.7.5	Does the model process union information? If yes, describe:	
dsc.4.7.5.1	Is union information included in the dataset? If yes, describe:	
dsc.4.7.5.2	Is union information included as a feature? If so, which feature:	
dsc.4.7.5.3	Is union information included as the target? If yes, describe:	
dsc.4.7.5.4	Is union information indirectly included in the dataset through a proxy? If yes, describe:	
dsc.4.7.5.5	Has a negative or positive bias been observed in relation to union information? If yes, describe:	
dsc.4.7.5.6	Are there any comments to the questions above?	
dsc.4.7.6	Does the model process health information? If yes, describe:	
dsc.4.7.6.1	Is health information included in the dataset? If yes, describe:	
dsc.4.7.6.2	Is health information included as a feature? If so, which feature:	
dsc.4.7.6.3	Is health information included as the target? If yes, describe:	
dsc.4.7.6.4	Is health information indirectly included in the dataset through a proxy? If yes, describe:	

dsc.4.7.6.5	Has a negative or positive bias been observed in relation to health information? If yes, describe:	
dsc.4.7.6.6	Are there any comments to the questions above?	
dsc.4.7.7	Does the model process information on sexuality? If yes, describe:	
dsc.4.7.7.1	Is information on sexuality included in the dataset? If yes, describe:	
dsc.4.7.7.2	Is information on sexuality included as a feature? If so, which feature:	
dsc.4.7.7.3	Is sexuality information included as the target? If yes, describe:	
dsc.4.7.7.4	Is sexuality indirectly included in the dataset through a proxy? If yes, describe:	
dsc.4.7.7.5	Has a negative or positive bias been observed in relation to sexuality? If yes, describe:	
dsc.4.7.7.6	Are there any comments to the questions above?	
dsc.4.7.8	Does the model process criminal record information? If yes, describe:	
dsc.4.7.8.1	Is criminal record information included in the dataset? If yes, describe:	
dsc.4.7.8.2	Is criminal record information included as a feature? If so, which feature:	
dsc.4.7.8.3	Is criminal record information included as the target? If yes, describe:	
dsc.4.7.8.4	Is criminal record information indirectly included in the dataset through a proxy? If yes, describe:	
dsc.4.7.8.5	Has a negative or positive bias been observed in relation to criminal record information? If yes, describe:	
dsc.4.7.8.6	Are there any comments to the questions above?	
dsc.4.7.9	Does the model process CPR (Personal Identification Number) information? If yes, describe:	
dsc.4.7.9.1	Is CPR information included in the dataset? If yes, describe:	
dsc.4.7.9.2	Is CPR information included as a feature? If so, which feature:	
dsc.4.7.9.3	Is CPR information included as the target? If yes, describe:	
dsc.4.7.9.4	Is CPR information indirectly included in the dataset through a proxy? If yes, describe:	
dsc.4.7.9.5	Has a negative or positive bias been observed in relation to CPR information? If yes, describe:	
dsc.4.7.9.6	Are there any comments to the questions above?	
dsc.4.7.10	Does the model process other personal data? If yes, describe:	

	# This category must be repeated in case that there more than one type of other personal information	
dsc.4.7.10.1	Is other personal data included in the dataset? If yes, describe:	
dsc.4.7.10.2	Is other personal data included as a feature? If so, which feature:	
dsc.4.7.10.3	Is other personal data included as the target? If yes, describe:	
dsc.4.7.10.4	Is other personal data indirectly included in the dataset through a proxy? If yes, describe:	
dsc.4.7.10.5	Has a negative or positive bias been observed in relation to other personal data? If yes, describe:	
dsc.4.7.10.6	Are there any comments to the questions above?	
dsc.11.6	Has bias (such as historical, representation, measurement, aggregation, evaluation, and implementation bias) been considered for the model? Why or why not?	
dsc.11.7	Has fairness been taken into account in the model's performance? (Including how the model performs on minorities and underrepresented classes.). Why or why not?	
dsc.11.8	What has been done, and what is the result?	
dsc.11.9	Does the model deliver data to other models (output data)? If yes, which:	
dsc.12.0	In what context is the model's output intended for use?	
dsc.12.1	In what contexts should the model not be used?	

Table 4 AIRA: Legal Part (Facilitator Part in Nagbøl et al. (2021))

Q#	Question	Answer
leg.1.0	Does the model solve the business need?	
leg.1.1	Which Artificial Intelligence Act category does the model belong to (Prohibited/High-risk/Non-high-risk)?	
leg.1.2	Does the model fulfill expectations regarding effect?	
leg.1.3	Is the model transparent enough?	
leg.1.4	Has there been a satisfactory check for differences in training and production data, as well as problematic biases?	
leg.1.5	Have appropriate safety measures been taken?	
leg.1.6	Is it ensured that relevant legislation is complied with?	

Evaluation Plan Framework

The Evaluation Plan framework adapted and revised from Nagbøl and Müller (2020) and Nagbøl et al. (Submitted) (see table 5). For description see Nagbøl and Müller (2020) and Nagbøl et al. (Submitted). For the design principles see Nagbøl et al. (Submitted).

Table 5 The Evaluation Plan framework adapted and revised from Nagbøl and Müller (2020) and Nagbøl et al. (Submitted)

Q#	Question	Answer
EP.1.0	Who should participate in the evaluation (e.g., application manager, relevant business unit, ML lab)?	
EP.1.1	Who owns the model/the solution (usually the business)?	
EP.1.2	When should the first evaluation meeting take place?	
EP.1.3	What is the expected meeting frequency (How often should you meet and evaluate)?	
EP.1.4	What is the current threshold setting for the AI system?	
EP.1.5	What is the basis for the evaluation (e.g., logging data, annotated evaluation data, i.e.., data where human categorization is compared with the model)?	
EP.1.6	Is data unbalanced to a degree where this must be taken into account when fabricating data for evaluation and retraining? If so, how?	
EP.1.7	What resources are needed (e.g., who can make evaluation data, evaluation data is provided internally or externally, how much needs to be evaluated, what is the cost in time/money)?	
EP.1.8	What is the expected resource need for the evaluation?	
EP.1.9	Is the model visible or invisible to external users?	
EP.2.0	Does the model receive input from other models? If so, which ones?	
EP.2.1	What are success and error criteria (e.g., When does a model perform good/bad, what percentage, business value, labor waste)?	
EP.2.2	Is there future legislation that will have an impact on the model's performance (e.g., the introduction of new requirements, abolition of requirements, or the like)?	
EP.2.3	Are there other future factors that affect the model's performance (e.g., bias, circumstances, data, standards, or the like)?	
EP.2.4	When should the model be retrained?	
EP.2.5	When should the model be muted or deactivated?	

Evaluation Support Framework

The Evaluation Support Framework (see table 6) is described in Nagbøl and Müller (2020).

Table 6 The Evaluation Support Framework adapted and revised from Nagbøl and Müller (2020)

Q#	Question	Answer
ES.0.9	Model name and version number	
ES.1.0	Date of evaluation	
ES.1.1	When was the model evaluated the last time?	
ES.1.2	What did the last evaluation show?	
ES.1.3	Who is participating in the meeting?	
ES.1.4	Who is conducting the current evaluation?	
ES.1.5	How many cases/documents have been reviewed in the evaluation (find minimum)	
ES.1.6	Is the data for evaluation satisfying?	
ES.1.7	What does the evaluation show?	
ES.1.8	Has performance on the model decreased?	
ES.1.9	Has performance on the model increased?	
ES.2.0	Has anything happened in the meantime (since last evaluation) that could have an impact on the model's performance?	
ES.2.1	What is the threshold setting?	
ES.2.2	What does the history of the threshold show?	
ES.2.4	Why has the threshold setting been changed?	
ES.2.5	Is there still a business need/case for the model? If not, should the model be shut down?	
ES.2.6	What value does the model provide?	
ES.2.7	Has the model fulfilled its purpose?	
ES.2.8	Is there future legislation or events that will have an impact on the performance of the model? (bias, the introduction of new requirements, abolition of requirements, etc.)	
ES.2.9	Should the model be retrained based on the evaluation?	

Retraining Framework

The Retraining Framework (see table 7) is described in Nagbøl and Müller (2020).

Table 7 The Retraining Framework adapted and revised from Nagbøl and Müller (2020)

Q#	Question	Answer
1.5.GS.0.9	Model name and version	
1.5.GS.1.0	Why should the model be retrained?	
1.5.GS.1.1	What did the last evaluation of the model show?	

1.5.GS.1.2	Own take on root cause (why retrain the model?) (e.g., change in document form, legislation, tenders, etc.)	
1.5.GS.1.3	Is there new training data available for retraining? (including human resources)	
1.5.GS.1.4	How critical is it to get the model retrained?	
1.5.GS.1.5	Is the model dependent on other models? Yes/No -Status of them	
1.5.GS.1.6	What does training data look like compared to the current situation? (e.g., change in document form, legislation, tenders, etc.)	
1.5.GS.1.7	Is it possible to recycle parts of training data vs. brand new training dataset (which of the previous training data can be used).	
1.5.GS.1.8	Is data unbalanced to a degree where it has to be taken into account when fabricating data for retraining? If so, how?	
1.5.GS.1.9	Observed suspiciousness (e.g., bias against industry, gender, ethnicity, forms of businessed, etc.) Is that a problem? Yes/No	
1.5.GS.2.0	Does the model deliver outputs to other models (input)? Yes/No - Status of them?	
1.5.GS.2.1	Have algorithms been developed that better solves the task since the model was put into operation?	
1.5.GS.2.2	"Concluding field" Has a decision been made regarding if the model should be retrained (Have all stakeholders agreed on that the model should be retrained)	

Summary of findings

Paper 1 (Asatiani *et al.*, 2021) answered the research question, *How can an organization exploit inscrutable AI systems in a safe and socially responsible manner?* The question was answered through a case study in the Danish Business Authority. Here were found AI requirements. The DBA had to dedicate attention to ensuring instrumental outcomes did not lead to ignoring human outcomes. Because the DBA was a public agency, it was required to make the fairest and bias-free decisions possible; diverse stakeholders such as managers, data scientists, systems developers, and caseworkers have shaped the approach to balance the explainability–accuracy trade-off.

The analysis based on the exploratory case study revealed three significant findings. First, it proved that the conceptual work on envelopment (Floridi, 2011a; Robbins, 2020) held

empirical validity in a knowledge-work context. We found the DBA had actively used boundaries, training data, and input and output data but not function envelopment. Second, we found that envelopment is a sociotechnical matter, with human agents having a central role in defining, maintaining, and renegotiating the envelopes. The envelopment methods mentioned by the interviewees to limit the AI system's capabilities were never solely technical. Instead, they were negotiated iteratively, involving several stakeholders' views, responsibilities to society, and implications for employees' workflow. Third, the articulation of the found connection between envelopment methods and model choice indicates that while envelopment does not alter the relationship between accuracy and explainability, it does allow for responsibly choosing from a broader range of models (see figure 4).

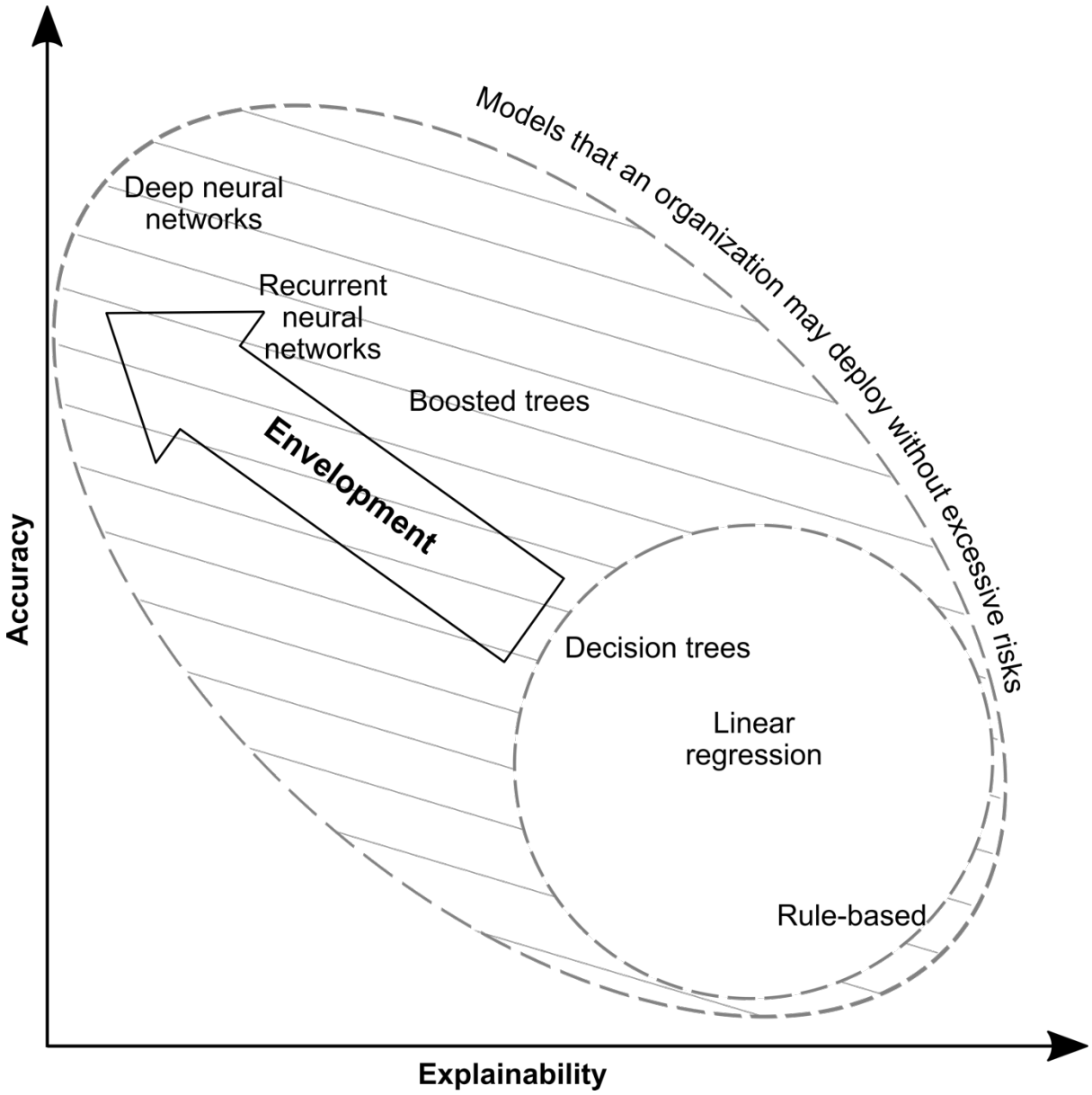


Figure 4 "How Envelopment Expands the Set of Models an Organization May Adopt Without Excessive Risks" from Asatiani et al. (2021, p. 340)

Paper 2 (Asatiani *et al.*, 2020) answers the research question *How can organizations reconcile the growing demands for explanations of how AI based algorithmic decisions are made with their desire to leverage AI to maximize business performance?* The paper is aimed at a practitioner audience with findings of identifying the six elements: the model, goal, training data, input data, output data, and environment of a hypothetical AI agent (see figure 5). The paper provides a six-dimensional framework (see table 8) corresponding to the six elements of the AI agent. The paper provides examples of how the DBA addressed the six dimensions in the framework (see figure 6).

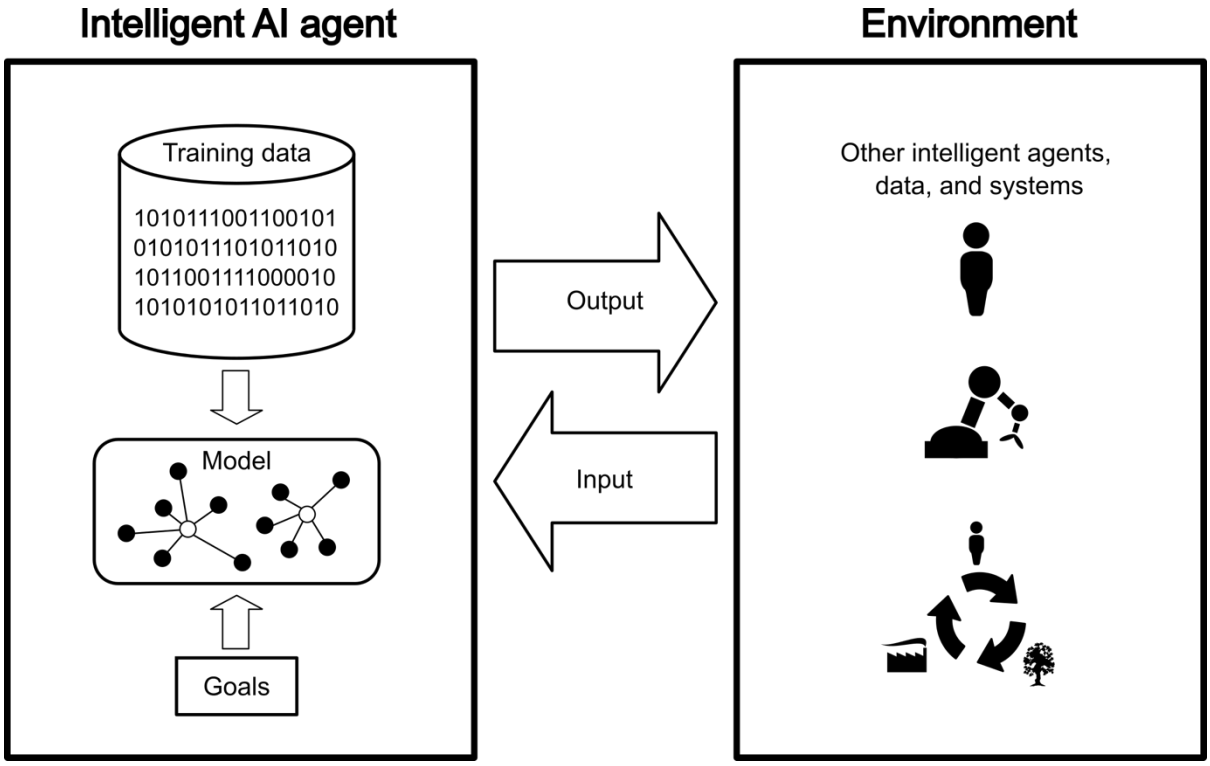


Figure 5 "The Six Elements of an Intelligent AI Agent" from Asatiani et al. (2020, p. 261)

Table 8 "Six-Dimension Framework for Explaining the Performance of AI Systems" from Asatiani et al. (2020, p. 263)

Dimension	Description	Example
1. Model	Explanation of the AI system's logic/behavior based on tracing its decision-making patterns.	A specific business-risk probability may be explained by the if-then sequence of steps taken by a business-risk estimation model.
2. Goals	Explanation of the AI system's logic/behavior derived from priorities or the strategic basis for a given decision.	The agent flags high probabilities of risk for companies that engage in reputation-compromising activities such as producing health-harming products or causing environmental damage, with the explanation lying in the fact that the model is trained and tested with performance metrics that give great weight to risking the organization's reputation.
3. Training Data	Explanation based on the characteristics of the training data.	The agent assigns exceptionally high probabilities of risk to certain types of business, such as medical practices, because of biased training data. Data on medical practitioners might have been collected in economically deprived areas while data from other businesses are geographically more diverse.

4. Input Data	Explanation based on the characteristics of the input data.	Unreliable business-risk probabilities can be explained by low-quality input data produced by inaccurate measurement of relevant risk factors.
5. Output Data	Explanation derived from humans' examination and verification of the output.	A human examines the validity of the AI agent's business-risk probability for a loan application and makes sure that the rationale for the decision can be explained to the applicant in meaningful terms.
6. Environment	Explanation that is based on the environment in which the AI agent operates.	Inappropriate risk estimations may be explained by the AI agent being fed risk-assessment data from environments that are not suitable for this purpose (e.g., using soccer-league scoring data to predict the risks of businesses not connected to soccer).

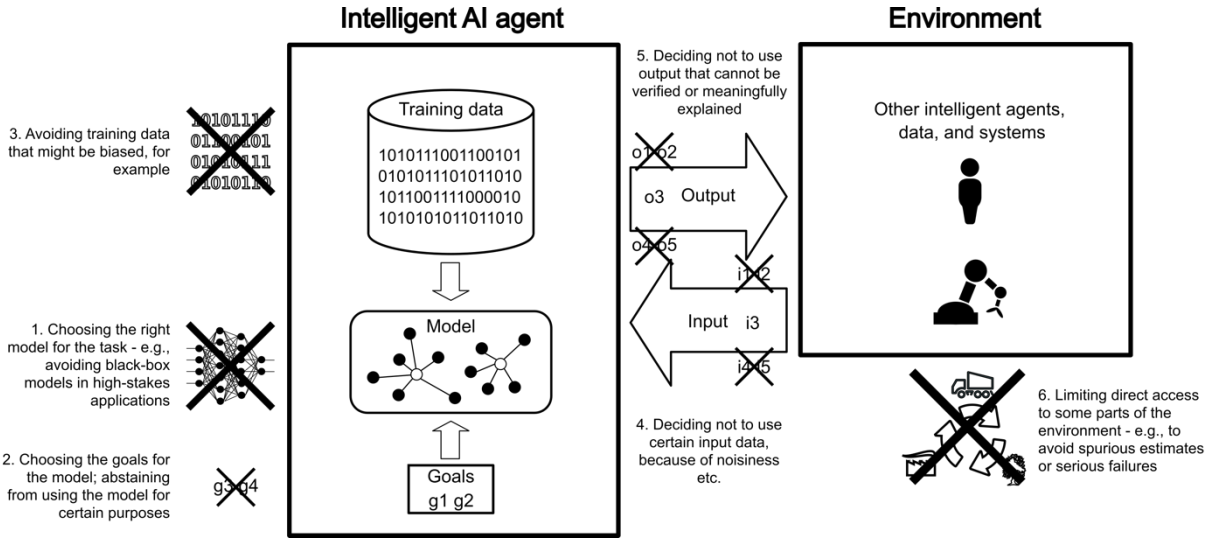


Figure 6 "The DBA's Approach Took Account of all Six Dimensions of the Framework" from Asatiani et al. (2020, p. 267)

Paper 3 (Nagbøl and Müller, 2020) answers the research question *How do we ensure that machine learning (ML) models meet and maintain quality standards regarding interpretability and responsibility in a governmental setting?* The key finding from paper 3 is that a framework for responsible AI should contain AI assessment, evaluation planning, evaluation, and retraining.

Paper 4 (Nagbøl, Müller and Krancher, 2021) answers the research question *How should procedures be designed to assess the risks associated with a new AI system?* The paper presented findings during three iterations: building, intervening, and evaluation. We found in the first iteration that an assessment tool was likely to work in the DBA context and expand the tool to include user stories from a business perspective and data privacy. Furthermore, we identified a desire to calculate a risk score. The second iteration yielded two main findings. The first was the need to utilize knowledge from different stakeholders such as domain experts and data scientists. The second was that the questionnaire in its entirety was too time-consuming; different stakeholders should fill out different parts, and the questionnaire should be filled out before the meeting and discussed at the meeting. We abandoned the idea of calculating a risk score. The third iteration found a need to improve the readability of some questions and consider preparation requirements for the informants. Finally, paper four articulated five design principles (see table 9) following the guidelines from Gregor et al. (2020).

Table 9 "Design principles for an artificial intelligence risk assessment tool" from Nagbøl et al. (2021, pp. 335–336)

Principle of…	Aim, implementer, and user	Mechanism	Rationale
1: Multi-perspective expert assessment	To perform a multi-perspective risk assessment (aim), organizations using AI should…	… ensure that the AI system is jointly assessed by users (domain experts) and developers (data scientists)	Risk assessment in socio-technical systems implies integrating knowledge from business and technical perspectives (Barki, Rivard and Talbot, 2001; Wallace, Keil and Rai, 2004)
2: Structured intuition	To motivate and engage diverse stakeholders to participate in risk assessment (aim), organizations using AI (implementers) should…	… prescribe aspects that need to be assessed, but not the specific methods or tools to be used for that assessment	Risk assessment needs to strike a balance between deliberate analysis and structure to ensure motivation and coverage of key risks (Moeini and Rivard, 2019)

3: Expected consequences	To make risk assessments based on expected real-world consequences instead of lab results (aim), organizations using AI (implementers) should…	… combine probabilities of outcomes of algorithmic decisions (e.g., true positive/negative rate) with their respective costs and benefits	Considering both risk probabilities and their impacts is a common practice in risk management (Boehm, 1991; Moeini and Rivard, 2019). Drawing on expected utility theory (Morgenstern and Von Neumann, 1944), we extend this idea to also take positive outcomes into consideration
4: Beyond accuracy	To account for risks beyond "false predictions" (aim), organizations using AI (implementers) should…	… evaluate AI systems not only in terms of predictive accuracy but also in terms of dimensions like interpretability, privacy, or fairness	We draw on Lipton's (2018) desiderata of interpretable ML (trust, causality, transferability, informativeness, and fair and ethical decision making) and the accompanying properties of interpretable models in terms of transparency and post-hoc explainability. The principle is further backed up by the EU GDPR
5: Envelopment of black boxes	To leverage the superior predictive power of complex "black box" AI systems with minimal risks, organizations	… envelop the training data, inputs, functions, outputs, and boundaries of their AI systems	In robotics, envelopes are three-dimensional cages built around industrial robots to make them achieve their purpose without

	using AI (implementers should…		harming human workers or destroying physical things (Floridi, 2011a). The idea has recently been transferred to ML by Robbins (2020) and Asatiani et al. (2021)

Paper 5's main findings consist of five challenges addressed by five design principles (see figure 7). The five challenges are choosing and preparing appropriate data, estimating resource needs and availability, maintaining an overview, prioritizing evaluations, and timing evaluations.

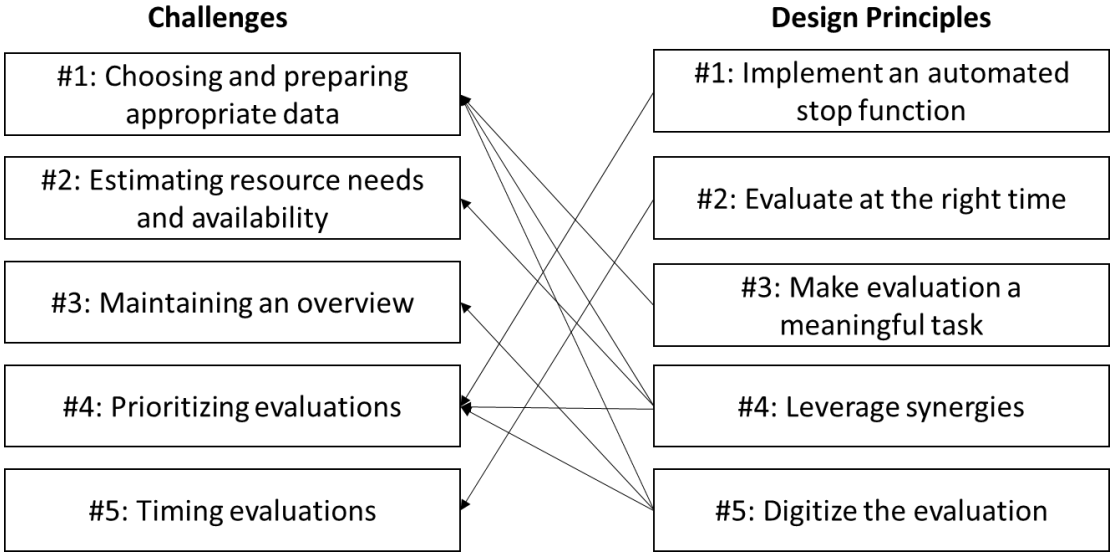


Figure 7 "Challenges of and Design Principles for AI Systems Evaluation" from Nagbøl et al. (Submitted, p. 11)

The five design principles are described in the table (taken from paper 5) according to the guidelines from Gregor et al. (2020).

Table 10 Design Principles from Nagbøl et al. (Submitted, pp. 15–16)

Principle	Aim	Mechanism	Rationale
#1: Implement an automated stop function	To enforce compliance with the Evaluation Plan…	…ensure that the AI system cannot run in production without being evaluated by	As (semi-)autonomous systems, AI systems can cause undesired consequences. Emergency stop measures as known from other dangerous machines like power

		humans as per the Evaluation Plan.	saws or lawn mowers can help to prevent some of these consequences.
#2: Evaluate at the right time	To make sure that the AI system is up to date when needed…	…consider event-based and frequency-based timing strategies in line with expected real-world changes.	According to representation theory (Recker *et al.*, 2019), the basic purpose of any information system, including AI-based systems. is to faithfully represent certain real-world phenomena. Hence, AI systems need to be re-evaluated and, if needed, re-trained whenever the real-world phenomenon they are representing changes.
#3: Make evaluation a meaningful task	To ensure motivated evaluators…	…design the annotation task so that it is an opportunity for autonomy, competence, and relatedness.	According to self-determination theory (Ryan and Deci, 2000), satisfying the basic psychological needs for autonomy, competence, and relatedness can increase people's intrinsic motivation for a given task.
#4: Leverage synergies between AI system evaluation, human training, human work, and AI system training	To reduce costs and make evaluation work less tedious ...	… recycle data between work, evaluation, and training activities.	According to representation theory (Recker *et al.*, 2019), information systems are representations of real-world work systems. Hence, the task of training and assessing an AI-based decision-making system (a type of information system) has important parallels to the task of training and assessing a human decision-making system, suggesting that synergies between these two can be leveraged, e.g., by reusing the products of human training efforts for AI training or assessment.
#5: Digitize the evaluation	To ensure compliance with Evaluation plans and maintain an overview …	… implement a digital platform that automatically collects data about evaluation activities and outcomes.	According to control theory (Eisenhardt, 1985), accurate information about a controlee's behavior makes it more likely that the controlee will engage in the desired behaviors. Digitizing the evaluation infrastructure helps make information about evaluation activities transparent and thus encourages evaluators (i.e., controlees) to comply with Evaluation Plans.

Discussion of thesis

Contribution of the five papers

The thesis asks the research question: *How can organizations ensure responsible use of artificial intelligence?* The question is answered through the five papers of the thesis contributing to both theory and practice.

This part of the thesis initiates with presenting and discussing the contributions of papers 1, 4, and 5 before ending with discussing the entire thesis. Paper 2 is absent in this part of the thesis due to the nature of the JAIS (Benbya, Pachidi and Jarvenpaa, 2021) and MISQ-E (Benbya, Davenport and Pachidi, 2020) double issue. The JAIS paper (Asatiani *et al.*, 2021) targeted the academic audience, while the MISQ-E paper (Asatiani *et al.*, 2020) targeted the practitioner audience. Therefore, paper 2's contribution is presented under practical implications. Paper 3 is not present here due to being a research-in-progress paper without a theoretical contribution.

Paper 1 (Asatiani *et al.*, 2021) has two key contributions. The first contribution is to introduce the concept of sociotechnical envelopment and provide evidence showing that the DBA has successfully applied boundary, training-data, and input and output envelopment. Meanwhile, function envelopment was not found, thereby finding that the concept of envelopment holds empirical validity and is sociotechnical. The second contribution is that while envelopment does not change the relationship between accuracy and explainability, it enables organizations to manage the trade-off between low explainability and high performance for inscrutable models, thereby allowing, to some extent, for the sacrifice of a bit of explainability for higher accuracy without risking harmful consequences. Sarker et al. (2019) warn in their review of sociotechnical approaches that IS research too often focuses on technologies' instrumental outcomes. They argue for addressing both the instrumental and humanistic outcomes. We found that the DBA did not only focus on instrumental outcomes, such as efficiency and higher precision. The DBA must ensure that using AI would not lead to the misuse of government power or the unnecessary surveillance of both citizens and companies. These actions would compromise the DBA's integrity and potentially harm public trust. The AI projects' humanistic outcomes were also an internal focus where caseworkers actively redesigned their workflow, identified the problem domain, and

developed the AI system. In summary, *"We propose theoretical implications for (1) describing organizational AI implementation as a balancing act between human and AI agency, and (2) conceptualizing sociotechnical envelopment as the primary tool for this crucial balancing act"* (Asatiani *et al.*, 2021, p. 341).

Paper 4 (Nagbøl, Müller and Krancher, 2021) has two types of contribution. The first contribution is the designed artifact AIRA that supports the DBA in assessing risks associated with new AI systems. The second is a theoretical contribution in the form of five design principles for AI risk management. The first three design principles are grounded in risk management theory and focus on involving diverse stakeholders; meanwhile, the last two are grounded in the literature related to interpretable and safe machine learning. The theoretical contributions went beyond the existing research in four ways. Firstly, our work emphasizes guiding communication between stakeholders of diverse expertise with a focus on interaction between developers and users of AI systems through the AIRA tools' three parts dedicated to domain experts, data scientists, and facilitators. Secondly, the AIRA tool supports its users in assessing risk in relation to benefits for a given AI system by concentrating on establishing a joint understanding among the stakeholders. Thirdly, the AIRA tools accentuate the incorporation of performance metrics beyond accuracy, including assessment of bias, fairness, and interpretability, benefitting not only preproduction risk identification but also postproduction risk monitoring. Fourthly, we contribute to a stronger theoretical foundation of AI documentation and assessment.

Paper 5 (Nagbøl, Krancher and Müller, Submitted) contributes to the area of knowledge concerning the evaluation of productive AI systems in organizations after going live. To the best of our knowledge, little work has been conducted within this area despite its importance in continuously avoiding harm and ensuring benefits. We contributed to the literature by building, implementing, and testing our designed Artifact the Evaluation Plan, which supports organizations structure and secure resources for future evaluation after go-live. In addition, we contributed by identifying five challenges concerning planning and evaluating AI systems post-go-live and providing five design principles for addressing the challenges. Thus, we found that foundational research in the forms of representation theory (Recker *et al.*, 2019) and control theory (Eisenhardt, 1985) was useful in guiding our evaluation infrastructure design. Our work also contributes to the literature by discussing the benefit of combining domain experts and AI experts when developing AI (Lebovitz, Levina and

Lifshitz-Assaf, 2021; Lou and Wu, 2021; Nagbøl, Müller and Krancher, 2021; van den Broek, Sergeeva and Huysman, 2021) and arguing that the collaboration should continue post-go-live.

Contributions to Risk Management

The current risk management literature is not tailored to AI systems' needs, especially after they go live. Bannerman (2008) states that risk management literature lacks the needs from practice. It is an accurate statement when it comes to those related to AI. The literature, to the best of my knowledge, does not address risk-associated needs concerning training data that increasingly deviates from production data over time, bias in data, or discriminating decisions. The literature also does not provide methods for bias detection, fairness measurement, and ongoing AI evaluation and maintenance. The risk management literature provides foundational knowledge to approaching risk. For example balancing the use of deliberate analysis and intuition (Moeini and Rivard, 2019), leading to our AI-specific principle of structured intuition (Nagbøl, Müller and Krancher, 2021) or supporting our work by emphasizing the necessity for activities such as risk assessment (B. W. Boehm, 1991). The risk management literature has mainly informed the thesis's work on approaching the object of focus—in this case, AI systems.

The thesis contributes to the stream of risk management literature by building on the work of Moeini and Rivard (2019) and Boehm (1991) in identifying AI-specific risks. To recall, "*[t]he most common definition of risk in software projects is in terms of exposure to specific factors that present a threat to achieving the expected outcomes of a project"* (Bannerman, 2008, p. 2119). These specific factors can be—but are not limited to—an inappropriate use of Black-box AI, opaque or unfair decisions as well as different forms of bias including historical, representation, measurement, aggregation, evaluation, and deployment.

These new risks introduced with AI change the approaches to categories described by Boehm (1991): risk assessment with the subcategories risk identification, risk analysis, and risk prioritization. Approaches to identifying and mitigating these risks are already present in streams of literature in IS and reference disciplines. These measures are often stand-alone measures addressing a single AI-related issue, such as bias and fairness (Suresh and Guttag, 2020), interpretability of machine learning (Lipton, 2017, 2018; Du, Liu and Hu, 2019),

envelopment (Robbins, 2020; Asatiani *et al.*, 2021), suitable inclusion of domain expertise (Lebovitz, Levina and Lifshitz-Assaf, 2021; Lou and Wu, 2021; van den Broek, Sergeeva and Huysman, 2021), lack of documentation of datasets and models (Mitchell *et al.*, 2019; Gebru *et al.*, 2020), or conventual performance metrics like ROC AUC (Spackman, 1989; Fawcett, 2006). It is a contribution in itself to combine those approaches into a holistic approach to responsible AI (Nagbøl and Müller, 2020). This approach suggests a pre-production risk assessment and evaluation providing the design principles of multi-perspective expert assessment, structured intuition, expected consequences, beyond accuracy, and envelopment of black boxes (Nagbøl, Müller and Krancher, 2021). It also demands that the evaluation must be planned for and continued post-production, and demanding retraining of AI systems when necessary. We are, to the best of my knowledge, the first to describe the challenges and solutions to post–go-live continuous evaluation of AI in a governmental setting (Nagbøl, Krancher and Müller, Submitted).

Contributions to Envelopment

The thesis contributes to the theory of envelopment of AI. Robbin suggests, inspired by the work of Floridi (2011b, 2011a), enveloping virtual AI by defining the five properties of training data, inputs, functions, outputs, and boundaries envelopes (Robbins, 2020). We contribute to this stream of literature by providing empirical evidence for boundary, training data, and input and output envelopment, that the envelopment is sociotechnical, and allows organizations, to some extent, to lower the requirements for interpretability of AI systems without jeopardizing safety (Asatiani *et al.*, 2021). Examining the properties of envelopment can provide explanations for the behavior of AI systems (Asatiani *et al.*, 2020). The thesis furthermore provides concrete suggestions for how to envelop AI systems (see artifact: AIRA) (Nagbøl, Müller and Krancher, 2021). Sociotechnical envelopment offers an approach that enhances responsible conduct when using both opaque and interpretable AI systems, hence, providing a solution to scenarios in which there is a demand for responsible use of AI, but lower accuracy will cause harm.

Methodological Reflections: Action Design Research as a Project

This part of the thesis reflects on working on an entire Ph.D. project period with ADR as a research method. The format of the collaborative Ph.D. created a setting where every other

workday, on average, was spent at the DBA. Therefore, it was decided from the beginning that ADR was a natural fit. Furthermore, choosing ADR as an approach provided several benefits, such as working scientifically with the design of IT artifacts and testing them and the ingrained theory in an authentic setting (Sein *et al.*, 2011), ascribing the work in the DBA scientific relevance.

The method presents difficulties from a longitudinal project perspective in managing and balancing the action, design, and research parts. A lack of attention toward these parts can potentially jeopardize or ruin the project. The engagement triangle (see figure 9) is drawn to summarize my experience with relying on ADR for my Ph.D. project.

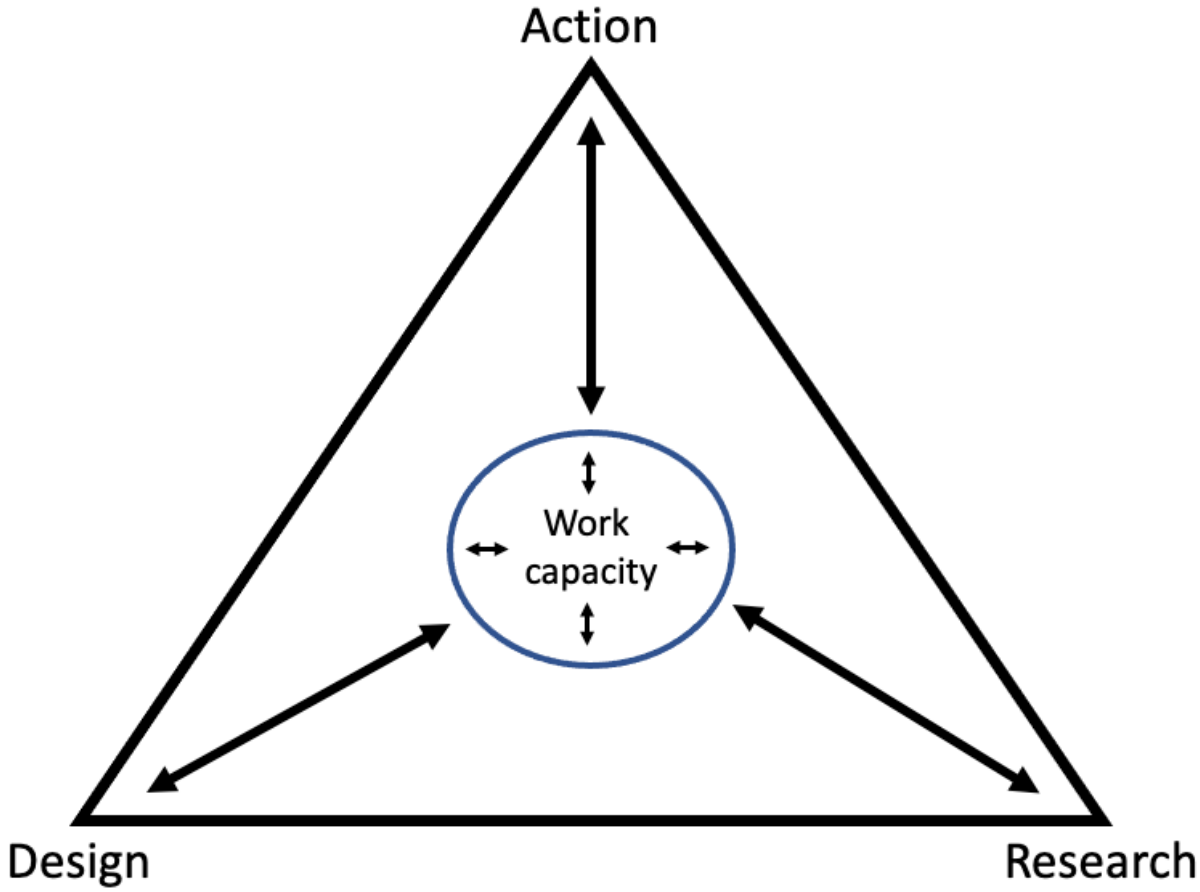


Figure 8 Engagement Triangle

The engagement triangle has each corner representing one of the aspects of ADR: Action, Design, and Research. The arrows represent engagement toward one of those aspects. The arrows are attached to work capacity representing the amount of work that I, as a user of the ADR method, provide. The amount of work capacity is adjustable within the limitations of the 24 hours in a day, work-life balance, and human fatigue. All three aspects can easily fill

up a work schedule. ADR has the principle of authentic and concurrent evaluation (Sein *et al.,* 2011) and that the organizational use shapes the artifact. The action dimension represents organizational engagement and involvement. If that aspect is not properly nurtured, the artifacts might be discarded or not used in the organization. Commitment here is important because that organization needs to be continuously convinced that your design exists and provides value. It cannot be expected that an organization will implement a design that will impact people's lives solely to satisfy the work for a Ph.D. thesis. The project needs organizational use. The research aspect represents the work's theoretical grounding and contribution. ADR has the principle theory ingrained artifacts (Sein *et al.,* 2011). Insufficiently nurturing the research can result in an unsatisfying theoretical foundation and ultimately leave the project without a scientific contribution. The design aspect touches on designing and physically constructing the artifact, which might be a resource-heavy endeavor. A lack of attention in this direction can cause a lack of artifact and, hence, a lack of project. I had to accept that more work could be done towards each aspect and, most importantly, determine how to balance and manage the engagement. Further research could address managing ADR for long-term projects.

Implications for Practice

The thesis provides two main contributions to practice. First, the X-RAI artifact is currently implemented and mandatory for all AI systems in the Danish Business Authority. The work on X-RAI artifacts has contributed to a high standard of AI ethics in the DBA. The quality standard is evident in the recognized consultancy company Gartner, which has conducted a case study and recommends the DBA's approach as they write in summary: "*The Danish Business Authority developed a concrete way to apply ethical guidelines to AI model development and assessment, once deployed. D&A and AI leaders can adopt the DBA's approach to ensure they develop and use their AI models in an ethically defensible way"* (Gartner, 2021, p. 1). The methodological guideline of ADR emphasizes that the design must solve an instance of a problem representative of a class of similar problems existing elsewhere and that the design principles embody the knowledge acquired by solving the instance of a problem relevant for solving similar problems (Sein *et al.*, 2011). The problems I address in this thesis is evidently represented elsewhere (Schwartz *et al.*, 2022). The questions asked in the X-RAI frameworks are a part of this thesis and available for others to use to solve problems wherever they see fit. The first two artifacts named the Artificial

Intelligence Risk Assessment (AIRA) tool (Nagbøl, Müller and Krancher, 2021) and the Evaluation Plan (Nagbøl, Krancher and Müller, Submitted) are accompanied by the design principles in the papers and included in this thesis, aiding the implementation and allowing for testing the generalizability of both theoretical outcomes and the designed artifact.

Second, the thesis provides empirical examples of sociotechnical envelopment from the Danish Business Authority, showcasing how AI systems can be enveloped in practice and how and when the use of envelopment to some extent allows for the use of less transparent AI systems without jeopardizing safety (Asatiani *et al.*, 2021). Based on the DBA case study, we found that addressing the framework's six dimensions (the model, goals, training data, input data, output data, and environment of an AI system) for explaining the AI system behavior enables successful and responsible deployment of AI systems, including Black-box algorithms. Furthermore, we provide four recommendations for practitioners, including technical and managerial approaches. The first recommendation is *to implement strict controls on the use of Black-box AI systems*, the second is *to use modular design to make it easier to explain the behavior of an AI system*, the third is *to avoid online learning if the need for an explanation is a priority*, and the fourth *facilitate continuous open discussion between stakeholders* (Asatiani *et al.*, 2020).

Limitations

The thesis has two kinds of limitations in the form of circumstances and scope.

The project has been impacted by the Covid-19 pandemic. Working from home has limited access to everyday life at the Danish Business Authority. Covid-19 put some work on hold in the DBA to support Danish society during a crisis. This especially relates to the ML lab and other departments that work with the Covid-19 compensation packages aiming toward saving Danish society from mass bankruptcy and the associated consequences. These circumstances have delayed the development of X-RAI and caused a lack of use of the artifacts. This has led to fewer opportunities for authentic evaluations demanded by ADR. The last two artifacts—the evaluation support framework and the retraining framework—are therefore still in an early version edition.

The thesis has many limitations regarding its scope. Traditional IT maintenance aspects are beyond the scope of this thesis theoretically and technologically. The work does not touch upon the construction and evaluation of infrastructure related to databases, cyber security, IT project management, data availability, or technical integrations. The work generally does not address engineering aspects related to hardware and servers, code efficiency, project planning, or managing cost. Risk management is not used in the conventual project management way, which is often described in the literature. The reason for this is that the focus is on AI-specific parts. The nature of AI enforces a continued assessment or evaluation of the AI system even after being put to use. The thesis does not address environmental issues related to the use of AI. The thesis does not address the relationship between vendor and government in IT development and how it impacts maintaining technical and domain competencies, which are important for the post-go-live continuous evaluation and retraining of AI systems. That might conflict with IT development approaches in organizations where AI systems are developed by vendors who leave the organization post-implementation. It has been argued that while some biases are problematic, some provide value. For example, Google searches for restaurants provide more relevant results when they are biased by location (Søgaard, 2016). That discussion is outside the scope of the thesis. The term "harm" in this thesis describes unintended harm. Hence, the term does not address AI systems purposely designed to harm people, such as military or similar AI systems. The discussion of whether these AI systems should be allowed is beyond the scope. Nevertheless, these AI systems most likely have the highest and most fatal unintended consequences of false positives or negatives, and adequate measures must be taken to avoid bias and discrimination. The thesis does not address futuristic topics that are not expected to be relevant in the foreseeable future.

Future Work

Future work will continue the development of X-RAI and make it an accessible open-source in a tool version in a git repository. That will allow for testing the generalizability in other organizational contexts by, for example, supporting the implementation, asking users about feedback, or allowing users to improve the tool.

Furthermore, the Artificial Intelligence Act (AIA) (European Commission, 2021) will be ingrained (Sein *et al.*, 2011) into the X-RAI artifact, thereby making X-RAI a tool that

supports AIA compliance. This thesis takes the first steps in delivering a European edition of the symbiotic legislation and tool relationship that exists between the Canadian directive on automated decision-making (Secretariat, Treasury Board of Canada, 2019) and the Algorithmic Impact Assessment tool (Secretariat, Treasury Board of Canada, 2020). The focus on the European Artificial Intelligence Act has led to the newest edition of the AIRA (see artifact) tool having the facilitator part as a legal part. The AIA is analyzed and compared to X-RAI. Future work will focus on reshaping X-RAI according to the findings in this analysis. Future research should address how to design tools to support the legal compliance of AI.

The DBA has decided on an approach where they are responsible for the entire process from the cradle to the grave of AI systems. Future research could address how cradle to grave in-house AI development impacts the public sector. Further work must be put into understanding what the domain experts learn from the AI systems. Do they, for example, inherit the bias of the AI system? How do you prevent Borg behavior? How does mutual learning between humans and AI systems shape future work?

Finally, large-scale AI implementation will introduce new problems, needs, and risks in organizations where an increasing part of the “labor” is carried out by AI systems.
A management tool will be built to accommodate these new problems, needs, and risks. The tool will showcase data from X-RAI and other AI-related data to enable management to gain insight into the performance, status, and evaluation of the AI systems in the DBA, as well as to support top management in developing the AI orientation (Li *et al.*, 2021) and provide a foundation to manage the organization based on knowledge grounded in the current state of AI in the organization. The management tool will support managing an organization with large-scale implementation of AI systems. The research question for that project is this: How should AI management infrastructure be designed to enable efficient control of a large amount of AI systems?

Literature

Allen, A. (2016) *The 'three black teenagers' search shows it is society, not Google, that is racist*, *the Guardian*. Available at: http://www.theguardian.com/commentisfree/2016/jun/10/three-black-teenagers-google-racist-tweet (Accessed: 12 November 2020).

Almeder, R. (2014) 'Pragmatism and science', in *The Routledge Companion to Philosophy of Science*. Second. Routledge.

Angwin, J. *et al.* (2016) *Machine Bias*, *ProPublica*. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing?token=1B8jKuq-H9G4ZEq4_95FZ7ZaZ9a3rKDs (Accessed: 11 November 2020).

Asatiani, A. *et al.* (2020) 'Challenges of Explaining the Behavior of Black-Box AI Systems', *MIS Quarterly Executive*, 19(4), pp. 259–278.

Asatiani, A. *et al.* (2021) 'Sociotechnical Envelopment of Artificial Intelligence: An Approach to Organizational Deployment of Inscrutable Artificial Intelligence Systems', *Journal of the Association for Information Systems*, 22(2), pp. 325–352. Available at: https://doi.org/10.17705/1jais.00664.

B. W. Boehm (1991) 'Software risk management: principles and practices', *IEEE Software*, 8(1), pp. 32–41. Available at: https://doi.org/10.1109/52.62930.

Bannerman, P.L. (2008) 'Risk and Risk Management in Software Projects: A Reassessment', *J. Syst. Softw.*, 81(12), pp. 2118–2133. Available at: https://doi.org/10.1016/j.jss.2008.03.059.

Barki, H., Rivard, S. and Talbot, J. (2001) 'An Integrative Contingency Model of Software Project Risk Management', *Journal of Management Information Systems*, 17(4), pp. 37–69.

Belgrave, L.L. and Seide, K. (2019) 'Coding for Grounded Theory', in *The SAGE Handbook of Current Developments in Grounded Theory*. London: Sage publications, pp. 167-185.

Benbya, H., Davenport, T. and Pachidi, S. (2020) 'Special Issue Editorial: Artificial Intelligence in Organizations: Current State and Future Opportunities', *MIS Quarterly Executive*, 19(4), pp. ix–xxi.

Benbya, H., Pachidi, S. and Jarvenpaa, S. (2021) 'Special issue editorial: Artificial intelligence in organizations: Implications for information systems research', *Journal of the Association for Information Systems*, 22(2), p. 10.

Berente, N. *et al.* (2021) 'Managing artificial intelligence', *MIS Q*, 45(3), pp. 1433–1450.

Boehm, B.W. (1991) ‘Software risk management: principles and practices’, *IEEE Software*, 8(1), pp. 32–41. Available at: https://doi.org/10.1109/52.62930.

van den Broek, E., Sergeeva, A. and Huysman, M. (2021) ‘WHEN THE MACHINE MEETS THE EXPERT: AN ETHNOGRAPHY OF DEVELOPING AI FOR HIRING.’, *MIS Quarterly*, 45(3).

Buolamwini, J. and Gebru, T. (2018) ‘Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification’, in *Proceedings of Machine Learning Research 81:1–15. Conference on Fairness, Accountability, and Transparency*, p. 15.

Charmaz, K. (2006) *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. London United Kingdom: SAGE Publications.

Commission, E. (2018) *COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE EUROPEAN COUNCIL, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Artificial Intelligence for Europe*. Brussels.

Doshi-Velez, F. and Kim, B. (2017) ‘Towards A Rigorous Science of Interpretable Machine Learning’, *arXiv:1702.08608 [cs, stat]* [Preprint]. Available at: http://arxiv.org/abs/1702.08608 (Accessed: 4 March 2021).

Du, M., Liu, N. and Hu, X. (2019) ‘Techniques for Interpretable Machine Learning’, *Commun. ACM*, 63(1), pp. 68–77. Available at: https://doi.org/10.1145/3359786.

Eisenhardt, K.M. (1985) ‘Control: Organizational and economic approaches’, *Management Science*, 31(2), pp. 134–149.

European Commission (2021) *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 (Accessed: 21 April 2022).

Fawcett, T. (2006) ‘An introduction to ROC analysis’, *Pattern recognition letters*, 27(8), pp. 861–874.

Floridi, L. (2011a) ‘Children of the Fourth Revolution’, *Philosophy & Technology*, 24(3), pp. 227–232. Available at: https://doi.org/10.1007/s13347-011-0042-7.

Floridi, L. (2011b) ‘Enveloping the world: the constraining success of smart technologies’, in *CEPE 2011: Crossing Boundaries Ethics in Interdisciplinary and Intercultural Relations. CEPE*, Milwaukee Wisconsin: INSEIT (2011), p. 6.

Fügener, A. *et al.* (2021) ‘Will Humans-in-the-Loop Become Borgs? Merits and Pitfalls of Working with AI’, *Management Information Systems Quarterly (MISQ)-Vol*, 45.

Gartner (2021) *Case Study: How to Apply Ethical Principles to AI Models (Danish Business Authority)*. Case Study G00749866. Gartner. Available at: https://www.gartner.com/en/documents/4004387.

Gebru, T. *et al.* (2020) ‘Datasheets for Datasets’, *arXiv:1803.09010 [cs]* [Preprint]. Available at: http://arxiv.org/abs/1803.09010 (Accessed: 16 November 2020).

Government of Canada (2020) ‘canada-ca/aia-eia’. Government of Canada - Gouvernement du Canada. Available at: https://github.com/canada-ca/aia-eia (Accessed: 17 November 2020).

Government, T.D. (2019) *National Strategy for Artificial Intelligence*. Ministry of Finance and Ministry of Industry, Business and Financial Affairs, p. 74.

Gregor, S., Kruse, L.C. and Seidel, S. (2020) ‘Research Perspectives: The Anatomy of a Design Principle’, *Journal of the Association for Information Systems*, 21(6), pp. 1622–1652. Available at: https://doi.org/10.17705/1jais.00649.

Hevner, A.R. (2007) ‘A three cycle view of design science research’, *Scandinavian journal of information systems*, 19(2), p. 4.

Hill, K. (2020) ‘Wrongfully Accused by an Algorithm’, *The New York Times*, 24 June. Available at: https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html (Accessed: 12 November 2020).

Kaplan, A. and Haenlein, M. (2019) ‘Siri, Siri, in my hand: Who’s the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence’, *Business Horizons*, 62(1), pp. 15–25. Available at: https://doi.org/10.1016/j.bushor.2018.08.004.

Keil, M. *et al.* (2008) ‘The influence of checklists and roles on software practitioner risk perception and decision-making’, *Agile Product Line Engineering*, 81(6), pp. 908–919. Available at: https://doi.org/10.1016/j.jss.2007.07.035.

Lashbrook, A. (2018) *AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind*, *The Atlantic*. Available at: https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/ (Accessed: 12 November 2020).

Lebovitz, S., Levina, N. and Lifshitz-Assaf, H. (2021) ‘Is AI ground truth really “true”? The dangers of training and evaluating AI tools based on experts’ know-what’, *Management Information Systems Quarterly* [Preprint].

Li, J. *et al.* (2021) ‘STRATEGIC DIRECTIONS FOR AI: THE ROLE OF CIOS AND BOARDS OF DIRECTORS.’, *MIS Quarterly*, 45(3).

Lipton, Z.C. (2017) ‘The Mythos of Model Interpretability’, *arXiv:1606.03490 [cs, stat]* [Preprint]. Available at: http://arxiv.org/abs/1606.03490 (Accessed: 24 November 2020).

Lipton, Z.C. (2018) ‘The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery.’, *Queue*, 16(3), pp. 31–57. Available at: https://doi.org/10.1145/3236386.3241340.

Lou, B. and Wu, L. (2021) ‘AI ON DRUGS: CAN ARTIFICIAL INTELLIGENCE ACCELERATE DRUG DEVELOPMENT? EVIDENCE FROM A LARGE-SCALE EXAMINATION OF BIO-PHARMA FIRMS.’, *MIS Quarterly*, 45(3).

Lundberg, S.M. and Lee, S.-I. (2017) ‘A Unified Approach to Interpreting Model Predictions’, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc. (NIPS’17), pp. 4768–4777. Available at: http://arxiv.org/abs/1705.07874.

Merriam-Webster.com (2022) ‘Definition of RESPONSIBLE’, *Merriam-Webster.com*. Available at: https://www.merriam-webster.com/dictionary/responsible (Accessed: 26 April 2022).

Mitchell, M. *et al.* (2019) ‘Model Cards for Model Reporting’, *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* ’19*, pp. 220–229. Available at: https://doi.org/10.1145/3287560.3287596.

Moeini, M. and Rivard, S. (2019) ‘Sublating Tensions in the IT Project Risk Management Literature: A Model of the Relative Performance of Intuition and Deliberate Analysis for Risk Assessment’, *Journal of the Association for Information Systems*, 20(3). Available at: https://doi.org/10.17705/1jais.00535.

Morgenstern, O. and Von Neumann, J. (1944) *Theory of Games and Economic Behavior*. Princeton University Press.

Nagbøl, P.R., Krancher, O. and Müller, O. (Submitted) ‘Challenges and Practices in the Evaluation of AI Systems in the Public Sector’, in.

Nagbøl, P.R. and Müller, O. (2020) ‘X-RAI: A Framework for the Transparent, Responsible, and Accurate Use of Machine Learning in the Public Sector’, in *Proceedings of Ongoing Research, Practitioners, Workshops, Posters, and Projects of the International Conference EGOV-CeDEM-ePart 2020. EGOV-CeDEM-ePart 2020*, p. 9. Available at: http://dgsociety.org/wp-content/uploads/2020/08/CEUR-WS-Proceedings-2020_Full-Manuscript.pdf#page=273.

Nagbøl, P.R., Müller, O. and Krancher, O. (2021) ‘Designing a Risk Assessment Tool for Artificial Intelligence Systems’, in L. Chandra Kruse, S. Seidel, and G.I. Hausvik (eds) *The Next Wave of Sociotechnical Design*. Cham: Springer International Publishing, pp. 328–339.

Purao, S., Rossi, M. and Sein, M.K. (2010) ‘On integrating action research and design research’, in *Design research in information systems*. Springer, pp. 179–194.

Recker, J. *et al.* (2019) ‘Information Systems as Representations: A Review of the Theory and Evidence’, *Journal of the Association for Information Systems*, 20(6), p. 5.

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) ‘“Why Should I Trust You?”: Explaining the Predictions of Any Classifier’, *arXiv:1602.04938 [cs, stat]* [Preprint]. Available at: http://arxiv.org/abs/1602.04938 (Accessed: 7 December 2020).

Robbins, S. (2020) ‘AI and the path to envelopment: knowledge as a first step towards the responsible regulation and use of AI-powered machines’, *AI & SOCIETY*, 35(2), pp. 391–400. Available at: https://doi.org/10.1007/s00146-019-00891-1.

Russell, S.J. and Norvig, P. (2010) *Artificial intelligence: a modern approach*. 3rd ed. Upper Saddle River: Prentice Hall (Prentice Hall series in artificial intelligence).

Ryan, R.M. and Deci, E.L. (2000) ‘Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being’, *American psychologist*, 55(1), p. 68.

Sarker, S. *et al.* (2018) ‘Learning from first-generation qualitative approaches in the IS discipline: An evolutionary view and some implications for authors and evaluators (PART 1/2)’, *Journal of the Association for Information Systems*, 19(8), p. 1.

Sarker, S. *et al.* (2019) ‘The Sociotechnical Axis of Cohesion for the IS Discipline: Its Historical Legacy and Its Continued Relevance’, *MIS Q.*, 43(3), pp. 695–720. Available at: https://doi.org/10.25300/MISQ/2019/13747.

Schwartz, R. *et al.* (2022) ‘Towards a Standard for Identifying and Managing Bias in Artificial Intelligence’.

Secretariat, Treasury Board of Canada (2019) *Directive on Automated Decision-Making*. Available at: https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592 (Accessed: 17 November 2020).

Secretariat, Treasury Board of Canada (2020) *Algorithmic Impact Assessment (AIA)*, *aem*. Available at: https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html (Accessed: 17 November 2020).

Sein, M. *et al.* (2011) ‘Action Design Research’, *Management Information Systems Quarterly*, 35(1), pp. 37–56.

Søgaard, A. (2016) ‘Biases we live by’, *Nordisk Tidsskrift for Informationsvidenskab og Kulturformidling*, 5(1), pp. 31–35.

Spackman, K.A. (1989) ‘SIGNAL DETECTION THEORY: VALUABLE TOOLS FOR EVALUATING INDUCTIVE LEARNING’, in A.M. Segre (ed.) *Proceedings of the Sixth International Workshop on Machine Learning*. San Francisco (CA): Morgan Kaufmann, pp. 160–163. Available at: https://doi.org/10.1016/B978-1-55860-036-2.50047-3.

Suresh, H. and Guttag, J.V. (2020) ‘A Framework for Understanding Unintended Consequences of Machine Learning’, *arXiv:1901.10002 [cs, stat]* [Preprint]. Available at: http://arxiv.org/abs/1901.10002 (Accessed: 10 February 2021).

Sweeney, L. (2013) ‘Discrimination in Online Ad Delivery: Google Ads, Black Names and White Names, Racial Discrimination, and Click Advertising’, *Queue*, 11(3), pp. 10–29. Available at: https://doi.org/10.1145/2460276.2460278.

Tavory, I. and Timmermans, S. (2014) *Abductive analysis: Theorizing qualitative research*. University of Chicago Press.

Wallace, L., Keil, M. and Rai, A. (2004) ‘Understanding software project risk: a cluster analysis’, *Information & Management*, 42(1), pp. 115–125. Available at: https://doi.org/10.1016/j.im.2003.12.007.

Weller, A. (2019) ‘Transparency: Motivations and Challenges’, *arXiv:1708.01870 [cs]* [Preprint]. Available at: http://arxiv.org/abs/1708.01870 (Accessed: 7 December 2020).

Journal of the Association for Information Systems (2021) **22**(2), 3**25**-352
doi: 10.17705/1jais.00664

RESEARCH ARTICLE

ISSN 1536-9323

Sociotechnical Envelopment of Artificial Intelligence: An Approach to Organizational Deployment of Inscrutable Artificial Intelligence Systems

Aleksandre Asatiani[1], Pekka Malo[2], Per Rådberg Nagbøl[3], Esko Penttinen[4], Tapani Rinta-Kahila[5], Antti Salovaara[6]

[1]University of Gothenburg, Sweden, aleksandre.asatiani@ait.gu.se
[2]Aalto University School of Business, Finland, pekka.malo@aalto.fi
[3]IT University of Copenhagen, Denmark, pena@itu.dk
[4]Aalto University School of Business, Finland, esko.penttinen@aalto.fi
[5]The University of Queensland, Australia, t.rintakahila@uq.edu.au
[6]Aalto University School of Arts, Design and Architecture, Finland, antti.salovaara@aalto.fi

Abstract

The paper presents an approach for implementing inscrutable (i.e., nonexplainable) artificial intelligence (AI) such as neural networks in an accountable and safe manner in organizational settings. Drawing on an exploratory case study and the recently proposed concept of envelopment, it describes a case of an organization successfully "enveloping" its AI solutions to balance the performance benefits of flexible AI models with the risks that inscrutable models can entail. The authors present several envelopment methods—establishing clear boundaries within which the AI is to interact with its surroundings, choosing and curating the training data well, and appropriately managing input and output sources—alongside their influence on the choice of AI models within the organization. This work makes two key contributions: It introduces the concept of sociotechnical envelopment by demonstrating the ways in which an organization's successful AI envelopment depends on the interaction of social and technical factors, thus extending the literature's focus beyond mere technical issues. Secondly, the empirical examples illustrate how operationalizing a sociotechnical envelopment enables an organization to manage the trade-off between low explainability and high performance presented by inscrutable models. These contributions pave the way for more responsible, accountable AI implementations in organizations, whereby humans can gain better control of even inscrutable machine-learning models.

Keywords: Artificial Intelligence, Explainable AI, XAI, Envelopment, Sociotechnical Systems, Machine Learning, Public Sector

Hind Benbya was the accepting senior editor. This research article was submitted on February 29, 2020 and underwent three revisions.

1 Introduction

Advances in big data and machine-learning (ML) technology have given rise to systems using artificial intelligence (AI) that bring significant efficiency gains and novel information-processing capabilities to the organizations involved. While ML models may be able to surpass human experts' performance in demanding analysis and decision-making situations (McKinney et al., 2020), their operation logic differs dramatically from humans' ways of approaching similar problems. Rapid growth in the volumes of data and computing power available has made AI systems increasingly complex, rendering their behavior inscrutable and, therefore, hard for humans to interpret and explain

(Faraj et al., 2018; Stone et al., 2016). While the economic value of such systems is rarely in doubt, broader organizational and societal implications, including negative side-effects such as undetected biases, have started to cause concerns (Benbya et al., 2020; Brynjolfsson & McAfee, 2014; Newell & Marabelli, 2015). Thus, humans' ability to explain how AI systems produce their outputs, referred to as "explainability" (e.g., Rosenfeld & Richardson, 2016), has become a prominent issue in various fields.

The inscrutability of AI systems leads to a host of ethics-related, legal, and practical issues. ML models, by necessity, operate mindlessly, meaning that they approach the work from a single perspective, with no conscious understanding of the broader context (Burrell, 2016; Salovaara et al., 2019). For example, ML models cannot reflect on the ethics or legality of their actions. Accordingly, an AI system may exhibit unintended biases and discrimination after learning to consider inappropriate factors in its decision-making (Martin, 2019). Through such problems during the training stage and beyond, an organization may (wittingly or not) end up operating in a manner that conflicts with its values (Firth, 2019), with models being susceptible to biases and errors connected with vexing ethics issues, such as discrimination against specific groups of people. Designing models with solid ethics in mind could provide means to identify, judge, and correct such biases and errors (Martin, 2019), but all of this is impossible if the model's actions are inscrutable. Alongside ethics matters, there are legislative factors that impose concrete and inescapable requirements for explainability (Desai & Kroll, 2017). Public authorities often must honor requirements for transparency in their actions, and private companies may also be compelled to explain and justify, for instance, how they use customer data. The European Union's General Data Protection Regulation (GDPR) serves as a prominent example of recent legislative action that promotes the rights of data subjects to obtain an explanation of any decision based on data gathered on them (European Union, 2016).

Yet producing an explainable AI system may not always be feasible. Inscrutability takes many forms, linked to such elements as intentional corporate or state secrecy, technical illiteracy, and innate characteristics of ML models (Burrell, 2016). This multifaceted nature, combined with limitations on human logic, means there are no simple solutions to explainability problems (Edwards, 2018; Robbins, 2020). For example, some legal scholars maintain that the GDPR's provision for a right to explanation is insufficient and could result in meaningless "transparency" that does not actually match user needs (Edwards & Veale, 2017): while there may technically be an explanation for a given decision, this might not be understandable for the person(s) affected. Though approaches such as legal auditing (O'Neil, 2016; Pasquale, 2015), robust system design (Rosenfeld & Richardson, 2019), and user education may improve explainability in some cases, they are unidimensional and inadequate for tackling the fundamental challenges presented by the mindless operation of AI (Burrell, 2016). In an organizational setting, information-technology (IT) systems affect a broad spectrum of stakeholders who display differing, often sharply contrasting, demands and expectations (Koutsikouri et al., 2018). Explanation of AI agents' behavior is further complicated by the environment wherein AI development takes place, with various incumbent work processes, structures, hierarchies, and legacy technologies. These challenges have prompted calls for human-centered and pragmatic approaches to explainability (Mittelstadt et al., 2019; Ribera & Lapedriza, 2019). This invites us to approach explainability from a sociotechnical perspective to account for the interconnected nature of technology, humans, processes, and organizational arrangements, and thereby give balanced attention to instrumental and humanistic outcomes of technology alike (Sarker et al., 2019).

It is against this backdrop that we set out to address the following research question (RQ): *How can an organization exploit inscrutable AI systems in a safe and socially responsible manner?* Our inquiry was inspired by a desire to understand how organizations cope with AI models' inscrutability when facing explainability demands. The sociotechnical nature of the problem became apparent during the early phases of a research project at the case organization. We observed a need to integrate the organization's social side (people, processes, and organizational structures) with its technical elements (information technology and AI systems) synergistically if the organization wished to take advantage of a wider array of AI models, including some of the inscrutable models available. This pursuit involved two types of goals, explainability- and performance-oriented goals, which, in the case of AI implementation, present conflicting demands. Here, we draw on Sarker et al.'s (2019) concepts of instrumental and humanistic outcomes of information-system implementation to analyze the well-known tradeoff between explainability and accuracy. In its development of powerful AI models, the organization sought instrumentally oriented outcomes (better performance and greater efficiency) but also needed to cater to humanistic outcomes by making sure that the use of such models would not diminish human agency or harm people affected by the models' use. As we drilled down to precisely how the organization addressed both sets of desired outcomes, *envelopment* emerged as an illuminating lens for conceptualizing the various approaches.

This concept—envelopment of AI—has recently emerged as a potentially useful approach to cope with the explainability challenges described above (Robbins, 2020). It suggests that, by controlling the training data carefully, appropriately choosing both input and output data, and specifying other boundary conditions mindfully, one may permit even inscrutable AI to make decisions, because these specific precautions erect a predictable envelope around the agent's virtual maneuvering space. Thus far, however, envelopment has been illustrated in only a handful of contexts (e.g., autonomous driving, playing Go, and recommending apparel) and on a conceptual level only; thus, relatively limited insights have been presented for tackling explainability challenges in complex real-world organizations. To address this gap, we describe how envelopment is practiced in one pioneering organization that has embarked on utilizing AI in its operations, and we show that envelopment is fundamental to enabling an organization to use inscrutable systems safely even in settings that necessitate explainability. Further, we deepen the concept of envelopment by showing how it emerges via sociotechnical interactions in a complex organizational setting. With the empirical findings presented here, we argue that the sociotechnical envelopment concept has widespread relevance and offers tools to mitigate many challenges that stand in the way of making the most of advanced AI systems.

2 Review of the Literature and Theory Development

This section offers a review of lessons already learned from organizational AI implementations and their sociotechnical underpinnings. Also, we address the properties of good explanations and provide a more detailed picture of the envelopment concept.

2.1 A Sociotechnical Approach to Organizational AI

The recent emergence and proliferation of new generations of ML tools have reawakened interest in organizational AI research (Faraj et al. 2018; Keding 2021; Sousa et al. 2019). Like human intelligence, AI is notoriously difficult to define as a concept. For the purposes of our study, we follow Kaplan and Haenlein (2019) in defining AI as a "system's ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals" (p. 17). Complementing conceptual works, empirical studies on the topic have started to appear (e.g., Ghasemaghaei, Ebrahimi, & Hassanein, 2018; Salovaara et al., 2019; Schneider & Leyer, 2019). The papers have increasingly shifted the position of AI research from a largely technical one to a perspective encompassing the *social* component (Ågerfalk, 2020). Whereas the technical facet involves the information systems (IS) angle, IT infrastructure, and platforms, the social aspect brings in people, work processes, organizational arrangements, and cultural and societal factors (Sarker et al., 2019). Although scholars have discussed issues such as replacing humans with machines versus augmenting humans' capabilities (e.g., Davenport, 2016; Jarrahi, 2018; Raisch & Krakowski, in press), there is still little critical empirical work investigating the human aspects involved with deploying and managing AI in organizations (Keding, 2021).

Research on organizations' implementation and use of AI and other forms of automated decision-making has highlighted some recurrent patterns. First, AI's mindless and, thereby, error-prone nature necessitates careful control of the AI's agency and autonomy in the implementation. Humans can serve as important counterweights in this equation (Butler & Gray, 2006; Pääkkönen et al., 2020; Salovaara et al., 2019). The division of labor and knowledge between humans and AI can be arranged in various ways whereby organizations can balance rigidity and predictability against flexibility and creative problem-solving (Asatiani et al., 2019; Lyytinen et al., in press). Second, organizations' AI agents interact with many types of human stakeholders, each with a particular dependence on AI and distinct abilities to understand its operation (Gregor & Benbasat, 1999; Preece, 2018; Weller, 2019). Studies indicate that AI is rarely considered a "plug-and-play" technology and that an organization deploying it requires a clear implementation strategy that takes into account the wide spectrum of stakeholders (Keding, 2021). For instance, since the impact of AI's implementation varies greatly between stakeholders, decisions to decouple stakeholders from the process of designing, implementing, and using it increase the likelihood of unethical conduct and breach of social contracts, often leading to the systems' ultimate failure (Wright & Schultz, 2018).

Collectively, the literature on organizational AI shows how important it is for organizations to balance the risks associated with AI against the efficiency gains that may be reaped. These considerations also show that organizational AI deployment entails a significant amount of coordination and mutual adaptation between humans and AI and is thus inescapably a matter of sociotechnical organization design (Pääkkönen et al., 2020). Those advocating a sociotechnical approach maintain that attention must be given both to the technical artifacts and to the individuals/collectives that develop and utilize the artifacts in social (e.g., psychological, cultural, and economic) contexts (Bostrom et al., 2009; Briggs et al., 2010). In a corollary to this, taking a sociotechnical stance is aimed at meeting instrumental objectives (e.g., effectiveness and accuracy of the model or other

artifact developed) and humanistic objectives (e.g., engaging users and retaining employee skills) alike (Mumford, 2006).

Sarker et al. (2019) have reviewed the intricate ways in which the social and the technical may become interwoven such that neither the social nor technical aspects come to dominate. They show that this relationship is quite varied, and they demonstrate this by presenting examples of reciprocal as well as moderating influence, inscription of the social in the technical, entanglement, and imbrication. For instance, from the perspective of reciprocal influence, technology and organizational arrangements may be seen to coevolve throughout an IS implementation as they mutually appropriate each other (Benbya & McKelvey, 2006). From the sociomaterial perspective of imbrication, in turn, humans and technologies are viewed as agencies whose abilities interlock to produce routines and other stable emergent processes.

2.2 Challenges of Inscrutable AI

As noted in the introduction, complex AI models often promise better performance than simple ones, but such models also tend to lack transparency, and their outputs can be hard or even impossible to explain. Writings on AI explainability often employ the interrelated concepts of transparency, interpretability, and explainability in efforts to disentangle the threads of this problem. *Transparency* refers to the possibility of monitoring AI-internal operations—e.g., tracing the paths via which the AI reaches its conclusions (Rosenfeld & Richardson, 2019; Sørmo et al., 2005). Its opposite is opacity, a property of "black-box" systems, which hide the decision process from users and sometimes even from the system's developers (Lipton, 2018). The two other concepts—*interpretability* and *explainability*—refer to the AI outputs' understandability for a human (e.g., Doshi-Velez & Kim, 2017; Miller 2019). On occasion, the terms are used interchangeably (e.g., Došilović et al., 2018; Liu et al., 2020) while sometimes authors employ separate definitions. Often, interpretability has strong technical connotations while explainability is more human centered in nature and hence a more sociotechnically oriented concept.

Many of the more traditional AI models, such as linear regression, with its handling of only a limited number of known input variables, and decision trees, which can display the if-then sequence followed, are considered explainable. However, more and more of today's AI models are so complex that explainability is rendered virtually impossible. For instance, when a traditional decision-tree model is "boosted" via a machine-learning technique called gradient boosting, its performance improves but its behavior becomes far more difficult to explain. Other examples of highly accurate models that lack explainability are deep and recurrent neural networks, complexly layered computing systems whose structure resembles that of the biological networks of a brain's neurons. Then, one deems them *inscrutable* (Dourish, 2016; Martin, 2019), referring to situations wherein the system's complexity outstrips practical means of analyzing it comprehensively. A recent open-domain chatbot developed at Google, which has 2.6 billion free parameters in its deep neural network (Adiwardana et al., 2020), is an extreme example of an AI system whose inner workings are inscrutable for humans even if they are transparent.

Unrestrained use of inscrutable systems can be problematic. Humans interacting with such systems are unable to validate whether the decisions made by the system correspond to real-world requirements and adhere to legal or ethics norms (Rosenfeld & Richardson, 2019). The issue is far from academic; after all, reliance on inscrutable systems could lead to systematic biases in decision-making, completely invisible to humans interacting with or affected by the system (Došilović et al., 2018).

In consequence, organizations intending to deploy AI systems face an *explainability-accuracy tradeoff* (Došilović et al., 2018; Linden et al., 2019; London, 2019; Martens et al., 2011; Rosenfeld & Richardson, 2019). On the one hand, complex models with greater flexibility, such as deep neural networks, often yield more accurate predictions than do simple ones such as linear regression or decision trees. On the other hand, simple models are usually easier for humans to interpret and explain. The tradeoff that seems to exist between explainability and accuracy forces the design to prioritize one over the other: an organization wishing to reduce the risks associated with inscrutable AI must settle for AI models with a high degree of explainability. Figure 1 illustrates this tradeoff, following depictions by Linden et al. (2019) and Rosenfeld and Richardson (2019).

One approach recently introduced to address the risks brought by black-boxed systems is envelopment. In recognition of its potential for managing the explainability-accuracy tradeoff, the following section delves into the suggestions that researchers have presented in relation to this approach.

2.3 Envelopment

As noted above, we identified envelopment (Floridi, 2011; Robbins, 2020) as a suitable sensemaking concept when examining the domain of organizational AI development. In its original context in robotics, a *work envelope* is "the set of points representing the maximum extent or reach of the robot hand or working tool in all directions" (RIA Robotics Glossary, 73; cited by Scheel, 1993, p. 30).

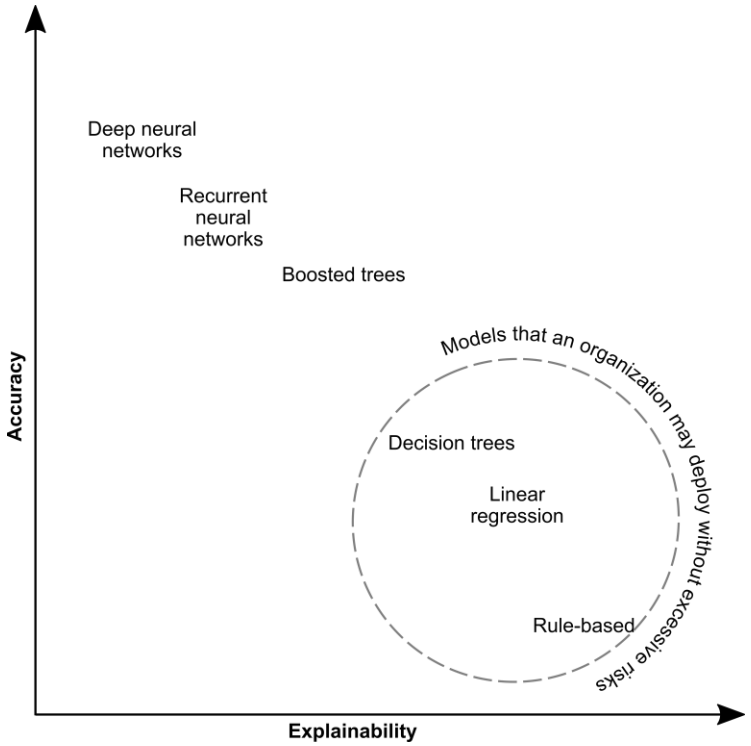


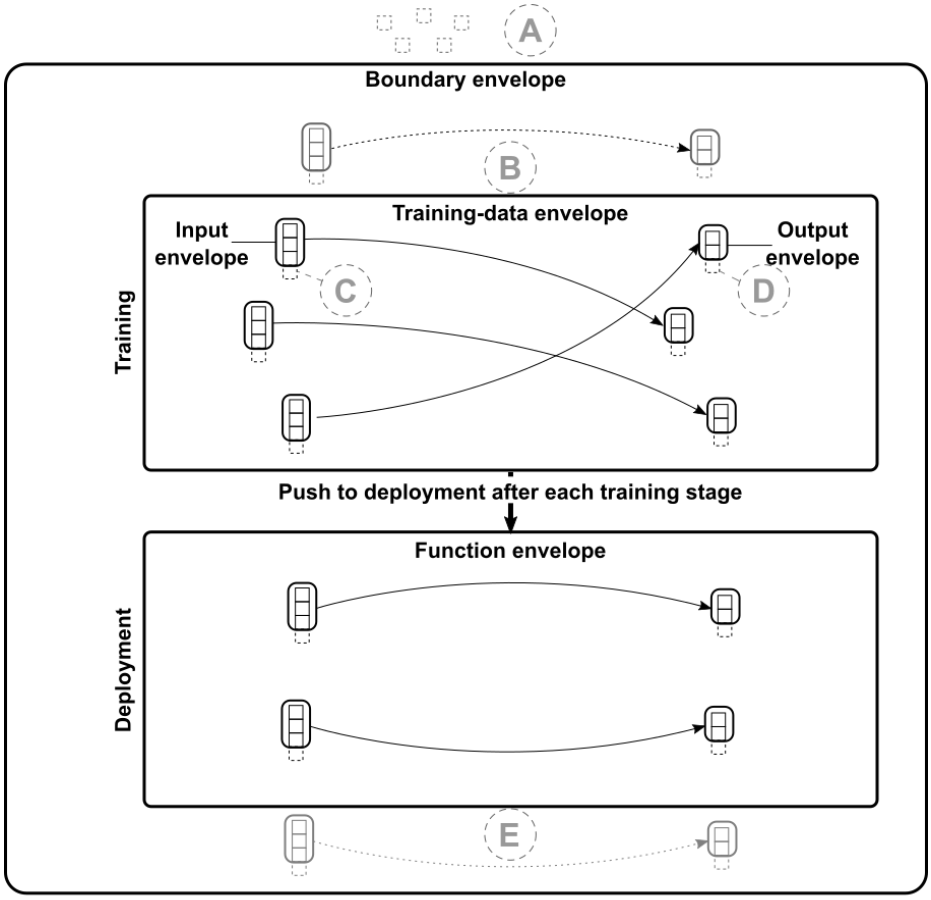
Figure 1. The Explainability-Accuracy Tradeoff

Robots' work envelopes, often presented as shaded regions on factories' floor maps and as striped areas on factory floors, are a practical solution for fulfilling what is known as the "principle of requisite variety" (Ashby, 1958)—i.e., meeting the requirement that the number of states of a robot's logic be larger than the number of environmental states in which it operates. If a robot acts in an environment whose complexity exceeds its comprehension, it will pose a risk to the surroundings. Work envelopes—areas that no other actors will enter—can guarantee that the physical environment of the robot is simplified sufficiently (i.e., that the number of possible states of the environment is reduced enough). Through this modification, the robot can handle those states that still need to be controlled, thereby fulfilling the principle of requisite variety. In addition to physical parameters, a robot's envelope may be specified by means of time thresholds, required capabilities/responsibilities, and accepted tasks (McBride & Hoffman, 2016, p. 79). These parameters are dynamic: when a robot faces new problems, the envelope parameters are adjusted to accommodate what the requisite variety now entails (p. 81).

Our research is a continuation of work wherein this concept has been applied to cases that involve humans and nonphysical work performed by AI agents. In this context, the envelope is not physically specified but relates to the realm of information processing. This domain change notwithstanding, there remains a need for collaboration with a human partner who maintains the envelope and thus guarantees the safety and correctness of the AI's operation (Floridi, 2011). Also, the underlying principle of requisite variety continues to persist, meaning that the AI should not be used for tasks it cannot master and that it should not be trained with data irrelevant to the tasks. Such undesired effects—"excessive risks" in Figure 1—can manifest themselves in several forms, among them erroneous input-action mappings, ethics dilemmas that an AI agent should not be allowed to tackle by itself, and behaviors that demonstrate bias (e.g., Robbins, (2020). Even if the realization of such risks does not impair the financial bottom line or operations' efficiency, it can result in problematic humanistic outcomes. For example, an AI system that processes job applications to identify the most promising candidates may increase the efficiency of an HR department, and consistently identify candidates that meet requirements for the position. At the same time, the system could consistently discriminate against certain groups of applicants who would otherwise qualify because of a bias in an underlying model. In such scenarios, AI actions may not impact the bottom line of the company, at least in the short term, but may be nevertheless problematic.

Envelopment can be advanced via several methods. Figure 2 presents our interpretation of the five methods that Robbins (2020) articulated. We summarize them below, then build on them in relation to our study. *Boundary envelopes* represent the most general of the

envelopment methods. The envelope delineates *where* the AI operates—for example, only analyzing images of human faces photographed in good lighting conditions. An AI model enveloped in this way will not encounter any tasks other than those carefully designated for it (condition A in Figure 2). Robbins (2020) takes the design of a robot vacuum cleaner as an example. Its boundary envelopment mechanism means that the robot does not need to be able to avoid threats that never exist in indoor domestic spaces (e.g., puddles of water). The benefit of boundary envelopment is that the AI does not need to incorporate methods to recognize whether the agent is being made to operate in scenarios that extend beyond its ability to comprehend the surroundings (i.e., requisite variety).



Legend:

An input or output vector of data. One vector element (dashed gray line) has been enveloped out and is not used in the model. Rectangles with bold stokes denote envelopes.

Examples of what the envelope scope excludes:

A Events and states-of-affairs in the world that the model does not need to "know" about. [Boundary]

B Input–output pairs that could be used in training data but are suspected of bias, errors, or represent cases for which not enough data exists yet and the model should not be allowed to learn from. [Training-data envelope]

C Input sources that would provide low-quality information. [Input envelope]

D Outputs that a model could provide but that are biased, not needed, or redundant. [Output envelope]

E Purposes for which the trained model will not be used (e.g., for ethics reasons), even if it would be capable of accurate performance. [Function envelope]

Figure 2. Illustration of AI Envelopment Methods Suggested by Robbins (2020)

Among the other envelopment methods are three that refer to the notion of *what content* the AI will manipulate (Robbins, 2020). The first of them is the *training-data envelope*, related to the curation of the correct input-output mappings with which the AI model is trained. Robbins cites biases and other representativeness problems ("B" in Figure 2) as particularly likely to propagate or uphold societal stereotypes if the envelope is not handled properly. *Input envelopes*, in turn, address the technical details of inputs to the AI. For example, in Robbins's example, a recommendation AI uses various pieces of weather and user data (e.g., temperature, real-time weather status, and the user's calendar) to produce clothing recommendations (e.g., the suggestion to wear a raincoat). For good results, the data should arrive from sources that are high quality, noise free, and of appropriate granularity. Input envelopment limits input channels to those that meet appropriate criteria in this regard and prevents poorly understood sources from affecting the model's behavior. The third envelopment method in the "what" category is the use of *output envelopes*. These define the set of actions that may be performed within the realm of the AI's operation. In the case of an autonomously driving car, the outputs might be specified as speeding up, turning the wheels, and braking. Even if speeding would be technically possible and sometimes useful, it presents risks to passengers and other traffic. Therefore, that output is enveloped out of an autonomous car's actions. In Figure 2, "C" and "D" illustrate the input- and output-envelopment methods described above.

The fifth and final method, use of a *function envelope*, addresses the question of *why* the AI exists and what goals and ethics it has been designed to advance. This category of envelopment is applied to limit the AI's use for malicious or otherwise problematic purposes, even in cases wherein it operates correctly. For example, the functions of conversational home assistants such as Echo or Alexa are limited to only a small set of domestic activities to avoid privacy infringements (Robbins, 2020). Such filtering out of functions is denoted as "E" in Figure 2.

Robbins suggests that with such variety of envelopment methods available, one can either overcome some problems connected with black-box AI or neutralize their effects. Our work is thus informed by the envelopment concept, and we consider its applicability in complex and emergent sociotechnical settings. In particular, we maintain that humans play an important role in an AI agent's envelopment and in how it is organized by striving to guarantee that the AI does not face tasks it is unable to process or interpret correctly—where the problems exceed its requisite variety (e.g., Salovaara et al., 2019). Next, we report on our case study.

3 The Case Study: Machine Learning in a Governmental Setting

To examine how an organization may tackle explainability challenges, we conducted an exploratory case study at a government agency that actively pursues the deployment of AI via several ML projects. We selected a case organization with both extensive capabilities to develop AI/ML tools and a commitment to accountability and explainability.

3.1 The Study Setting

The Danish Business Authority (DBA) is a government entity operating under the Ministry of Industry, Business, and Financial Affairs of Denmark. It has approximately 700 employees and is based in Copenhagen, with satellite departments in Silkeborg and Nykøbing Falster. The authority is charged with a wide array of core tasks related to business, clustered around enhancing the potential for business growth throughout Denmark. The DBA maintains the digital platform VIRK, through which Danish companies can submit business documents and that allows the DBA to maintain an online business register (containing approximately 809,000 companies, with roughly 812,000 registrations in all and together filing about 292,000 annual statements per year). The DBA has maintenance and enforcement remits related to laws such as Denmark's Companies Act, Financial Statements Act, Bookkeeping Act, and Act on Commercial Foundations. In the past, the DBA also collaborated with Early Warning Europe (EWE)—a network established to help companies and entrepreneurs across Europe—to produce support mechanisms for companies in distress. The ML projects analyzed in our study are related to the DBA's core tasks—for example, understanding VIRK users' behavior and checking business registrations and annual statements for mistakes and evidence of fraud.

The idea of using ML at the DBA originated in 2016. The agency embarked on AI-related market research, which culminated in several data-science projects and the establishment of the Machine Learning Lab ("the ML Lab" from here on) in 2017. One factor creating the impetus for establishing the ML Lab was tremendous growth in the quantities of various types of documents processed by the DBA. Rather than engage and rely on external consultants, the DBA opted to hire its own data engineers and data scientists. The main reasons for this in-house approach were cost-management concerns and a desire to retain relevant knowledge within the agency. Creating ML solutions internally by combining technologies such as Neo4j graph database management, Docker containers, and Python offers a better fit for the organization than commercial off-the-shelf solutions. Also, the ML Lab's role is

restricted largely to experimentation and development surrounding proof-of-concept models. If a solution is deemed useful and meets the quality criteria set, its deployment is offloaded to external consulting firms, which then put the model into production use. This decision was primarily based on DBA culture, in which vendors take responsibility for the support and maintenance functions related to their code: the ML models follow the same governance as other IT projects within the DBA.

Hence, DBA operations related to ML are divided between two main entities: a development unit (the ML Lab) and an implementation unit (external consultants). The ML Lab's role is to collaborate closely with domain experts (hereafter "case workers") to develop functional prototypes as part of a proof of concept. The lab's main objective is to prove that the problems identified by the case workers can be solved by means of ML. In combination, the proof of concept and documentation such as the evaluation plan form the foundation for the DBA steering committee's decision-making on whether to forward the model to the implementation unit. Different stakeholders are accountable for different parts of the process. The ML Lab is responsible for developing the prototype, and the case workers provide domain knowledge to the lab's staff as that prototype is developed. The case workers also answer for the ML models' operational correctness, being charged with evaluating each model and with its retraining as needed. The steering committee then decides which models will enter production use and when. Finally, the implementation unit is accountable for implementing the model and overseeing its technical maintenance.

3.2 Data Collection

Interviews and observations at the DBA served as our main data sources. We used purposive sampling (Bernard, 2017) and selected the case organization by applying the following criteria. The organization needed to have advanced AI and ML capabilities, in terms of both resources and know-how. It also had to be committed to developing explainable systems. Finally, the researchers needed access to the AI/ML projects, associated processes, and relevant stakeholders. The last criterion was especially important for giving us a broader perspective on the projects and for enabling the verification of explainability claims made by the informants. The DBA met all of these criteria.

To gain access to the DBA, we used the known-sponsor approach (Patton, 2001): we had access to a senior manager at the DBA working with ML initiatives within the organization, who helped us arrange interviews at the early stages of data collection. Piggybacking on that manager's legitimacy and credibility helped us establish our legitimacy and credibility within the DBA from the start (Patton, 2001). In addition, one of the authors had a working relationship with the organization at the operations level, allowing us to arrange interviews further along in the data-collection work. This helped us to establish mutual trust with the informants and prevented us from being seen as agents of the upper management.

We collected and analyzed data in a four-stage iterative process (presented in Table 1), in which the phases overlapped and earlier stages informed subsequent stages. To prevent elite bias, we sought to interview a wide range of DBA employees with varying tenure at several levels in the hierarchy (Miles et al., 2014; Myers & Newman, 2007). Phase 1 was explorative in nature. Its purpose was to establish research collaboration and create a picture of the DBA's current and future ML projects and visions from a data-science and case-work perspective. The second phase was aimed at gaining in-depth understanding of the DBA's various ML projects and the actors involved. In this phase, we focused on the ML Lab and its roles and responsibilities in the projects, along with explainability in relation to ML. Then, in Phase 3, we interviewed all ML Lab employees as well as two case workers who acted in close collaboration with the lab. The final phase involved validating the interpretations from our analysis and obtaining further insight into the technical infrastructure supporting the lab.

We conducted semi-structured interviews in all phases, taking place from August 2018 to October 2020. Initial impressions are important for establishing trust between researchers and informants (Myers & Newman, 2007); hence, we always presented ourselves as a team of impartial researchers conducting an academic study. At the start of each interview, we explained the overall purpose of the study and our reasons for selecting the informant(s) in question to participate. We promised anonymity and confidentiality to all the informants and asked for explicit consent to record the interviews. Also, we explained the right to withdraw consent at any time during the interview or after it, up to the time of the final publication of a research article. We made sure to address any concerns the informants expressed about the procedure and answered all questions.

The interviews were conducted in English, with one of the authors, a native Danish speaker, being present for all of them and clarifying terminology as necessary. In addition, the informants had the opportunity to speak Danish if they so preferred. The choice of English as the primary language was made in consideration of the fact that most members of the research team did not speak Danish, whereas all informants were highly proficient in English. Though we recognize potential downsides to conducting interviews in a language that is not native to the interviewees, we accepted the remaining risk for the sake of enabling the whole research team to be involved in the data-collection process and data analysis. All interviews were audio-recorded and transcribed, yielding 167,006 words of text.

Table 1. The Four Phases of Gathering the Data

Phase number, theme, and date range	Method and duration	Informant's pseudonym and role	Focus of outcomes
1. ML projects overall, August-September 2018	Group interview (105 minutes)	James (ML Lab team leader / chief data scientist); Mary (chief consultant)	Responsibilities of the DBA; organization structure
2. ML Lab functions, October 2018 to January 2019	Personal interview (90 minutes)	James	The role of explainability in ML projects; allocation of tasks among stakeholders (the ML Lab, implementation unit, and case workers)
	Group interview (83 minutes)	David; John (both Early Warning Europe external case workers)	
	Personal interview (70 minutes)	Daniel (an internal case worker)	
	Personal interview (59 minutes)	Steven (a data scientist at the ML Lab)	
	Personal interview (51 minutes)	Mary	
	Personal interview (116 minutes)	James	
3. Explainability in ML projects, September 2019	Personal interview (51 minutes)	Steven	Practical means to address explainability issues; the sociotechnical environment of model development
	Personal interview (54 minutes)	Thomas (a data scientist at the ML Lab)	
	Personal interview (50 minutes)	Linda (a data scientist at the ML Lab)	
	Personal interview (48 minutes)	Michael (a data scientist at the ML Lab)	
	Personal interview (52 minutes)	Mark (a data scientist at the ML Lab)	
	Personal interview (53 minutes)	Joseph (a data scientist at the ML Lab)	
	Personal interview (54 minutes)	Jason (a team leader at the ML Lab)	
	Personal interview (48 minutes)	Susan (a data scientist at the ML Lab)	
	Personal interview (62 minutes)	William (an internal case worker)	
	Personal interview (54 minutes)	Daniel	
4. Verification of interpretations from analysis, December 2019 to October 2020	Personal interview (55 minutes)	Jason	Validation of interpretations via interview feedback and an assessment exercise involving mapping via project templates
	Assessment exercise (time N/A)	Steven; Mary; Thomas; Linda; Michael; Mark; Joseph; Jason; Susan	
	Personal interview (27 minutes)	Jason	
	Personal interview (32 minutes)	Steven	
	Personal interview (49 minutes)	Daniel	

In addition to interviews, we employed participant observation and document analysis. Hand-written field diaries kept by the Danish-speaking author provided background information. These go back to September 2017, when he became involved with ML at the DBA. Covering work as an external consultant and then a collaborative PhD student funded equally by the IT University of Copenhagen and the DBA, the diary material comprises observations, task descriptions, and notes taken at meetings. The diaries extended over the full duration of our research period, including the time when most ML projects were either very early in their development or had not even begun. Accounting for approximately every other workday at the DBA, the doctoral student's observations give a realistic view of day-to-day work life at the case organization. We used the field diaries for memory support, to fill gaps in the interview data, and as a reference for basic information about key informants, organization structure, and organizational processes and work practices. In addition, the diaries helped to corroborate some claims made by informants. Similarly, the document analysis addressed the entire time span of interest. This work included analyzing documentation and user stories extracted from the DBA's Jira system, a project management tool. The document analysis also extended to accessing the DBA's Git repository (used in version control) and verifying which model was

applied in each project. In addition, the collaborative doctoral researcher had access to a personal email account at the organization and could search old conversations and start new ones if decisions made during ML projects needed further explanation. Finally, to verify the interpretations arising in the course of the authors' analysis, we asked the ML Lab data scientists to fill in an outline document for each of the ML projects alongside the authors in an assessment exercise. This exercise produced an *input–ML-model–output* framework that allowed us to verify the ML projects' fundamentals and establish uniform project descriptions characterizing, for example, the data fed into the model, the type of ML model employed, and the nature of the output produced. Appendix A provides a summary of this framework.

3.3 Data Analysis

Overall, our analysis approach can be considered abductive: it began as inductive but was later informed by a theoretical lens that emerged as a suitable sensitizing device (Sarker et al., 2018; Tavory & Timmermans, 2014). We coded all interview data in three stages, utilizing coding and analysis techniques adopted from less procedure-oriented versions of grounded theory (Belgrave & Seide, 2019; Charmaz, 2006). In practice, this entailed relying on constant comparative analysis to identify initial concepts. The processes of data collection and analysis were mutually integrated (Charmaz, 2006), constantly taking us between the specific interview and the larger context of the case organization (Klein & Myers, 1999). Later, we linked the emerging concepts to higher-level categories. Similarities can be seen between our approach to using elements of grounded theory for qualitative data analysis and methods established in earlier IS studies (e.g., Asatiani & Penttinen, 2019; Sarker & Sarker, 2009).

The three stages of coding produced concepts (first-order constructs), themes (second-order constructs), and aggregate dimensions (see Appendix C), paralleling the structure proposed by Gioia, Corley, and Hamilton (2013). In the first stage, we performed open coding with codes entirely grounded in our data. This involved paragraph-by-paragraph coding, using in vivo codes taken directly from the informants' discourse (Charmaz, 2006) with minimal interpretation by the coders. For example, the extract: "There would be a guidance threshold. Actually, no. For this model, there would be some guidance set by us, yeah. And then case workers will be free to move it up and down" was assigned two codes: "case workers' control thresholds" and "guidance threshold." Two of the authors performed open coding independently, after which the two sets of codes were revisited, compared, and refined. Conceptually similar codes were merged into the set of concepts.

In the second stage, we analyzed the results from the open coding and started to look for emerging themes. We iterated between the open codes and interview transcripts, coding data for broader themes connecting several concepts (axial coding). While these themes were at a higher level than the in vivo codes from the first stage, they still were firmly grounded in the data. All the authors participated in this stage, which culminated in the codes identified being compared and consolidated to yield the second-order constructs—the themes.

In the third stage, we applied theoretical coding to our data. That term notwithstanding, the goal for this stage was not to validate a specific theory. Rather, we wanted to systematize the DBA's approaches to tackling explainable AI challenges where building a transparent system was not an option. For this, the envelopment framework of Robbins (2020) served as a sensitizing lens to help us organize the themes that emerged in the second stage of analysis. The decision was data-driven—we had not anticipated finding such strong focus on envelopment at the case organization, but the first two stages of analysis inductively revealed that the DBA's strategy resembled an envelopment rather than a method whereby the DBA would attempt to guarantee explainability in all of its AI model implementations. All authors participated in this stage of the work, performing coding independently. Then, the codes were compiled, compared, and synthesized into a single code set.

4 Findings

Our findings draw from the DBA ML Lab's work in eight AI projects, denoted here as Auditor's Statement, Bankruptcy, Company Registration, Land and Buildings, ID Verification, Recommendation, Sector Code, and Signature (see Appendix A for project details). While every project had a distinct purpose, each was aimed at supporting the DBA's role in society as a government business authority. At the time of writing this paper, many of these projects had been deployed and entered continuous use. The DBA had faced pressure to be highly efficient while remaining a transparent and trustworthy actor in the eyes of the public, and AI-based tools represented an efficient alternative to the extremely resource-intensive fully human-based processing of data. At the same time, the use of such tools presented a risk of coming into conflict with the DBA's responsibility to be transparent. To situate the set of envelopment methods employed by the DBA in this context, we begin by analyzing the DBA's viewpoint on requirements for the AI systems to be used in the agency's operations. This sets the stage for discussing the envelopment methods that the DBA developed to address the challenges of the explainability-accuracy tradeoff (see Figure 1) introduced by its development of ML solutions.

4.1 Requirements for AI at the DBA

Our interviews showed that, given the drive to improve its operations by using AI models, the DBA must devote significant attention to making sure instrumental outcomes do not come bundled with ignoring humanistic ones. Two factors have shaped the organization's quest to find balance in terms of the explainability-accuracy tradeoff: its positions as a public agency and diverse stakeholder requirements.

First, as a public agency, the DBA has significant responsibility for making sure that its decisions are as fair and bias-free as possible. Recent discussion surrounding regulations such as the GDPR has brought further attention to the handling of personal data and to citizens' rights to explanation. These reasons have impelled the DBA to be sure that the organization's ML solutions respond to explanability requirements sufficiently. This comment from a chief consultant on the DBA annual statements team, Mary, addresses transparency's importance:

> *I think in Denmark, generally, we have a lot of trust towards systems …. I'm very fond of transparency. I think it's the way to go that it's fully disclosed why a system reacts [the way] it does. Otherwise, you will feel unsafe about why the system makes the decisions it does … For me, it's very important that it's not a black box.*

Still, the DBA has ample opportunities to benefit from deploying AI in its operations, in that it has access to vast volumes of data and boasts proactive case workers who are able to identify relevant tasks for the AI. Sometimes inscrutable models clearly outperform explainable ones, so the agency has a strong incentive to seek ways of expanding the range of AI models that are feasible for its operations, in pursuit of higher accuracy and better performance. However, it needs to do so without incurring excessive risks associated with inscrutable models:

> *If the output of the algorithm is very bad when using the [explainable] models and we see a performance boost in more advanced or black-box algorithms, we will use [the more advanced ones]. Then, we will afterwards check like "okay, how to make this transparent, how to make this explainable…"* (Steven, ML Lab)

Secondly, the quest for explainable AI is made even more complex by the diversity of explanation-related requirements among various DBA stakeholders. The internal stakeholders comprise several distinct employee categories, including managers, data scientists, system developers, and case workers. Externally, the DBA interacts with citizens and the companies registered in Denmark, as well as with the IT consulting firms that maintain the agency's AI models deployed in the production environment.

Each of these stakeholders requires a specific kind of explanation of a given model's internal logic and outputs. While an expert may consider it helpful to have a particular sort of explanation for the logic behind the model's behavior, that explanation may be useless to someone who is not an expert user. For a nonexpert user, a concise, directed, and even partially nontransparent explanation may have more value than a precise technical account. David, a case worker with Early Warning Europe, offered an example: "When [a data scientist] explained this to us, of course it was like the teacher explaining … brain surgery to a group of five-year-olds."

These two factors together explain why expanding the scope of candidate models can pose problems even if more accurate models are available and technically able to be brought into use. Because of the different stakeholders' various needs, a suitable level of explainability is hard to reach. Therefore, approaches that could broaden the range of models—visualized as a circle with a dashed outline in Figure 1—are sorely needed.

Our findings indicate that envelopment offers a potential solution to the explainability-accuracy tradeoff. With a variety of envelopment methods, the risks of inscrutable AI may be controlled in a manner that is acceptable to the different stakeholders, even when technical explanations are not available. As Steven stated:

> *Often, we [are] able to unpack the black box if necessary and unpack it in a way that would be more than good enough for our case workers to understand and to use it and also for us to explain how the model came to the decision it did.*

Next, we discuss how the DBA has succeeded in this by enveloping its AI systems' boundaries, training data, and input and output data. We then consider our findings with regard to the connection between the choice of AI model and envelopment.

4.2 Boundary Envelopment

The notion of boundary envelopment suggests that an AI agent's limits can be bounded by well-defined principles that demarcate the environment within which it is allowed to process data and make decisions. One example of boundary envelopment at the DBA is the document filter implemented in the Signature project. It filters out images that are not photographs of a paper document. The need for such a filter was identified when an external evaluator tested the model with a picture of a wooden toy animal and the model judged the image to be a signed document because it

was operating beyond its intended environment. Having not been trained to analyze images other than scans and photographs of black-and-white documents, the model returned unpredictable answers. By limiting the types of input images to ones that the model had been trained to recognize, the filter created in response acts as a boundary envelope guaranteeing the requisite variety for the AI model that constitutes the next element in the information-processing pipeline. Thus, the AI model was enveloped in two ways: technically, via the development of a filter for its input data, and socially, via a change in workflow, whereby documents now undergo screening before they are assessed for completeness.

Both social and technical dimensions of envelopment were evident also in other instances at the case organization. The following quotes exemplify how the DBA orchestrates its AI agents' boundary-creation work and makes sure that its AI solutions speak to very different stakeholders' concerns. To ensure that AI systems' abilities and limitations are controlled and therefore enveloped, the DBA decided to divide its AI development into a process of incremental stages by introducing multiple small-scale solutions, each dedicated to a certain set of relatively simple and well-defined actions. The following comment summarizes this method:

> *Well, I'm working at an organization where, luckily, the management wants us to develop results fast or fail fast, so they are happy with having small solutions put into production [use] rather than having large projects fail …. We decided to use an event-driven architecture, because when dealing with complex systems, it's better to allow an ordered chaos than try to have a chaotic order. By having an event-driven architecture, you can rely on loosely coupled systems, and by having sound metadata it will help you create order in the chaos of different systems interacting with the same data.* (Jason, ML Lab)

Thus, from a purely technical angle, the event-driven architecture and loosely coupled systems constitute a technique in which the various components of a larger architecture operate autonomously and malfunctions are limited to local impacts only. For instance, erroneous decisions are less likely to be passed onward to other systems, and if this somehow does occur, the loose coupling allows the DBA to rapidly curb the failure's escalation. Each component is therefore operating in its own envelope, and larger envelopes are created to control AI components' operation as a network.

However, as highlighted by the reference above to envelopes that meet various stakeholders' needs, boundary envelopes do not serve a technical purpose alone. The following extract from the data shows how important the understanding of these boundaries is for those human stakeholders that are tasked with judging the correctness of the model's operation when, for example, the complexity of the environment exceeds the model's comprehension capability:

> *We have around 160 rules. We have technical rules that look into whether the right taxonomy is being used, whether it is the XBRL format, and whether it is compliant. We also have business rules. For example, do assets and liabilities match? Some rules only look at technical issues in the instance report. Other rules are what we called full-stop rules … filers are not allowed to file the report until they have corrected the error. We also have more guidance[-type] rules, where we say, "It looks like you're about to make a mistake. Most people do it this way. Are you sure you want to continue filing the report?" And then [users] can choose whether to ignore the rule [or not].* (Mary)

In addition to the technical issues connected with accounting for multiple kinds of failure, the comment attests to boundary envelopes' social dimension. The boundaries are clearly explained to internal users at the DBA, who can overrule the models if necessary. Moreover, customer-facing models operate within an environment that has clearly defined rules constraining their operation. Wherever nonexpert employees interact directly with a model, these rules are explained to them, and the human always has the power to ignore the models' recommendations if they seem questionable.

Thus, importantly, for every customer-facing AI model at the DBA, the final boundary envelope is a human. A decision suggested by an AI model is always verified by a case worker. In simple terms, human rationality creates a boundary that envelops the model's operation. This serves a dual purpose: it denies any model the power to make unsupervised decisions while it also makes certain that every DBA decision is compliant with legal requirements. According to Jason:

> *The agency can be taken into court when we dissolve a company, when we end a company [forceably] by means of the law. And we, in that situation, in court, will have to provide … full documentation of why that decision has been made. Now, legally speaking, as soon as there's a human involved, as there always is, we always keep a human in [the] loop, [so we are on the safe side]. In that context, it's only legally*

> *necessary to present that human's decision. But we want to be able to explain also decision support, so that's why we need explainability in our model and information chain. Explainability, on the microscale, is beneficial to understanding [the] organization on a sort of macroscale.*

In other instances, expert case workers are allowed to set thresholds for the model in question, to make certain it produces the most useful and precise recommendations. This has a knock-on effect in facilitating DBA workers' acceptance of the relevant model:

> *For some [of our] models, there would be some guidance threshold set by us. And then case workers are free to move it up and down.* (Susan, ML Lab)

> *The ability to "mute" a model or change the threshold has been a major cultural factor in [the] business adaptation of this technology.* (Jason)

In summary, envelopment of boundaries involves both resolving technical issues (understanding the limits of the model's abilities, etc.) and addressing social factors (providing the various stakeholders with sufficient explainability and, thereby, affording trust in the model's accuracy, etc.).

4.3 Training-Data Envelopment

The crucial importance of the data used in AI systems' training is widely acknowledged in the AI/ML community. If trained on different data sets, two models with otherwise identical structure produce vastly different outputs (Alpaydin, 2020; Robbins, 2020). Accordingly, close control of the training data and the training process form an important aspect of envelopment: if the spectrum of phenomena that the training data represent is considered with care, one can better understand what the model will—and will not—be able to interpret.

Since the DBA wants to avoid any undesired outcomes from an uncontrolled model roaming freely on a sea of potentially biased training data, the organization has decided to maintain full control over the learning process; thus, it abstains from using online-learning models, which continue learning autonomously from incoming data. This aids the DBA in protecting its systems from the unintended overfitting and bias that less tightly controlled training data could more easily introduce. The training may be implemented in a controlled, stepwise manner:

> *We have taken a conscious decision not to use [online-learning] technologies, meaning that we train a model to a certain level and then we accept that it will not become smart until we retrain it.* (Jason, ML Lab)

Avoidance of models that learn "on the fly" has a downside in that models' training at the DBA is a highly involved periodic process that requires human expertise. Successful training-data envelopment therefore entails numerous stakeholders at the agency cooperating periodically to assess the needs for retraining and to perform that retraining. Paying attention to training data stimulates internal discussion of the data's suitability and of possible improvements in detecting problematic cases that are flagged for manual processing.

To plan retraining appropriately, data scientists at the ML Lab communicate with case workers regularly with regard to analyzing the models' performance and new kinds of incoming data. Though time-consuming, this process supports employees' mutual understanding of how the models arrive at specific results. A case worker described the effect as follows:

> *I'm not that technically [grounded a] person, but doing that—training the model and seeing what output actually came out from me training the model…—made my understanding of it a lot better.* (William, Company Registration)

Through interaction during the retraining steps, the stakeholders gain greater appreciation of each other's needs:

> *In the company team, we would very much like [a model that] tells us, "Look at these areas," areas we didn't even think about: "Look at these because we can see there is something rotten going on here," basically. Other control departments would rather say, "We have seen one case that look[s] like this; there were these eight things wrong. Dear machine, find me cases that are exactly the same." And we have tried many times to tell them that that's fine. We had a case years ago where there were a lot of bakeries that did a lot of fraud, but now it doesn't make sense to look for bakeries anymore, because now these bakeries … are selling flowers or making computers or something different.* (Daniel, Company Registration)

In summary, training-data envelopment involves social effort in tandem with the purely technical endeavor of preparing suitable input-output mappings in machine-readable form that the AI can then be tasked with learning. For the training-data envelopment to succeed, the screening and ongoing monitoring of a model's performance requires the cooperation of many different stakeholders. Only this can guarantee that biases and

other deficiencies in the data are reduced—and that the model remains up to date. Otherwise, as the environment changes around the model, its boundary envelope becomes outdated. Training-data envelopment helps address this alongside issues of bias.

4.4 Input and Output Envelopment

Input and output determine, respectively, what data sources are used to create predictions and what types of decisions, classifications, or actions are created as the model's output. Any potential inputs and outputs that exhibit considerable noise, risk of bias, data omissions, or other problems are enveloped out of the AI's operation through these decisions. The selection of input sources is thus closely tied to conceptions of data quality. In the concrete case of the ID-recognition model PassportEye, the benefits of input control in conditions of poor and variable end-user-generated content became clear to the lab's staff:

> *I think our main problem was that, yeah, we had to go a little bit back and forth because the input data was [of] very varied quality. Mostly low quality. Out of the box, PassportEye actually returned very bad results, and that reflects the low quality of the input data, because people just take pictures in whatever lighting, [against] whatever background, and so on. So we actually figured out a way to rotate the images back and forth to get a more reliable result. Because, it turned out, PassportEye was quite sensitive to angle of an image. We didn't write it [the image analysis software], so this is maybe one of the risky parts when you just import a library instead of writing it yourself.* (Thomas, ML Lab)

As for output envelopment, the interplay between social and technical is more prominent here. Instead of simply preventing production of outputs that may be untrustworthy, the DBA takes a more nuanced approach. Output of appropriate confidence ratings and intervals from the models is a subject of active deliberation at the DBA. Estimates such as probabilities that a financial document is signed are important for the agency's case workers, who need them for identifying problematic cases. When an AI model yields a clearly specified and understandable confidence value, the case worker's attention can be rapidly drawn to the model's output as necessary:

> *If there's no signature, [the case workers] will simply reject it. Because the law says this document has to be signed, so the human will look at the papers and say, "It's not here. You will not get your VAT number, or your business number, because you didn't sign the document."* (James, ML Lab)

When able to verify judgments on the basis of confidence ratings, the case worker can act in an accountable manner in the interactions with DBA clients (e.g., companies that have submitted documents) and respond convincingly to their inquiries. As Steven explained:

> *If a person calls and asks, "Why was my document rejected?" then a case worker will say, "That's because you have not signed it." "How do you know that?" "I have looked at the document. It is not signed." So they don't have to answer, "Well, the neural network said it because of a variable 644 in the corner." That's why you can get away [with] using a neural network in this case, regardless of explainability.*

However, sometimes it is trickier to verify the model's output unequivocally, in which case the organization strives to understand the AI model's behavior by consulting domain experts who understand the social context of the model's output. As Steven put it, "When [it is] harder to determine if the model is right or wrong, we can push the cases to the case workers and say, 'Please look at this.'"

These examples of input and output envelopment demonstrate clear interplay between the social and the technical. While an opaque model is able to process a large quantity of unstructured data efficiently and produce recommendations on whether to accept or reject particular documents, this process is closely guided by case workers who rely on organizational objectives and legislative limitations to be sure the AI-produced decisions are in line with their needs. Thus, final decisions are produced at the intersection of actions by humans and AI.

4.5 The Implications of Envelopment for Model Choice

Having demonstrated the use of several envelopment methods in concert at the DBA, we now turn to their implications for the choice of a suitable AI model. Overall, the adoption of envelopment practices has enabled the DBA to use models that could otherwise pose risks. Different AI models are based on different architectures, which has ramifications for what the models can and cannot do. Models differ in, for example, their maturity, robustness to noise, ability to unlearn and be retrained quickly, and scalability. These qualities are dependent on the choice of the model type. For instance, robustness against noise is often easier to achieve with neural networks, while abilities of quick unlearning and retraining may be more rapidly exploited with decision trees. Depending on the needs for accuracy and/or explainability associated with a given model type, alongside the use case, suitably chosen envelopment methods can be implemented as

layers that together guarantee safe and predictable operation.

Boundary envelopment has given the DBA more degrees of freedom in choosing its models by limiting the AI agent's sphere of influence. This has allowed the staff to take advantage of complex models that, were it not for envelopment, could be rendered problematic by their lack of explainability. Jason characterized this as follows: "You can sort of say we're feeding the dragon, organization-wise, with one little biscuit at a time, so we can produce models that can be brought into production and are indeed put into production." In this way, human agents adjust the organization's processes and structures in order to contain the technological agent's operations safely.

Similarly, understanding and controlling data through training-data and input-data envelopment combined guarantee that the model's behavior is within safe limits and that the DBA possesses sufficient understanding of how the outputs are generated, even in the absence of full technical traceability. As James at the ML Lab mused:

> *Here's a new data set. What can we say about it? What should we be aware of? That's becoming increasingly important also as we are using more data connected to people's individual income, which is secret in Denmark .... Our experience with the initial use of the model ... has emphasized that this model and the data it [encompasses] needs some additional governance to safeguard that we're not going outside our initial intentions ... We've revisited some of the metadata handling that's built into the platform to ensure that we get the necessary data about how the model behaves in relation to this case handling so we can survey model output.*

With regard to output, provided that a human is able to judge its validity, one can easily opt for black-boxed models that yield superior performance. The following comment by James demonstrates how exercising output control has enabled the use of an inscrutable model: "I don't have to be able to explain how I got to the result in cases such as identifying a signature on a paper. You can just do deep learning because it's easy to verify by a human afterward."

The interviews illustrate how a need for new models may arise in response to new legislative initiatives, a new organizational strategy, or changes in taxpayer behavior. An incumbent model may have to be retrained or even entirely overhauled if metrics for accuracy or explainability indicate that it is no longer performing satisfactorily (e.g., its classifications are no longer accurate or they start leading to nonsensical estimates that cannot be explained). James gave an example illustrating the use of a boundary envelope to "mute" a model in such a case while it was directed to retraining or replacement: "The caseworkers found that the output of the model was not of quality that they could use to anything, so they muted the model. That comes back to us. We take the model down. Retrain it…." Through this process, humans decreased the AI's agency in the work process by muting it and renegotiating its agency via retraining or replacement.

4.6 Summary

The concept of envelopment has helped us flesh out our view of the conceptual and practical mechanisms of countering challenges posed by inscrutable AI. The subsections above provide empirical evidence for several distinct envelopment methods in an organizational setting. It is worth noting that, while we found evidence of the DBA actively applying boundary, training-data, and input- and output-data envelopment, we did not observe discussions about the last of the five envelopment methods listed by Robbins (2019): function envelopment, which the reader may recall refers to deciding that an AI agent will not be used for certain purposes even though it could do so accurately. Behind this decision may be ethics considerations, for instance. We believe that the lack of discussion of topics related to function envelopment at the DBA can be explained by the goals for each system having already been narrowly specified based on government regulations for every process.

We summarize the findings as follows. Considering, first, that the DBA has been able to implement several AI-based solutions successfully in its operations and, second, the evidence of envelopment in the DBA's practices (both in general and pertaining to the various methods), the concept of envelopment appears to effectively capture some of the ways in which the explainability-accuracy tradeoff presented in Figure 1 can be managed in AI implementation. Specifically, our findings indicate that, although envelopment does not change the relationship between accuracy and explainability, it allows organizations to choose from a wider range of AI models without facing an insurmountable risk of harmful consequences (e.g., wildly unpredictable outcomes). Envelopment can permit an organization to compromise some explainability for the sake of greater accuracy without needing to worry, as long as this takes place within some limits of predictable behavior. The principal benefit of envelopment is depicted in Figure 3 below.

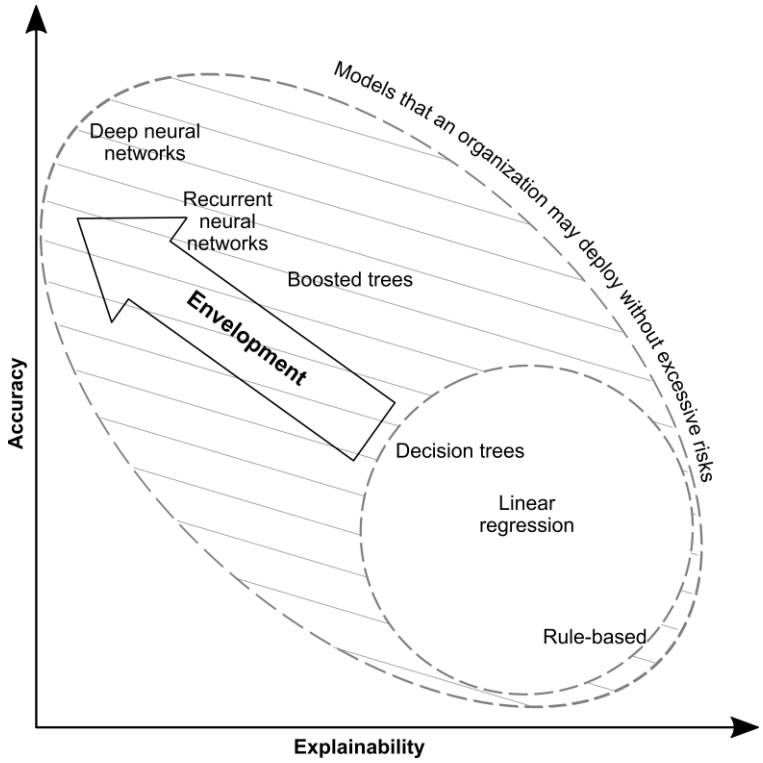


Figure 3. How Envelopment Expands the Set of Models an Organization May Adopt Without Excessive Risks

Second, in terms of the sociotechnical perspective, regardless of which envelopment method they were discussing, the interviewees never spoke of a purely technical solution for limiting AI agents' capabilities. Analysis revealed that, rather than in isolation, such actions were always carried out via iterative negotiations that took into account several stakeholder views, responsibility to society, and particular implications for the personnel's work processes.

5 Discussion

In this research, we asked: *How can an organization exploit inscrutable AI systems in a safe and socially responsible manner*? We sought answers to this question by conducting a case study of a publicly funded organization that regularly deploys AI to improve its operations, which are of importance for society. As described above, the study and analysis of the results built on the concept of envelopment as a possible approach to balancing accuracy with explainability and finding good harmony between efficiency and safety.

The analysis presented above clearly identified three significant findings. First, the case study showed that AI's envelopment, as a concept, holds empirical validity in an organizational knowledge-work setting. This complements prior envelopment literature (see Floridi, 2011; Robbins, 2020), which is of a purely conceptual nature. Second, we demonstrated that envelopment is far more than a technical matter—to be effective, it has to be situated at the intersection of the technical and the social. Our study showed how social factors pervade all aspects of envelopment and that human agents are an integral part of envelopment, responsible for defining suitable envelopes as well as maintaining and renegotiating them. Finally, the analysis articulated connections between envelopment methods and the choice of ML model. Together, these findings demonstrate the utility of envelopment—*sociotechnical* envelopment in particular—as an approach to understanding the ways in which AI's role in an organization can be conceptualized and the ways in which its responsibilities can be defined and managed. We discuss specific implications for theory and practice next.

5.1 Implications for Theory

Attending to the considerations described above allows for deeper sociotechnical discussion of enveloping AI, anchored in the DBA case as an example. This is possible via synthesis of prior literature and our empirical results. Sarker et al.'s (2019) review of sociotechnical approaches in IS research, discussed near the beginning of this paper, warns that today's IS work is in danger of too often being focused on technologies' instrumental outcomes, since they are easier to measure and evaluate. Sarker and colleagues suggest that sociotechnically oriented IS scholars would do well to address both the instrumental and humanistic outcomes of systems.

In the case of the DBA, any given AI deployment's possible instrumental outcomes would indeed be easier to analyze and declare than its humanistic outcomes, since they tie in with typical reasons for automating processes, such as aims of increased efficiency and higher precision. However, we saw that such instrumental outcomes are not the only consideration at the DBA: it was deemed crucial that AI projects not lead to misuses of government power or unnecessary profiling/surveillance of either citizens or private enterprises. Such outcomes would be problematic from a humanistic perspective and would compromise the organization's integrity as a public authority, potentially introducing erosion of public trust. Moreover, AI projects have humanistic outcomes even internally to the DBA. They expand case workers' opportunities to redesign their work processes—in fact, most of the agency's projects are undertaken in light of their proposals—and case workers are also directly involved in AI development processes. This serves to increase workplace democracy, empowerment, and occupational well-being. The DBA's AI envelopment is clearly a sociotechnical process: the technical specification of limits for AI's operations takes place via a social process wherein the case workers and other stakeholders are central actors.

The fact that the DBA's AI development is typically triggered by case workers suggests that the organization has adopted an emergent mode of operation. Case workers identify practical domain problems for the ML Lab to work on and they also participate in the AI models' development. In the search for a suitable model, ML experts and case workers analyze the capabilities and constraints entailed by various ML models, then match them interactively with the properties of the problems to be solved. When suitable models are not found for the problem at hand, the problem is broken into an alternative structure. Another approach, in such cases, is to adapt the case workers' role in resolution to mesh with the AI system's capabilities.

We propose theoretical implications for (1) describing organizational AI implementation as a balancing act between human and AI agency, and (2) conceptualizing sociotechnical envelopment as the primary tool for this crucial balancing act. Addressing the first implication builds on considering how AI development processes consist of action sequences in which case workers and AI systems, as partnered agents, carry out tasks together. The desired level of agency (that is, a suitable balance between humans and AI systems) is determined in the course of developing models and governed by the capabilities and constraints of the possible AI solutions. AI technologies' powerful information-processing capabilities offer an abundance of opportunities for numerous kinds of implementation (Kaplan & Haenlein, 2019). At the same time, thanks to ready availability of scalable computing resources, AI places few constraints on data-processing capacity (Lindebaum et al., 2020). Therefore, there are multitudes of possibilities for using such technology. However, because of the complexity of many AI models, the technology presents constraints with regard to its ability to provide technical explanations for its workings. Therefore, AI's potential still must be curbed appropriately: for example, it is necessary to find an acceptable explainability-accuracy tradeoff and, to this end, one must also establish the required level of meaningful explainability for a given context (Ribera & Lapedriza, 2019; Robbins, 2019), which takes place via negotiations across the agency among social actors. Hence, AI implementations tend to involve a balancing act between human and AI agency to arrive at a suitable level of agency for the AI. In this context, the power balance between the two parties is more equal than in many other human-technology relationships (e.g., implementing enterprise resource planning systems) in which the technology's workings are known and its capabilities seem less likely to represent unexpected negative consequences for stakeholders.

This discussion leads us to the second implication: conceptualization of sociotechnical envelopment. Two-pronged envelopment of this nature emphasizes the social dimension that is missing from existing envelopment literature (Floridi, 2011; Robbins, 2020) by focusing on the interaction of human and AI agencies, instead of on merely limiting or adjusting an AI system's capabilities. In doing so, we have been able to extend discussion on envelopment by revealing how envelopes can be constructed and maintained in a sociotechnical setting. We posit that this sociotechnical view of envelopment may offer a powerful tool for success in the balancing act between human and AI agency by offering a rich mechanism through which AI capabilities can be curbed in settings where ethics, safety, and accountability are vital to operations. This should help to offset the impact of uncertainty introduced by the inscrutability of AI and thus allow organizations to obtain efficiency gains from AI systems that offer powerful capabilities but lack explainability.

5.2 Practical Implications

For managers, whose expertise often lies in managing humans rather than AI agents, the envelopment methods presented and illustrated in this paper offer a suitable vocabulary and toolbox for handling AI development.[1] Through a process of analyzing the risks a given AI

[1] For more detailed managerial recommendations based on the case of the DBA please refer to Asatiani et al. (2020).

solution creates for business, ethics, consumer rights (e.g., the right to explanation), and environmental safety, a manager may be able to apprehend the organization's needs for envelopment. On this basis, sociotechnical approaches may be implemented and aligned with operations management and AI solution development, all in a manner that renders the models more understandable to stakeholders and addresses AI interpretability needs specific to data scientists.

A word of caution is crucial, however. Even in the presence of envelopment, one should not accept black-box models without having devoted significant effort to finding interpretable models. While a black-box model may initially appear to be the only alternative, there are good reasons to believe that accurate yet interpretable models may exist in many more domains than now recognized. Identifying such models offers greater benefit than does the sociotechnical envelopment of a black-box model. For every decision problem involving uncertainty and a limited training data set, numerous nearly optimal, reasonably accurate predictive models usually can be identified. This assertion stems from the so-called Rashomon set argument (Rudin, 2019), under which there is a good chance that at least one of the acceptable models is interpretable yet still accurate. Another recommended approach that simplifies envelopment is to strive for "gray-box models," as exemplified by the creation of "digital twins" that can simulate real, physical processes (see El Saddik, 2018; Kritzinger et al., 2018). Gray-box ML solutions are modeled in line with laws, theories, and principles known to hold in the given domain. For example, such an approach can establish a structure for a neural network, whereupon the free parameters can be trained more quickly to achieve high performance, without any reduction in explainability.

Another practical benefit of adopting envelopment as a tool for AI implementation is its relationship to technical debt. In an AI context, at least two kinds of debt can be identified. The first is related to selecting models that do not offer the best accuracy for the problems at hand (Cunningham, 1992; Kruchten et al,. 2012), as occurs if an organization needs to ensure explainability in its implementation. The other source, connected with documentation, applies to software development in general: organizations may decide to expedite their implementation efforts if they decide to relax the requirements for documenting their decisions and code (see Allman, 2012; Rolland et al., 2018). This may backfire if employee turnover rears its head and no one remains who can explain the underlying logic of the AI system. After all, answers only exist in individuals' heads or buried in code.

Envelopment may offer a means to address both types of debt: debt resulting from risk-averse choices in AI implementation that lag behind the problem's development, and debt occurring because of decisions to relax documentation requirements. Since envelopment involves carefully making and documenting decisions, it may serve as a practice whereby design decisions are rendered explicit; for example, implicit assumptions about the problem and model may be recorded. Envelopment, therefore, not only supports documentation but, by enabling the use of more accurate models, it can also decrease the accumulation of technical debt rooted in a conservative model-choice strategy.

5.3 Limitations and Further Research

Our research has some limitations. First, we used purposive sampling and studied a government unit as our empirical case since we presumed it would provide an empirically rich setting for gathering data on the use of AI. This choice, while supplying ample evidence of the envelopment strategies employed, did restrict us to studying such strategies in the specific setting of a public organization. Further research could examine envelopment of AI in a larger variety of contexts. For example, private firms driven by differently weighted objectives might use other types of envelopment strategies or employ the ones we studied in different ways. Moreover, our study did not find evidence pertaining to function envelopment—likely because the purposes of AI's use at the DBA are already strictly mandated by laws and regulations. Indeed, there was seldom reason to discuss whether the DBA's AI solutions should be applied to purposes for which they were never designed. Second, while our access to the case organization permitted in-depth analysis of the envelopment strategies applied, we could not examine their long-term implications. Further research is needed to probe the impacts of these envelopment strategies over time. Finally, while we were granted generous access for conducting interviews and analyzing secondary material, our corpus of interview data is naturally limited to what the informants expressed. To mitigate the risks associated with informant bias, we strove to obtain multiple views on all critical pieces of evidence associated with envelopment strategies. For example, we interviewed every employee working at the DBA's ML Lab, with the aim of harnessing several perspectives on each project.

With regard to both the utility of this paper and outgrowths of the efforts presented here, we wish to emphasize the value of developing a fuller understanding of the various methods by which AI and ML solutions can be controlled in order to harness the strengths they bring to the table. Envelopment strategies and their deeper examination can offer a practical means toward this end. Although the application of envelopment at the DBA was not grounded in the literature conceptualizing these

practices (e.g., Floridi, 2011; Robbins, 2020), given DBA developers' awareness of this prior work, more informed harvesting of the methods' potential could follow. Alongside such opportunities, future research could investigate whether the dynamics between humans and AI agents discussed here carry over to contexts other than AI implementation. We believe that similar logic might be identifiable, albeit in different forms, in other contexts where safe, ethical, and accountable IS implementation is crucial.

6 Conclusion

We find considerable promise in our definition and operationalization of sociotechnical envelopment in an organizational context. The findings shed light on specific instances of envelopment and they aid in identifying particular socially and technically oriented approaches to envelopment. We have been able to offer, as a starting point, a tantalizing glimpse of the capabilities and limitations of various sociotechnical envelopment approaches for addressing issues related to the safer use of AI for human good.

Acknowledgments

We are grateful to the Danish Business Authority and Early Warning Europe for the opportunity to conduct this study. We wish to thank the special issue editors and three anonymous reviewers whose insightful comments and constructive criticism helped us to greatly improve the quality of our paper. We also thank the roundtable participants at the ICIS 2019 JAIS/MISQE Special Issue Session for their feedback on our project proposal. Naturally, all remaining errors are ours.

References

Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., et al. (2020). Towards a human-like open-domain chatbot. https://arxiv.org/pdf/2001.09977v1.pdf.

Ågerfalk, P. J. (2020). Artificial intelligence as digital agency. *European Journal of Information Systems*, *29*(1), 1-8.

Allman, E. (2012). Managing technical debt. *Communications of the ACM*, *55*(5), 50-55.

Alpaydin, E. (2020. *Introduction to Machine Learning*, (4th ed.). MIT Press.

Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T., & Salovaara, A. (2020). Challenges of explaining the behavior of black-box AI systems. *MIS Quarterly Executive*, *19*(4), 259-278.

Asatiani, A., & Penttinen, E. (2019). Constructing continuities in virtual work environments: A multiple case study of two firms with differing degrees of virtuality. *Information Systems Journal*, *29*(2), 484-513.

Asatiani, A., Penttinen, E., Rinta-Kahila, T., & Salovaara, A. (2019). Implementation of automation as distributed cognition in knowledge work organizations: Six recommendations for managers. *Proceedings of the 40th International Conference on Information Systems*.

Ashby, W. R. (1958). Requisite variety and its implications for the control of complex systems. *Cybernetica*, *1*(2), 83-99.

Belgrave, L. L., & Seide, K. (2019). Coding for grounded theory. In A. Bryant and K. Charmaz (eds.), *The SAGE Handbook of Current Developments in Grounded Theory*, (pp. 167-185). SAGE.

Benbya, H., Davenport, T. H., & Pachidi, S. (2020). Special issue editorial: Artificial intelligence in organizations: Current state and future opportunities. *MIS Quarterly Executive*, *19*(4), ix-xxi.

Benbya, H., & McKelvey, B. (2006). Using coevolutionary and complexity theories to improve IS alignment: A multi-level approach. *Journal of Information Technology*, *21*(4), Springer, 284-298.

Bernard, H. R. (2017). *Research methods in anthropology: Qualitative and quantitative approaches*. Rowman & Littlefield.

Bostrom, R., Gupta, S., & Thomas, D. (2009). A meta-theory for understanding information systems within sociotechnical systems. *Journal of Management Information Systems*, *26*(1) 17-48.

Briggs, R. O., Nunamaker, J. F., & Sprague, R. H. (2010). Special section: Social aspects of sociotechnical systems. *Journal of Management Information Systems*, *27*(1), 13-16.

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. Norton.

Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data and Society*, *3*(1), 1-12.

Butler, B. S., & Gray, P. H. (2006). Reliability, mindfulness and information systems. *MIS Quarterly*, *30*(2), 211-224.

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. SAGE.

Cunningham, W. (1992). The WyCash portfolio management system. In *Addendum to the Proceedings on Object-Oriented Programming Systems, Languages, and Applications*, 29-30.

Davenport, T. (2016). Rise of the strategy machines. *MIT Sloan Management Review*, *58*(1), 29-30

Desai, D. R., & Kroll, J. A. (2017). Trust but verify: A guide to algorithms and the law. *Harvard Journal of Law & Technology*, *31*(1), 1-63.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. https://arxiv.org/pdf/1702.08608v2.pdf

Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. *Proceedings of the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*.

Dourish, P. (2016). Algorithms and their others: Algorithmic culture in context. *Big Data & Society*, *3*(2), 1-11.

Edwards, L., & Veale, M. (2017). Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke Law and Technology Review*, *16*(1), 18-84.

Edwards, P. N. (2018). We have been assimilated: Some principles for thinking about algorithmic systems. *Proceedings of the IFIP WG 8.2*

Working Conference on the Interaction of Information Systems and the Organization.

European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and the Council. *Official Journal of the European Union, L 119*(1), 1-88.

Faraj, S., Pachidi, S., & Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. *Information and Organization, 28*(1), 62-70.

Firth, N. (2019). Apple card is being investigated over claims it gives women lower credit limits. *MIT Technology Review*. https://www.technologyreview.com/2019/11/11/131983/apple-card-is-being-investigated-over-claims-it-gives-women-lower-credit-limits/

Floridi, L. (2011). Children of the fourth revolution. *Philosophy and Technology*, *24*(3), 227-232.

Ghasemaghaei, M., Ebrahimi, S., & Hassanein, K. (2018). Data analytics competency for improving firm decision making performance. *The Journal of Strategic Information Systems, 27*(1), 101-113.

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2012). Seeking Qualitative Rigor in inductive research: Notes on the Gioia methodology. *Organizational Research Methods, 16*(1), 15-31.

Gregor, S., & Benbasat, I. (1999). Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Quarterly, 23*(4), 497-530.

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons, 61*(4), 577-586.

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons, 62*(1), 15-25.

Keding, C. (2021). Understanding the interplay of artificial intelligence and strategic management: Four decades of research in review. *Management Review Quarterly*, *71*(1), 91-134.

Klein, H., & Myers, M. M. D. (1999). A set of principles for conducting and evaluating interpretive field studies in information systems. *MIS Quarterly*, *23*(1), 67-93.

Koutsikouri, D., Lindgren, R., Henfridsson, O., & Rudmark, D. (2018). Extending digital infrastructures: A typology of growth tactics. *Journal of the Association for Information Systems, 19*(10), 1001-1019.

Kritzinger, W., Karner, M., Traar, G., Henjes, J., & Sihn, W. (2018). Digital twin in manufacturing: A categorical literature review and classification. *IFAC-PapersOnLine, 51*(11), Elsevier, 1016-1022.

Kruchten, P., Nord, R. L., & Ozkaya, I. (2012). Technical debt: From metaphor to theory and practice. *IEEE Software, 29*(6), 18-21.

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.

Linden, A., Reynolds, M., & Alaybeyi, S. (2019). *5 Myths about explainable AI.* Gartner Research.

Lipton, Z. C. (2018). The mythos of model interpretability. *ACM Queue, 16*(3), 1-27.

Liu, N., Du, M., & Hu, X. (2020). Adversarial machine learning: An interpretation perspective. https://arxiv.org/pdf/2004.11488.pdf.

London, A. J. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Center Report, 49*(1), 15-21.

Lyytinen, K., Nickerson, J. V, & King, J. L. (in press). Metahuman systems = humans + machines that learn. *Journal of Information Technology*. https://doi.org/10.1177/0268396220915917.

Martens, D., Vanthienen, J., Verbeke, W., & Baesens, B. (2011). Performance of classification models from a user perspective. *Decision Support Systems, 51*(4), 782-793.

Martin, K. (2019). Designing ethical algorithms. *MIS Quarterly Executive, 18*(2), 129-142.

McBride, N., & Hoffman, R. R. (2016). Bridging the ethical gap: From human principles to robot instructions. *IEEE Intelligent Systems, 31*(5), 76-82.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. C., Darzi, A., & others. (2020). International evaluation of an AI system for breast cancer screening. *Nature, 577*(7788), 89-94.

Miles, M. B., Huberman, M. A., & Saldana, J. (2014). Drawing and verifying conclusions. In *Qualitative data analysis: A methods sourcebook* (pp. 275-322). SAGE.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1-38.

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

Mumford, E. (2006). The story of socio-technical design: Reflections on its successes, failures and potential. *Information Systems Journal, 16*(4), 317-342.

Myers, M. D., & Newman, M. (2007). The qualitative interview in IS Research: Examining the craft. *Information and Organization, 17*(1), 2-26.

Newell, S., & Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of "datification." *Journal of Strategic Information Systems, 24*(1), 3-14.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Pääkkönen, J., Nelimarkka, M., Haapoja, J., & Lampinen, A. (2020). Bureaucracy as a lens for analyzing and designing algorithmic systems. *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Pasquale, F. (2015). *The black box society*. Harvard University Press.

Patton, M. Q. (2001). *Qualitative Evaluation and Research Methods* (3rd ed.). SAGE.

Preece, A. (2018). Asking "why" in AI: Explainability of intelligent systems—perspectives and challenges. *Intelligent Systems in Accounting, Finance and Management, 25*(2), 63-72.

Raisch, S., & Krakowski, S. (in press). Artificial intelligence and management: The automation-augmentation paradox. *Academy of Management Review*. https://journals.aom.org/doi/10.5465/2018.0072

Ribera, M., & Lapedriza, A. (2019). Can we do better explanations? A proposal of user-centered explainable AI. In . In *Joint Proceedings of the ACM IUI 2019 Workshops*.

Robbins, S. (2020). AI and the path to envelopment: Knowledge as a first step towards the responsible regulation and use of AI-powered machines. *AI & Society*, 25, 391-400.

Robbins, S. (2019). A misdirected principle with a catch: Explicability for AI. *Minds and Machines, 29*(4), 495-514.

Rolland, K. H., Mathiassen, L., & Rai, A. (2018). Managing digital platforms in user organizations: The interactions between digital options and digital debt. *Information Systems Research, 29*(2), 419-443.

Rosenfeld, A., & Richardson, A. (2019). Explainability in human-agent systems. *Autonomous Agents and Multi-Agent Systems, 33*, 673-705.

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.

El Saddik, A. (2018). Digital twins: The convergence of multimedia technologies. *IEEE MultiMedia, 25*(2), 87-92.

Salovaara, A., Lyytinen, K., & Penttinen, E. (2019). High reliability in digital organizing: Mindlessness, the frame problem, and digital operations. *MIS Quarterly, 43*(2), 555-578.

Sarker, S., Chatterjee, S., Xiao, X., & Elbanna, A. (2019). The sociotechnical axis of cohesion for the IS discipline: Its historical legacy and its continued relevance. *MIS Quarterly, 43*(3), 695-719.

Sarker, S., Xiao, X., Beaulieu, T., & Lee, A. S. (2018). Learning from first-generation qualitative approaches in the IS discipline: An evolutionary view and some implications for authors and evaluators (Part 1/2). *Journal of the Association for Information Systems, 19*(8), 752-774.

Sarker, Saonee, & Sarker, Suprateek. (2009). Exploring agility in distributed information systems development teams: An interpretive study in an offshoring context. *Information Systems Research, 20*(3), 440-461.

Scheel, P. D. (1993). Robotics in industry: A safety and health perspective. *Professional Safety, 38*(3), 28-32.

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.

Sørmo, F., Cassens, J., & Aamodt, A. (2005). Explanation in case-based reasoning-perspectives and goals. *Artificial Intelligence Review, 24*, 109-143.

Sousa, W. G. de, Melo, E. R. P. de, Bermejo, P. H. D. S., Farias, R. A. S., & Gomes, A. O. (2019). How and where is artificial intelligence in the public sector going? a literature review and

research agenda. *Government Information Quarterly*, *36*(4), 101392.

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Julia, H., Kalayanakrishnan, S., Kamar, E., Kraus, S., Leyton-Brown, K., Parkes, D., Press, W., Saxenian, A., Shah, J., Tambe, M., & Teller, A. (2016). *Artificial intelligence and life in 2030: One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel*. Stanford University. https://ai100.stanford.edu/sites/g/files/sbiybj9861/f/ai_100_report_0831fnl.pdf

Tavory, I., & Timmermans, S. (2014). *Abductive analysis: Theorizing qualitative research*. University of Chicago Press.

Weller, A. (2019). Transparency: Motivations and Challenges. *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning*.

Wright, S. A., & Schultz, A. E. (2018). The rising tide of artificial intelligence and business automation: developing an ethical framework. *Business Horizons*, *61*(6), 823-832

Appendix A: The DBA's ML Projects

Project name	Project description (use case within the DBA and end users)	Purpose	Input	Output	Model and tool
Auditor's Statement	The Auditor's Statement model speeds up verification that the valuations of company assets given in an auditor's statement are correct and that the statement does not feature violations. The algorithm is used by internal DBA case workers.	Prevent misreporting of company assets	Text from auditor's statements that present asset valuations	Probability of violations in asset valuations	Random forest, bag of words
Bankruptcy	The Bankruptcy model predicts company distress and insolvency and ties in with the Early Warning Europe (EWE) initiative. The algorithm is used not at the DBA but by external consultants in the EWE community in Denmark and elsewhere in the European Union. The DBA is not responsible for actions and consequences related to the tool.	Identify companies in distress, to enable timely intervention	Data from the business registry and annual statements	Probability of bankruptcy	Scikit-learn, gradient boosting
Company Registration	The Company Registration model is aimed at detecting fraud-indicating behavior among newly registered Danish companies. The algorithm is used by internal DBA case workers.	Prevent abusing incorporation to commit fraud	Data from the business registry, annual statements, and VAT reports	Probability of fraudulent actions	XGBoost
Land and Buildings	The Land and Buildings model predicts violations of accounting policies related to property holdings and long-term investments. The algorithm is used by internal DBA domain experts.	Prevent violations of accounting policy	Text about accounting policies, from the auditor's statement	Probability of violations of accounting policies	Random forest, bag of words
ID Verification	The ID Verification model expedites processing of the documents submitted, by supplying a text string from the machine-readable portion of an ID document and comparing it against input data from the user. The algorithm is used by internal DBA case workers.	Facilitate processing of documents	Pictures of IDs submitted to the DBA	JSON string with text from the machine-readable portion of the ID	PassportEye
Recommendation	The Recommendation model improves the user experience of the DBA's virk.dk online portal by focusing on personalized content and optimized interfaces. The algorithm improves the portal's usability for external customers (end users).	Improve usability of the online portal	Telemetry data from virk.dk	Recommendation of relevant content	[Not decided by the time of this study]
Sector Code	The Sector Code model speeds up verifying a company's industry-sector code. At present, 25% of the company codes are incorrect. The algorithm is used by internal DBA case workers.	Prevent misreporting of industry-sector codes	Activity-description text from a company's annual statements	Probability distribution over the set of sector codes	Neural network
Signature	The Signature model, in combination with the associated document filter, speeds up verification of whether a company-establishment document is signed or not. The algorithm, used by internal DBA case workers, returns three probabilities: of whether the document is physically signed, whether it is digitally signed, and whether the signature is missing.	Facilitate the process of establishing a company	An image of a company-establishment document	Probability of whether a document is signed or not	Neural network (ResNet16)

Appendix B: The Interview Protocol

Personal background

Could you tell us about your academic and professional background?

How long have you been part of the DBA, and how long have you held your current position?

Could you tell us about projects you are involved in at the DBA?

ML and AI projects at the DBA

Could you list machine-learning and AI projects currently being carried out by the ML Lab?

Could you describe ML/AI projects that you are involved with?

What types of algorithms and models are used in these projects?

What is the rationale behind using these models?

In your own words, could you please explain…

- Which data go into the system and what type of output the algorithm provides?
- How well you understand how the algorithm works?
- How you interpret the output?

Use of black-box models and explainability

How explainable are the decisions of the AI used in the projects you are involved in?

Who is able to understand how the AI produces its outputs (data scientists, developers, case workers, …)?

Have you encountered a case in which you needed to explain a particular AI decision? Could you describe the case in detail?

Has this explanation been documented? Could you provide documents?

Could you give a concrete example of a typical decision your AI makes?

How would you explain the resulting decision if requested to do so…

- By qualified auditors?
- By an affected organization?
- By the general public?

What would be the procedure for requesting the explanation, and for delivering it?

Is explanation embedded in the algorithm (or predefined protocol)'s design, or is it *ad hoc* / emergent?

Explainability requirements

How does the requirement for explainability manifest itself in algorithm development?

- Do you use different machine-learning platforms for projects that require explainable AI?

Have you had any issues or problems with explainability (in development, in relations with external stakeholders, DBA-internally, or with regard to managers)?

- Have explanations been requested? By whom?
- Have you been able to provide satisfactory explanations upon request?
- Have you experienced inability to provide explanations to a stakeholder or to obtain explanations from one?

How should explainability be taken into account in system development?

What design principles were applied in development of PROJECTX (cost, time, etc.)?

How was the design of PROJECTX organized (following a waterfall model, in sprints, etc.)?

Was explainability a system requirement in the AI design?

- What did this mean for the design process?

- If explainability was initially specified as a system requirement, did it materialize in the final design as was intended? That is, did the final design's explainability correspond to what was envisioned?

Describe the process of crafting an explanation:

- Who creates it?
- How often, and for whom?
- What are the steps?

Were any of the design principles in conflict with explainability during the design phase?

- If so, how did you navigate through the issue?

Have you noticed conflicts related to differing understandings of the work done by the algorithm?

- Could you give examples?
- Is such conflict acceptable, or do contradictions need to be reconciled?
- How are they reconciled?
- What do you consider the best way to resolve conflicts?

Reasons for developing explainable AI and its implications

What are the main reasons for the requirement to explain AI?

Why do you need explainability?

- For internal purposes: for finding out how to improve your AI, or to double-check its outputs?
- For external purposes: to be accountable as a governmental authority with defendable unbiased processes?

External pressure for explainability:

- Do you have to be able to explain AI decisions to clients (taxpayers)? How, and at what level of detail?
- Which regulations, internal policies, outside pressure, etc. force you to explain the AI's decisions?
- Who are the main actors for whom you craft explanations? Could you name them and provide examples of what those explanations are like?

How do explainability requirements constrain the process of AI development? Could you describe these constraints?

- Do you have to limit your use of AI approaches because of a need for explainability?

How does needing to produce explainable systems affect the systems' performance?

Overall, how does explainability influence your ability to achieve organizational objectives?

Appendix C: The Coding

Concepts (first-order)	Themes (second-order)	Aggregate dimensions	Example quotations
• Case workers' control of thresholds • Guidance on threshold-setting • The thresholds' dependence on the code	Thresholds	Boundary envelopes	"But we're involved more or less the whole way because if suddenly there is a problem or suddenly there is 'Okay, we can deploy this, but do you want the machine to do this or this? Do you want it to have a marker saying this case cannot go further, or do you just want it to go through and [we] have a special marker where we can look it up later?'... So we are involved the whole way, but at some points we are more [in the goals or in practice] helping or [asking] 'Can we do...?'"
• Conversion of probabilities into flags • The AI flagging only basic flaws in documents	Flags		
• Designing AI that is easier to hand over • Basic AI tools with wide applicability	Division of a task into smaller parts		
• Simple algorithms' ease of explanation • An explainability/performance tradeoff not always existing—simple models work just fine	Choosing of interpretable algorithms		
• Close communication links for reducing misunderstandings during development • Communication with developers	Social dialogue		
• Understanding of input data as important • Quality of inputs	Input control	Input and output envelopes	"An example could be that our model [for whether a document is] signed or not, as it is now, if the model forecasts that the document is signed, then it gets a special code, 'document signed, everything is okay,' and if it's not signed, then it gets another marking, for 'document not signed.' These cases we go through, and then you can see that was correct and that was not correct. In that case, there isn't really any- we don't need to know- I don't need to know as [a case worker] why the model said 'signed' or 'not signed,' because I can see instantly if it's right or not right."
• Compensation for explainability-induced lower performance, via control over the output's use • Acceptability of having a black box if checking the outputs is simple	Output control		
• Verification as an aid to establishing trust in ML— • a human holding ultimate responsibility • Simple algorithms that a human expert can follow and reproduce	Human verification		
• External stakeholders' involvement in early stages of development • Establishment of feedback channels between technical and business teams	Human feedback	Model-choice envelopes	"We have around 160 rules. We have technical rules that look into whether the right taxonomy is being used, whether it is the XBRL format, and whether it is compliant. We also have business rules. For example, do assets and liabilities match? Some rules only look at technical issues in the instance report. Some rules are what we called full-stop rules: … filers are not allowed to file the report until they have corrected the error. We also have more guidance[-type] rules, where we say, 'It looks like you're about to make a mistake. Most people do it *this* way. Are you sure you want to continue filing the report?' And then [users] can choose to ignore the rule."
• Governance of AI development • In-house development, to improve understanding	Continuous-improvement procedure		
• Internal accumulation of training data • Data "red herrings" • Training on in-house data	Knowledge of data	Training-data envelopes	"I think it's important with these models to look at them often to see if something is changing. And, maybe, train them again. Because I think there might be some issues, with the robustness. We haven't gotten this system into production yet, but I think it's on its way."
• Challenges of creating models • The dangers of training a model on the open internet • Training of models in stages	Phased training of a model		

About the Authors

Aleksandre Asatiani is an assistant professor in information systems at the Department of Applied Information Technology, at the University of Gothenburg. He is also an affiliated researcher with the Swedish Center for Digital Innovation (SCDI). His research focuses on artificial intelligence, robotic process automation, virtual organizations, and IS sourcing. His work has previously appeared in leading IS journals such as *Information Systems Journal*, *Journal of Information Technology*, and *MIS Quarterly Executive*.

Pekka Malo is a tenured associate professor of statistics at Aalto University School of Business. His research has been published in leading journals in operations research, information science, and artificial intelligence. Pekka is considered as one of the pioneers in the development of evolutionary optimization algorithms for solving challenging bilevel programming problems. His research interests include business analytics, computational statistics, machine learning, optimization and evolutionary computation, and their applications to marketing, finance, and healthcare.

Per Rådberg Nagbøl is a PhD fellow at the IT University of Copenhagen doing a collaborative PhD with the Danish Business Authority within the field of information systems. He uses action design research to design systems and procedures for quality assurance and evaluation of machine learning, focusing on accurate, transparent, and responsible use in the public sector from a risk management perspective.

Esko Penttinen is a professor of practice in information systems at Aalto University School of Business in Helsinki. He holds a PhD in information systems science and an MSc in Economics from Helsinki School of Economics. Esko leads the Real-Time Economy Competence Center and is the co-founder and chairman of XBRL Finland. He studies the interplay between humans and machines, organizational implementation of artificial intelligence, and governance issues related to outsourcing and virtual organizing. His main practical expertise lies in the assimilation and economic implications of interorganizational information systems, focusing on application areas such as electronic financial systems, government reporting, and electronic invoicing. Esko's research has appeared in leading IS outlets such as *MIS Quarterly*, *Information Systems Journal*, *Journal of Information Technology*, *International Journal of Electronic Commerce*, and *Electronic Markets*.

Tapani Rinta-Kahila is a postdoctoral research fellow at the UQ Business School and Australian Institute for Business and Economics, at the University of Queensland in Australia. He holds a doctoral degree in information systems science from the Aalto University School of Business. His research addresses issues related to IT discontinuance, organizational implementation of artificial intelligence and automation, and the dark side of IS.

Antti Salovaara is a senior university lecturer at Aalto University, Department of Design and an adjunct professor in the Department of Computer Science at the University of Helsinki. He studies human-AI collaboration and online trolling and the methodology of user studies. His research has been published both in human-computer interaction and information systems journals and conferences, including *CHI*, *Human Computer Interaction* and *International Journal of Human-Computer Studies*, as well as *MIS Quarterly* and *European Journal of Information Systems*.

MIS Quarterly Executive

Volume 19 | Issue 4 Article 7

December 2020

Challenges of Explaining the Behavior of Black-Box AI Systems

Aleksandre Asatiani

Pekka Malo

Per Rådberg Nagbøl

Esko Penttinen

Tapani Rinta-Kahila

See next page for additional authors

Challenges of Explaining the Behavior of Black-Box AI Systems

Authors

Aleksandre Asatiani, Pekka Malo, Per Rådberg Nagbøl, Esko Penttinen, Tapani Rinta-Kahila, and Antti Salovaara

MIS Quarterly Executive

Challenges of Explaining the Behavior of Black-Box AI Systems

There are many examples of problems resulting from inscrutable AI systems, so there is a growing need to be able to explain how such systems produce their outputs. Drawing on a case study at the Danish Business Authority, we provide a framework and recommendations for addressing the many challenges of explaining the behavior of black-box AI systems. Our findings will enable organizations to successfully develop and deploy AI systems without causing legal or ethical problems.[1,2]

Aleksandre Asatiani
University of Gothenburg (Sweden)

Pekka Malo
Aalto University School of Business (Finland)

Per Rådberg Nagbøl
IT University of Copenhagen (Denmark)

Esko Penttinen
Aalto University School of Business (Finland)

Tapani Rinta-Kahila
The University of Queensland (Australia)

Antti Salovaara
Aalto University School of Arts, Design and Architecture (Finland)

Organizations Need to Be Able to Explain the Behavior of Black-Box AI Systems

Huge increases in computing capacity and data volumes have spurred the development of applications that use artificial intelligence (AI), a technology that is being implemented for increasingly complex tasks, from playing Go to screening for cancer. Private and public businesses and organizations are deploying AI applications to process vast quantities of data and support decision making. These applications can help to reduce the costs of providing various services, deliver new services and improve the safety and reliability of operations.

However, unlike conventional information systems, the algorithms embedded in AI applications can be "black boxes." Previously, those who developed applications could completely explain how an algorithm worked. Given an input, they could tell you what the output would be and why, because the systems applied human-made rules. That is no longer true for AI-based applications. The application creates internal structures that determine outputs, but these are inscrutable to outside observers, and even the programmers cannot tell you why a specific output was generated. Many AI systems leverage machine learning,

1 Hind Benbya is the accepting senior editor for this article.

2 The authors thank Hind Benbya and the members of the review team for their insightful feedback that has greatly improved the quality of this article. We are grateful to the Danish Business Authority for sharing their time and allowing us to conduct this study.

DOI: 10.17705/2msqe.00037

where a model learns how to act by detecting patterns in data by employing only general principles for how such patterns can be found. The actual process of finding those patterns may remain hidden and there is no human input or intervention in the process.

As a consequence, information systems (IS) researchers are striving to find ways to improve the transparency of algorithms embedded in AI applications—i.e., to provide the ability to explain the rationale or logic behind algorithmic decisions to human stakeholders. IS researchers and academics refer to this area as the "explainability"[3] of black-box AI algorithms.

The ability to explain how AI algorithms reach their decisions is a legal requirement in Europe. The European Union's General Data Protection Regulation (GDPR) mandates an individual's right to explanation. From an ethical point of view, the ability to explain can help to identify and defuse problematic biases. For example, Amazon's face-recognition and recruitment models were found to develop racial and gender biases.[4] Similarly, from a safety perspective, the ability to explain can help to identify the source of the problem in cases where an AI application has—from the users' point of view—made a mistake. Explanations can help to prevent such problems from reoccurring.

Thus, in their search for greater performance, organizations must deploy AI applications in a legal, ethical and safe manner, which means they must have the ability to explain how the applications make their decisions. This is especially true in the public sector, where public trust and confidence in AI-based decisions are of paramount importance.

Although there have been several attempts to produce technical explanations that allow humans to understand the behavior of AI applications, this is not always feasible because of the inductive reasoning applied by many AI applications. Technology giants (including Google and IBM) are beginning to offer AI solutions that are, at best, partially explainable and, in the U.S., the Defense Advanced Research Projects Agency has a program dedicated to the task of developing explainable AI. An inability to provide sufficient and meaningful explanations creates barriers for the successful deployment of AI applications in an organization, and therefore hinders the potential benefits of higher operational efficiency and accuracy.

Explaining the behavior of AI systems requires more than purely technical measures. Organizations must also consider what the outputs from the systems mean for human stakeholders.[5] A recent report[6] on using AI to combat public-sector fraud suggests that *"where a technical explanation for an AI tool is not possible, practical or meaningful, an ability to explain the priorities or strategic basis for a decision may suffice and may even be more meaningful ... depending upon the context."* Acquiring the ability to explain thus requires a managerial solution; however, there is a scarcity of such solutions.

Our research therefore addressed the question: How can organizations reconcile the growing demands for explanations of how AI-based algorithmic decisions are made with their desire to leverage AI to maximize business performance? This article presents the findings of our research, which are based on a case study of the Machine Learning Lab at the Danish Business Authority. (Details of the study are provided in Appendix A.)

First, we describe the six elements of a hypothetical intelligent AI agent—the model, goals, training data, input data, output data and environment. We then present a framework with six dimensions, each corresponding with one of the elements that will enable organizations to explain how AI-based algorithms reach their decisions. We then illustrate how this framework helped our case organization, the Machine Learning Lab at the Danish Business Authority, to responsibly and successfully exploit apparently unexplainable black-box AI. The lessons from this case are valuable both for IS researchers and

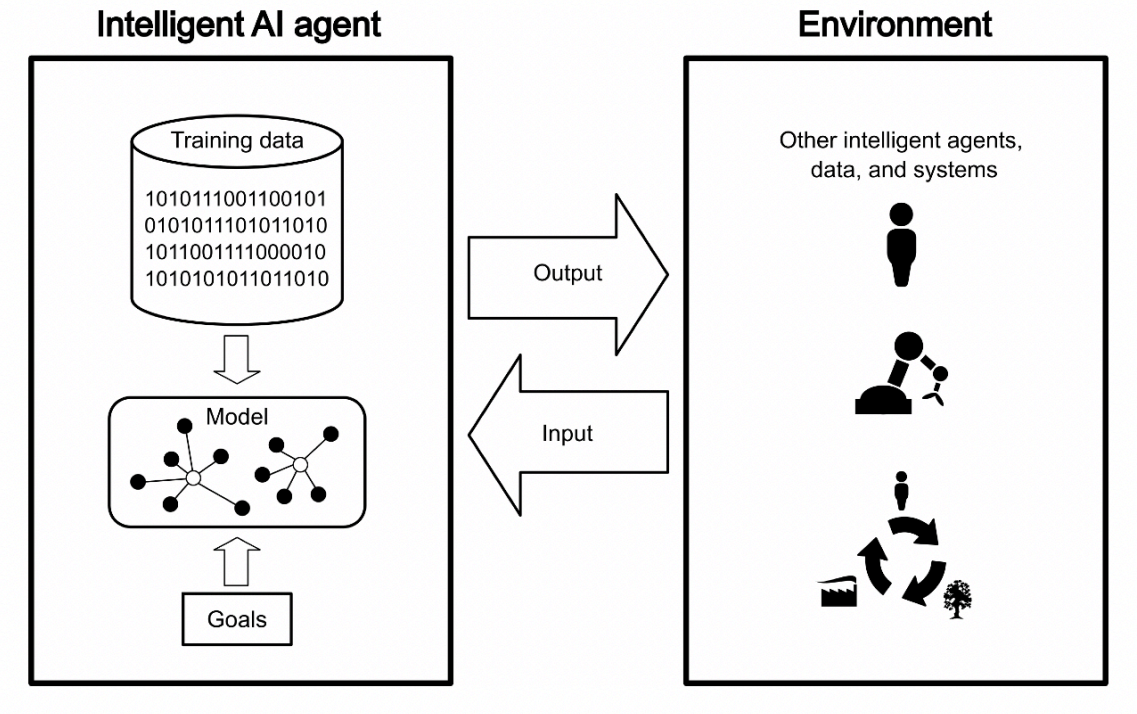
3 For a description of explainability, see Rosenfeld, A. and Richardson, A. "Explainability in Human–Agent Systems," Autonomous Agents and Multi-Agent Systems (33), May 2019, pp. 673-705.

4 See, for example, Vincent, J. "Gender and Racial Bias found in Amazon's Facial Recognition Technology (Again)," *The Verge*, January 25, 2019, available at https://www.theverge.com/2019/1/25/18197137/amazon-rekognition-facial-recognition-bias-race-gender.

5 For more information, see Martin, K. "Designing Ethical Algorithms," *MIS Quarterly Executive* (18:2), May 2019, pp. 129-142.

6 The Use of Artificial Intelligence to Combat Public Sector Fraud: Professional Guidance, Serious Fraud Office [U.K.], in collaboration with New Zealand's Serious Fraud Office, February 2020. This report was prepared by members of the International Public Sector Fraud Forum.

Figure 1: The Six Elements of an Intelligent AI Agent



designers of black-box AI applications, and for organizations that deploy such applications.

The article concludes with four recommendations derived from our analysis of the case study. These recommendations provide executives with a toolbox for proactively managing issues concerned with explaining how AI algorithms work and thus helping them to reap the potential benefits of AI applications.

The Six Elements of an Intelligent AI Agent

Our framework is described by reference to a hypothetical autonomous intelligent AI agent (which is depicted in Figure 1). According to Russel and Norvig, an intelligent agent, whether human or machine, pursues goals by processing data and interacting with other agents in the environment.[7] The intelligent AI agent we reference has six main elements: the model, goals, training data, input data, output data and environment. These elements form the dimensions of our framework.

Typically, developers of a machine-learning-based application construct the AI model by defining its mathematical formulation, setting goals for it, and training it to reach those goals. The model, which is, in essence, a mathematical function that relates an input to an output, has parameters whose originally unknown values are specified via a suitable training algorithm. Obviously, the choice of model has implications for understanding and being able to explain how the AI agent works: a human may be able to follow the if-then paths of simpler models, but this might not be possible with more complex models.

The *goals* of an intelligent AI agent are performance metrics (such as accuracy levels or average prediction error rate) that allow developers and other stakeholders to evaluate

7 For more on this definition of an intelligent agent, see Russell, S. J. and Norvig, P. *Artificial Intelligence: A Modern Approach*, 3rd edition, Prentice Hall, 2010.

whether the agent is satisfying the performance criteria set for it. By scrutinizing the agent's goals, people may be able to explain its behavior (for instance, testing the agent's output against various performance metrics might reveal imbalances in the original training data).

To prepare the intelligent AI agent for use, the algorithm is run on a set of *training data*, which the algorithm uses to identify suitable values for the model's as-yet-unspecified parameters. Supervised or unsupervised learning approaches may be used, depending on the business problem the AI agent is being used for.

In supervised learning, the AI agent is trained from a labeled dataset with data organized into predefined categories. For example, if the AI agent is to make predictions of a company's success or failure, the training data might include figures from annual financial statements that could be expected to predict business success. Once trained, a set of test data not used in its training is input to the AI agent. The agent's performance can then be evaluated against its goals (e.g., its ability to predict business risks accurately).

In contrast to supervised learning, unsupervised learning makes sense of the input data independently; it does not make use of neatly categorized training data. With unsupervised learning, the AI agent searches for hidden patterns (e.g., uncovering sources of business failures from combinations of financial and/or other indicators). In most applications, unsupervised learning is best described as an exploratory or descriptive tool.

Regardless of whether the learning approach is supervised or unsupervised, the body of data used to train the AI agent shapes its capabilities and is therefore integral to the model used. Some explanations of the behavior of the AI agent are rooted in biases found in the training data, which, for example, may reveal why the agent discriminates for or against certain groups of people.

Once the AI agent has been trained and validated, it is deployed for real-world use. Actual *input data* (e.g., figures from companies' annual statements) is fed into the black-box algorithm, which then produces *output data* (e.g., a probability of a business failing). Examination of the input and output data can reveal explanations for the AI agent's behavior. For instance, imprecise recording of input data may point to why there are flaws in the agent's output data. Comparing the output data to other available information can also help in tracing the agent's decision logic and finding blind spots in its operations.

The final element of the AI agent is the *environment* in which it operates. The environment determines the sources and validity of the incoming data, and the agent influences the environment via its outputs (e.g., the resultant risk assessment of a company's future may shape the actions of the company). Such feedback loops are especially important in "reinforcement learning," where the AI agent learns from interacting with its environment by trial and error and receives rewards for good performance. If the AI agent is deployed in a different environment, it is unlikely to operate correctly (e.g., a system trained to identify business risks may not perform well in non-business settings). Thus, an AI agent's inappropriate behavior might be explained by it being deployed in an environment for which it was not trained.

A Framework for Explaining the Behavior of Black-Box AI Systems

The above discussion suggests that the ability to explain the behavior of an AI agent can be enhanced by examining and suitably designing each of the six elements. Thus, as summarized in Table 1, our framework for explaining the behavior of black-box AI systems has six dimensions, each of which corresponds to one of the elements of the hypothetical AI agent.

Dimension 1: The AI System's Model

A core element of the ability to explain how an AI system operates is a thorough understanding of the model used—specifically, how it turns inputs into outputs. At the technical level, gaining this understanding can be fairly easy for simple, rule-based systems or certain machine-learning models such as decision trees and regressions. However, technical explanations may not be practical or even possible with more complex models where logical decision rules cannot be extracted, such as deep neural networks (layered computing systems whose structure resembles

Table 1: Six-Dimension Framework for Explaining the Performance of AI Systems

Dimension	Description	Example
1. Model	Explanation of the AI system's logic/behavior based on tracing its decision-making patterns.	A specific business-risk probability may be explained by the if-then sequence of steps taken by a business-risk estimation model.
2. Goals	Explanation of the AI system's logic/behavior derived from priorities or the strategic basis for a given decision.	The agent flags high probabilities of risk for companies that engage in reputation-compromising activities such as producing health-harming products or causing environmental damage, with the explanation lying in the fact that the model is trained and tested with performance metrics that give great weight to risking the organization's reputation.
3. Training Data	Explanation based on the characteristics of the training data.	The agent assigns exceptionally high probabilities of risk to certain types of business, such as medical practices, because of biased training data. Data on medical practitioners might have been collected in economically deprived areas while data from other businesses are geographically more diverse.
4. Input Data	Explanation based on the characteristics of the input data.	Unreliable business-risk probabilities can be explained by low-quality input data produced by inaccurate measurement of relevant risk factors.
5. Output Data	Explanation derived from humans' examination and verification of the output.	A human examines the validity of the AI agent's business-risk probability for a loan application and makes sure that the rationale for the decision can be explained to the applicant in meaningful terms.
6. Environment	Explanation that is based on the environment in which the AI agent operates.	Inappropriate risk estimations may be explained by the AI agent being fed risk-assessment data from environments that are not suitable for this purpose (e.g., using soccer-league scoring data to predict the risks of businesses not connected to soccer).

that of the biological networks of neurons in brains). Although the model's designers and developers most certainly understand the underlying mathematical formulation of their models, even they may find it very difficult, if not impossible, to explain the model's behavior once it has been trained and is used to process actual data.

The difficulty of providing a technical explanation is compounded in models where not only millions of parameters are learned from training data but the underlying structure or model topology is adjusted automatically by the training algorithm. The inability to explain how such an AI application has made a decision has caused problems in high-profile contexts, such as police trying to detect potential offenders before they have committed a crime.[8] Models that have been trained using unsupervised learning are typically more difficult to explain than supervised ones because of the lack of a priori labeling and benchmarking standards. Although reinforcement learning models can be assessed

8 See, for example, "Rules Urgently Needed to Oversee Police Use of Data and AI – Report," *The Guardian*, February 23, 2020, available at https://www.theguardian.com/uk-news/2020/feb/23/rules-urgently-needed-oversee-police-use-data-ai-report.

against various criteria, their trial-and-error-based learning logic makes them particularly challenging to explain.

Other dimensions of the framework can be used to explain the behavior of a black-box AI system. Whether these explanations are sufficient depends on several factors, including legislation, the impact of the decisions on stakeholders and ethical issues.

Dimension 2: The AI System's Goals

In setting goals for an AI system, developers need to translate high-level business objectives into concrete performance metrics that can be used to steer the development of the agent. Well-chosen metrics serve as the primary means for comparing the performance of competing inscrutable models against each other. Ideally, they can also help to explain a model's behavior by revealing situations where it performs well and where it fails. Consider, for instance, the example mentioned above of using a neural network for detecting potential offenders: If this AI application successfully identifies a high proportion of future offenders, it is deemed to have high accuracy. However, it might still produce an unacceptably large number of false positives within some groups (e.g., certain ethnic groups may be overrepresented) while failing to predict actual offenders in other groups, because the accuracy metric does not account for imbalances in the distribution of ethnicity data. Testing the AI system against a performance metric that does address this possibility helps to reveal such imbalances and, thus, provides explanations for the underlying logic. Such testing shifts the emphasis to training data, as discussed next.

Dimension 3: Training Data

The way in which an AI system performs is determined by the characteristics of the data used to train it. A biased training dataset leads to biased decisions even if there is nothing wrong with the functionality of the algorithm taught by the data. A good example is the AI system used in U.S. courts to predict convicts' risk of recidivism, which was found to have a racial bias.[9] The data used to train an AI system tends to reflect biases in the real world, which causes the AI system to adopt the same biases and therefore produce biased outputs. Even when the algorithm is too complex to be explained meaningfully, awareness of the characteristics of the training data can shed light on how and why it translates the input data into outputs in the way it does.

Dimension 4: Input Data

Insufficient attention to the quality of input data can result in difficulties in explaining the behavior of an AI system, as illustrated by a scandal in Australia. A simple AI system was deployed for identifying social-welfare debt and initiating debt collection from citizens.[10] Poor input data quality, stemming from pairing two incompatible data sources, caused the AI system's debt calculations to be incorrect. Although the algorithm was technically explainable, neither government workers nor citizens had been informed of the incompatibility of the sources the AI system was drawing on. Their impression was that the system was a black box, which made it difficult for them to prove the incorrectness of the debt calculations. As a consequence, workers and affected citizens suffered significant stress. A thorough understanding of the limitations resulting from matching incompatible datasets would have mitigated the problems that ensued.

Dimension 5: Output Data

The problems with the Australian AI system were aggravated by a decision to remove human workers from the debt-collection loop: the lack of human oversight of the AI system's outputs enabled erroneous debt claims to be sent to citizens. Having humans check the outputs becomes all the more important with a black-box AI system that employs opaque decision-making logic. The Russian proverb "trust but verify" is very apt: even if the model performs well, it may need a human gatekeeper.[11] Although the algorithm itself may be opaque, scrutinizing the

9 See Buranyi, S. "Rise of the Racist Robots – How AI is Learning All our Worst Impulses," *The Guardian*, August 8, 2017, available at https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses.

10 Bajkowski, J. "Federal Court bins Robodebt's Defective Algorithm," *iTNews*, November 27, 2019, available at https://www.itnews.com.au/news/federal-court-bins-robodebts-defective-algorithm-534677.

11 See Desai, D. R. and Kroll, J. A. "Trust but Verify: A Guide to Algorithms and the Law," *Harvard Journal of Law & Technology* (31:1), Fall 2017.

viability of its output can help humans provide explanations that are sufficiently meaningful.

Dimension 6: Environment

Understanding the boundaries of the environment in which an AI system operates can help to explain its decisions, even when it is not possible to explain the workings of the underlying algorithm. The importance of defining and knowing the environmental boundaries for an AI system is illustrated by the well-publicized case of Amazon's Alexa operating beyond its intended use context by recording personal conversations and emailing them to another Alexa user.[12] Although Alexa's actions seemed inexplicable at first, approaching them from the perspective of environmental boundaries helps to explain what was going on: although the AI system "thought" it was operating in a particular environment (i.e., taking orders from its human owner), it was, in fact, receiving input data from a context in which it should not have been operating (a private conversation between two humans).

These examples quoted above for each of the six dimensions of our framework suggest that explanations of the behavior of an AI system should holistically take account of all six dimensions. This is precisely the approach adopted by the Machine Learning Lab at the Danish Business Authority (DBA), as described below. This case study identified novel tools for tackling the challenges of explaining how AI applications reach their decisions, even though the inner workings were not always entirely explainable. This approach has enabled the DBA to implement AI applications responsibly and legally.

Machine-Learning AI Applications at the Danish Business Authority

The Danish Business Authority is an agency within Denmark's Ministry of Industry, Business and Financial Affairs. It has approximately 700 employees, divided between the headquarters in Copenhagen and two satellite departments in Silkeborg and Nykøbing Falster. Its primary responsibility is to enhance opportunities for business growth in Denmark, but it also has specific regulatory obligations, such as fraud prevention and supervision of companies without imposing an unnecessary administrative burden on the Danish business community. One of the DBA's obligations is to maintain and apply laws such as Denmark's Companies Act, Financial Statements Act, Bookkeeping Act and Commercial Foundation Act.

To facilitate the activities associated with these obligations, the DBA operates a multi-agency online platform called Virk (https://virk.dk). Citizens can use Virk, for example, to establish or shut down business enterprises, handle various registrations and submit documents such as financial reports electronically. The online business register contains approximately 809,000 companies, with 812,000 registrations, and filings of 292,000 annual reports. Annual reports are submitted in two formats: PDF documents to be read by humans, and documents in structured data format XBRL (eXtensible Business Reporting Language) to be automatically machine-processed. The sheer volume of data presents the DBA with ample opportunities to pursue machine learning for such core tasks as supporting companies' legal compliance, checking annual reports for signs of fraud, and identifying companies early on their route to distress so that timely support can be given.

Because of the large data volumes involved, the DBA established its Machine Learning Lab in 2017 to implement machine-learning projects for greater efficiency and scalability. The lab's team leader and chief data scientist, "James,"[13] stated the following:

> *"We are, in essence, trying to use [machine learning] as a force multiplier for our colleagues performing the controls but also trying to lessen the manual workload and reserving the human decision making for the more creative or advanced tasks."*

The lab uses technologies such as Neo4j's platform, Docker and Python[14] for the development, application and support of machine-learning AI applications, rather than

12 See Warren, T. "Amazon Explains How Alexa Recorded a Private Conversation and Sent it to Another User," *The Verge*, May 24, 2018, available at https://www.theverge.com/2018/5/24/17391898/amazon-alexa-private-conversation-recording-explanation.

13 Pseudonyms are used for all informants to protect their identity.

14 For information about Neo4j and its products, see https://neo4j.com/company/.

commercial off-the-shelf solutions. The lab develops functional prototypes of machine-learning applications that are capable of solving business problems specified by case workers: "*We are focusing on meeting the information need that the business has,*" James explained. A DBA steering committee decides whether to move a prototype to production use. When the committee decides in favor of implementation, an external vendor then implements the machine-learning application for real-world deployment. "David," an Early Warning Europe[15] case worker, elaborated on the importance of this type of governance:

> *"It's really easy to end up on the front page of a tabloid newspaper. … This is why [we] make sure the model is only handed over from the partner organizations to stakeholders through a package of management consultancy training, capacity-building, documentation, all these support services, where we make sure that at least they know the logic of using it."*

At a higher level, the lab is engaged in a wider dialogue about the use of AI in government and was recently involved in the Danish National Strategy for AI, with a particular focus on the transparent application of AI in the public sector.[16]

Denmark, in general, and specifically the DBA, is considered to be at the forefront of e-government initiatives globally. According to a recent UN report,[17] Denmark is a world leader in e-government development. Within the EU, Denmark is ranked first for the provision of e-government services for businesses,[18] and it was also ranked fourth in the EU's Digital Economy and Society Index (DESI),[19] where it was listed among the leaders in digital public services. Furthermore, Europe's Digital Progress Report specifically highlighted the DBA's Virk portal, noting that roughly 96% of Danish businesses make use of Virk. However, the high level of digitization and digitalization driven by the DBA in Denmark has not been accompanied by adverse media comments about digital government experienced by other countries. For these reasons, we consider the DBA to be a legitimate source for best practice in organizational use of AI. Conducting a case study of the DBA's development and implementation of AI applications enabled us to learn from a well-performing organization in the field of government IT. Details of the case study are in Appendix A.

How the Danish Business Authority Applied the Framework

The DBA's approach to explaining the behavior of its AI applications is characterized by limiting the capacities of AI agents while still obtaining the desired outputs from the applications.[20] The approach took account of all six dimensions of the framework described above: the choice of AI model, the goals of the AI application, the training data, input and output data, and boundaries of the environment in which the AI application operates. The actions taken by the DBA in all of these dimensions are summarized in Figure 2.

The key benefit of holistically managing the six very different dimensions is gaining a better understanding of, and control over, the outputs from AI applications, which enables the organization to prevent or at least mitigate any undesired outcomes. By establishing and knowing the boundaries of an AI system's operation, the organization has a better understanding of the system's capacity to act. Within these boundaries, AI solutions can be harnessed to maximum advantage—even those with models that are seemingly inexplicable. Thus, instead of being

15 Early Warning Europe provides free, impartial and confidential counselling to companies in distress. For more information, see https://www.earlywarningeurope.eu/.

16 See National Strategy for Artificial Intelligence, 2019, available at https://eng.em.dk/media/13081/305755-gb-version_4k.pdf

17 E-Government Survey 2020: Digital Government in the Decade of Action for Sustainable Development, United Nations Department of Economic and Social Affairs, August 24, 2020, available at https://publicadministration.un.org/egovkb/en-us/Reports/UN-E-Government-Survey-2020.

18 See eGovernment Benchmark 2019: trust in government is increasingly important for people, European Commission, October 18, 2019, available at https://ec.europa.eu/digital-single-market/en/news/egovernment-benchmark-2019-trust-government-increasingly-important-people.

19 The Digital Economy and Society Index (DESI), European Commission, 2019.

20 An example of limiting an AI system's capacity provided in Robbins, S. "AI and the Path to Envelopment: Knowledge As a First Step Towards the Responsible Regulation and Use of AI-Powered Machines," *AI & Society* (35), April 10, 2019, pp 391-400.

Figure 2: The DBA's Approach Took Account of all Six Dimensions of the Framework

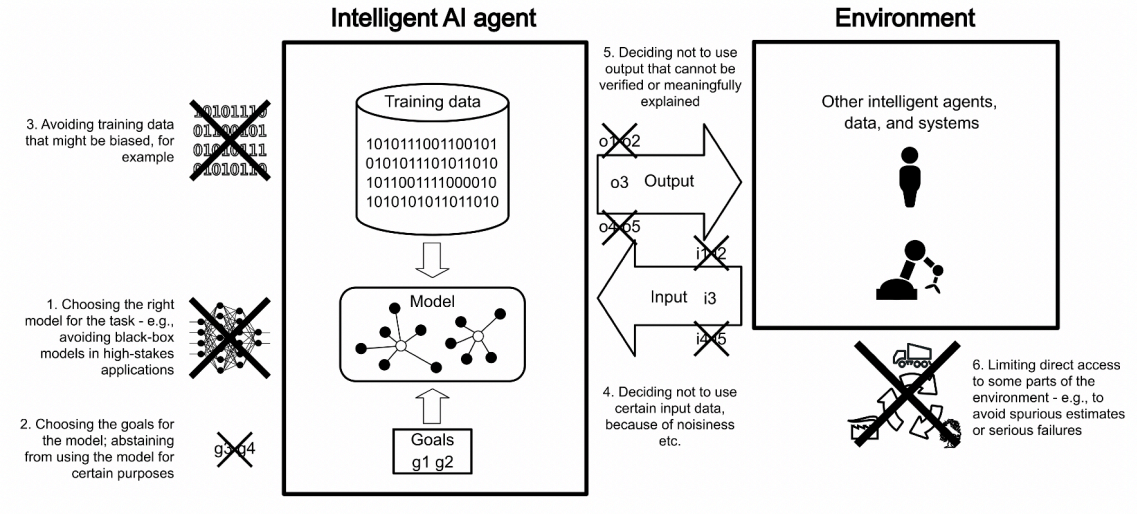


Table 2: Examples of AI Applications at the DBA

Project	Project Description	Goal	Input	Output	Model
Company Registration	To detect fraudulent behavior among newly registered Danish companies.	To prevent fraudulent companies from being established.	Data from the business registry, annual reports and VAT reports.	Probability of fraudulent behavior.	Gradient boosting (XGBoost).[21]
Signature	When coupled with its document filter, to speed up verification of whether company founding documents are signed or not.	To facilitate the process of founding a company.	Scanned images of the founding documents.	Probability of a document being signed or not.	Residual network (ResNet-16).[22]

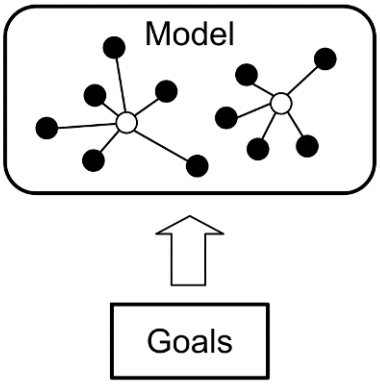
viewed as a method for producing technical explanations, the DBA's approach provides mechanisms for understanding and controlling the behavior of AI applications.

Below, we discuss in detail how the DBA's approach took account of each of the six dimensions. To demonstrate how the authority's actions were implemented, we provide examples from two ongoing projects at the DBA. These two projects, which exploit AI in different ways, are summarized in Table 2. (A full list of the AI projects being undertaken by the DBA is given in Appendix B.) The purpose of the first project, "Company Registration," is to prevent people from establishing companies for fraudulent purposes—i.e., creating companies that were never intended for the stated business objectives, but instead have ulterior, fraudulent motives behind them. In contrast, the aim of the second project, "Signature," is to facilitate the process of creating legitimate companies by detecting the absence of signatures from the documents

21 Gradient boosting is a machine-learning technique for regression and classification problems. XGBoost is an open-source software library for gradient boosting frameworks.

22 The residual-network technique uses machine-learning based on deep neural networks and is especially powerful in image-detection tasks. ResNet-16 is a residual-network technique whose architecture has 16 neural network layers.

Figure 3: Factors to Consider When Choosing and Controlling Training Data



- Consider how explainable the AI use case needs to be
- Select a model whose structure is not too open-ended, to avoid excessive flexibility that could allow learning from harmful spurious correlations
- Use structures that mirror the nature of the underlying problem

- Choose concrete and unambiguous performance metrics that reflect underlying business goals
- Analyze the pros and cons of the performance metrics carefully, and discuss the choices with various stakeholders

When contemplating the choice of model, ask:

- *Is this level of complexity necessary for achieving the required functionality?*
- *Could sufficient performance be obtained by means of a simpler alternative?*
- *Will the main users of the model need to explain its functioning to other people?*

required to found a company. Together, these two projects illustrate how the DBA's approach took account of all six dimensions of the framework for explaining the behavior of black-box AI systems.

Choosing the AI Model and Setting Goals (Dimensions 1 and 2)

To ensure that an AI application meets users' requirements, developers must carefully choose the system's model and goals (i.e., performance metrics), taking account of the need to be able to explain the outcome in a specific use case (see Figure 3).

Clearly, the demand for an explainable model depends on the type of project. For the Company Registration application, the DBA opted for an explainable model, because users must be able to understand readily why the algorithm has raised a red flag for a newly registered company:

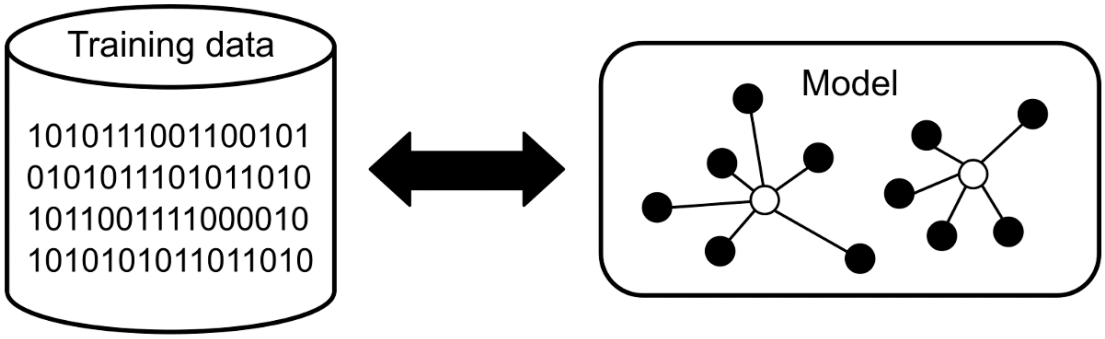
> *"We need to communicate the results and our findings to the case workers, so we try to use algorithms that are not complicated … or at least algorithms that can fairly easily give you some sense of which are the most important factors and which are not. So, [we need] explainable algorithms. … I guess that the more difficult it is for the case worker to actually see right away what the right answer is, the more important it is for the algorithm to be able to explain itself."* "Mark," a data scientist at the DBS Machine Learning Lab

The model chosen for an AI application has direct implications for how explainable the outputs from the application will be. Sometimes, though not always, choosing the model requires a tradeoff between performance and the transparency of the model selected. In most cases at the DBA, however, performance losses resulting from transparency demands have been negligible, as emphasized by James:

> *"We … [compared] a number of models, and gradient boosting came out as number one. We could have chosen deep learning [or] a deep neural network, but we chose not to, because we find [it would be] too complex to explain."*

In the Signature application, however, which is essentially an image-recognition application

Figure 4: Factors to Consider When Establishing Controls for Input and Output Data



- Gather an adequate set of training data
- Explore patterns and how they might influence the AI system's behavior
- Critically assess variables and their measurement
- Identify biases, and apply corrections if needed
- Prevent uncontrolled self-learning from potentially biased incoming data

When controlling training data, ask:

- *What kind of data exactly is needed for making the decisions?*
- *Are we using that sort of data, or something that is of lower quality and perhaps even biased, merely because it is more easily accessible?*
- *Is the data well suited to the type of AI model that we are using?*

using neural networks, it is perfectly acceptable to use a black-box approach. This is because users can easily verify whether the model works correctly or not without the need to understand its internal logic in great depth.

When choosing the goals and performance metrics for an AI application, our DBA interviewees emphasized that there is no silver bullet. Mark (a data scientist), reflected on how to prioritize among multiple performance metrics:

> *"... you could focus on the precision of the model. For example, how well does it predict [compared to our predictions] of ... fraudulent behavior in the future? How many would be correctly classified? But if [the case workers] have enough time on their hands, it might be [worthwhile looking retrospectively to] see how many of the companies [predicted to commit fraud] actually [do]. But that would give probably more work to the case workers. ... It depends on the situation, and it's a dialogue with the case workers exactly [as to] which metrics are the most important ones in each case."*

Clearly, the successful choice of metrics is highly problem-specific and requires both thorough understanding of the nature of the underlying data and solid domain expertise.

Understanding and Controlling the Training Data (Dimension 3)

Training data plays a key role in determining how an AI application works once deployed (see Figure 4). At the DBA, managers were well aware of the need for high-quality training data: "*It is my head on the line if it seems that the data is not good enough or [the data] is biased*" ("Steven," a data scientist at the Machine Learning Lab).

Controlling training data requires access to sufficient quantities of data and in-depth knowledge of the data, including any inherent biases and limitations. For example, if training data is limited to smaller companies, the implications of this bias should be assessed and the model's applicability may be narrower than initially assumed (the AI application might be suitable only for smaller firms). To ensure the application is relevant for companies of all sizes, any such biases in the training data will need to be corrected. To guarantee high-quality training data for both the Signature and Company Registration AI applications, the DBA opted to tag a large body of data manually, using domain experts as consultants in this process. In the words of Steven:

> *"We had tagged data, we had around 6,000 tagged documents, so we had a pile of [documents] that had not been used in training or in developing, so we just made sure that those were the ones we tested on and made sure that they had a fair distribution of different [outcomes]. … We asked domain specialists, 'Is this an accurate picture, or is it not?' and they said it was, so that's what we [decided we were] going with."*

The Machine Learning Lab's methods for controlling training data enable it to trace changes in an AI application's behavior. This is especially important for countering "data drift"—changes (or drift) in underlying data-generating processes that mean an AI application trained on historical data alone is unable to produce equally valid outputs as the future unfolds. The DBA has experienced some data-drift problems as fraudulent companies change their behavior over the years. For example, the strategies that sham companies use to commit tax fraud tend to evolve over time. This problem needs to be addressed by critical evaluation of training data and possibly by revising or updating the data used.

Responding to the challenge of data (and concept) drift also has implications for the choice of model. Developers can choose from a wide spectrum of models, ranging from offline batch-learning models, which treat data as a static pool and become smarter only when given a new batch of data to learn from, to online self-learning models that learn autonomously from a growing pool of data. For the latter models, the same input can produce different outputs at different times because the system learns "on the fly" and adapts to new information. Online self-learning models can be an appealing option for countering data drift, because of their ability to adapt. "Daniel," a case worker who uses the Company Registration AI application, said that *"we would very much like models that tell us, 'Look at these areas,' areas we didn't even think about. 'Look at these because … there [seems to be] something rotten going on here.'"*

To retain control over training data, the DBA has opted for batch training, not self-learning. This approach to controlling training data helps it to minimize uncertainty stemming from the data and aids in evaluating the outputs from partly or entirely inscrutable systems. In the words of "Jason," a team leader at the Machine Learning Lab. *"… we have made a conscious decision not to use self-learning technologies—i.e., that we'll train a model [on a certain dataset], and then we accept that it will not become smart until we retrain it."*

Controlling Input and Output Data (Dimensions 4 and 5)

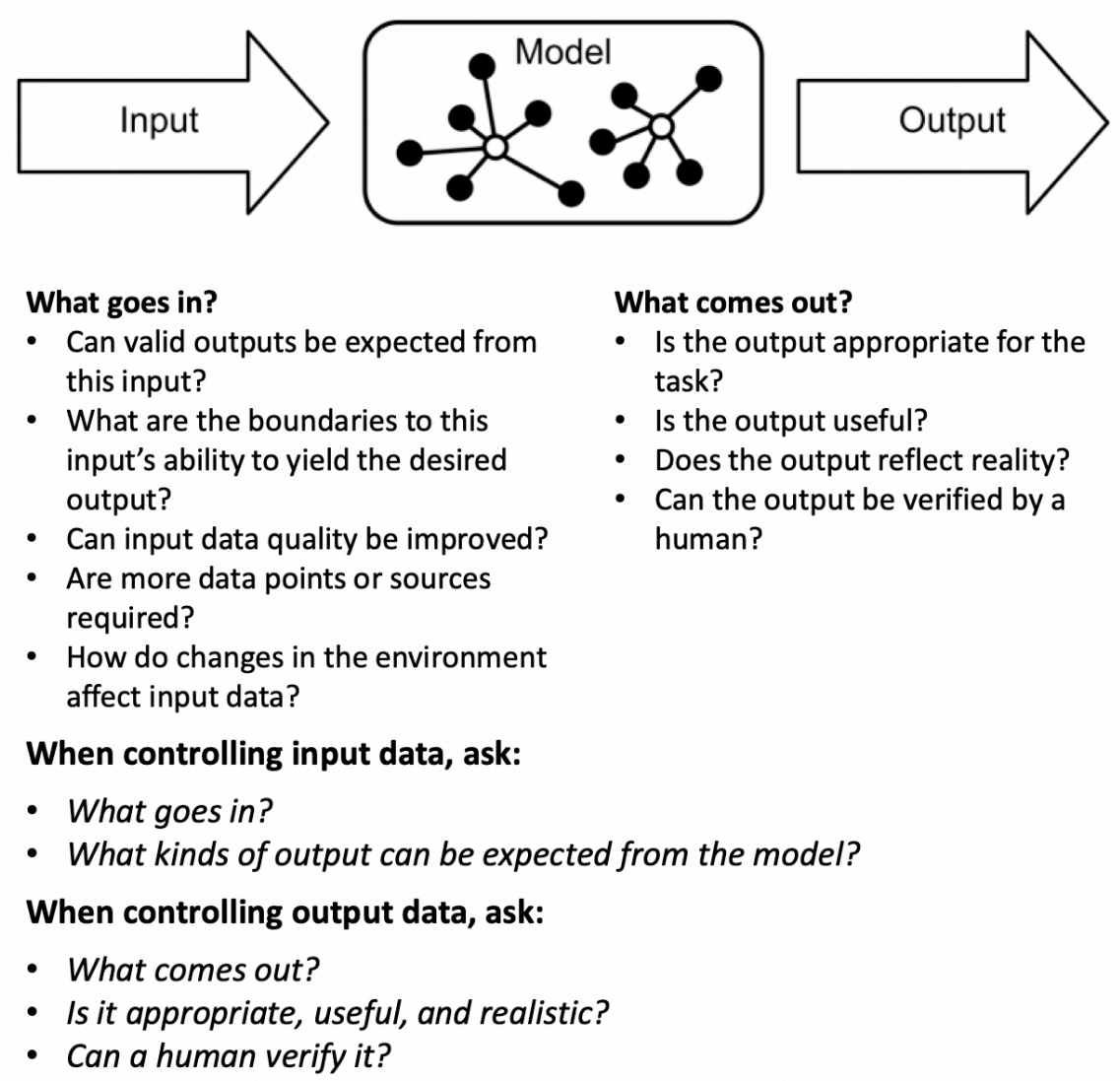
At the DBA, controlling input and output data focuses on understanding what goes into and what comes out of an AI application (see Figure 5). Similar to controlling training data, input control emphasizes the quality of the data processed by the model. In the words of Jason:

> *"When we have a good understanding of where our data comes from, what has influenced [that] data, the causal relation between [input and output data], we understand where, how, and why something happened."*

Low-quality input data can lead to biased or nonusable outputs even if the model has been properly trained. In some cases, the DBA has been able to improve the usability of the output data by preprocessing the input data. For instance, in a project involving citizen-uploaded photos of personal identification documents, rotating the photos before feeding them into the model improved the model's performance significantly.

Controlling output data involves verifying the results produced by an AI application. These

Figure 5: Factors to Consider When Establishing Controls for Input and Output Data

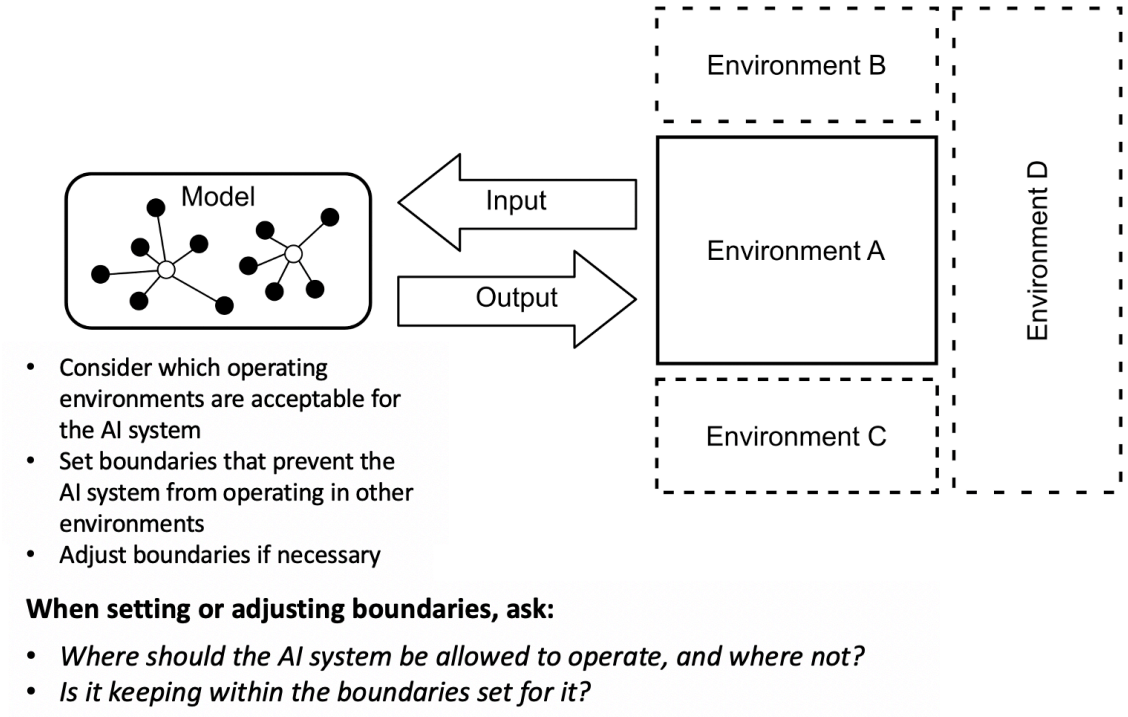


actions may be automated or done manually. For example, case workers using the Signature application manually check documents that the AI application judged to be incomplete, and where the model has expressed low confidence in the correctness of its decision. These outputs are very easy to verify—case workers can determine the completeness of a document by glancing through the relevant fields. This demonstrates that human verification of the output from an AI application does not always require special knowledge of the inner workings of the model, as Steven explained:

> *"If a person calls and asks, 'Why was my document rejected? then a case worker will say, 'That's because you haven't signed it.' 'How do you know that?' 'I have looked at the document. It isn't signed.' So they don't have to answer, 'Well, the neural network said it's because of a variable 644 in the corner.' That's why you can get away with using a neural network in this case, [even though you can't explain how it works]."*

Controlling the inputs to and outputs from an AI application allows the use of inscrutable

Figure 6: Factors to Consider When Setting the Environment Boundaries for an AI Application



models that are hard to explain, provided the organization has the ability to judge the quality of the input and output data.

Setting an AI Application's Environment Boundaries (Dimension 6)

An environment-centered approach to explaining the behavior of an AI application consists of setting clear boundaries for the application's area of operation (see Figure 6). One of the external vendor's testers of the Signature application discovered that the algorithm trained to detect signatures on scanned images of the documents required to found a company would accept an image of a wooden toy animal as valid input and classify it as a signed document. In other words, the application was operating outside its intended environment. To ensure that the application only operated within its appropriate boundary, the DBA created a filter to determine whether the image received is indeed of a document before the image is input into the AI system.

To simplify boundary setting, the DBA designed a software architecture comprising many simple models that operate in highly specific areas, performing very specific actions. This architecture confines each AI application to a limited area, within which its outputs can be easily analyzed. This architecture also limits the damage a malfunctioning application can cause, because the impact is contained in one area. Jason explained, *"By having an event-driven architecture, you can rely on loosely coupled systems, and having sound metadata will help you create order in the chaos of different systems interacting with the same data."*

The architecture also offers a safe and legally compliant way of using black-box AI systems where necessary. The fact that none of the DBA's AI applications make any final decisions affecting citizens or organizations imposes operational boundaries for the applications and also links boundary setting with output control. In many AI applications at the DBA, users have some degree of control over the extent of the operating

environment. For example, in the Company Registration application, case workers are able to adjust critical thresholds for the application to make sure they yield the most useful and precise recommendations possible. Jason explained that this has also facilitated workers' acceptance of the models: *"I was surprised to see the idea of [a] control tower. The ability to mute a model or change the threshold has been a major cultural factor in [the] business adoption of this technology."*

In summary, the environment boundaries of the DBA's AI applications are set through combinations of technological mechanisms (e.g., system design) and managerial controls (e.g., of case workers). However, expert users can gradually develop better rules and tune the boundary thresholds. It is also noteworthy that the boundary for an AI application need not coincide with the boundary between the organization and the external environment. An internal boundary can limit an AI application's effect on the organization's internal operations. For example, the outputs from the Signature project are passed on to another internal agent who continues the processing of the documents deemed by the application to have a valid signature.

In conclusion, the DBA case demonstrates that taking account of all six dimensions of the framework for explaining the behavior of AI systems enables the successful and responsible deployment of various AI applications, even black-box algorithms that are not technically explainable.

Recommendations for Explaining the Behavior of Black-Box AI Systems

Based on our analysis of the DBA case, we provide four recommendations for practitioners. These recommendations encompass both managerial and technological approaches for tackling the challenges of explaining the behavior of black-box algorithms used in AI applications.

1. Implement Strict Controls on the Use of Black-Box AI Systems

Taking account holistically of all six dimensions of the framework described above enables the use of inexplicable black-box AI systems without compromising the safety of operations. Decisions to use such systems depend on the application context and on whether a comprehensive set of control measures is available for the specific application. For instance, Jason stated that a neural network *"has a higher degree of precision but [lacks] transparency; ... we only apply them in areas with low impact or an otherwise objective relation to falseness."*

Our analysis indicates that the use of a black-box AI system, such as a deep neural network, is permissible if:

1. There is minimal possibility of the inscrutability of the system resulting in increased hazards for human stakeholders' wellbeing
2. Using a black-box system does not violate any laws that require the workings of the system to be explained to users
3. The impact of the AI system can be strictly bounded within an internal environment and its output can be controlled by humans.

For instance, the DBA's Signature application rejects a document only if a human can verify the AI application's decision as valid and assume responsibility for the actions that follow. The involvement of human workers can make this approach costly, but the benefits for the DBA, mainly in the form of efficiency gains, have outweighed the additional costs. The DBA case workers can easily screen the problematic documents out of the workflow and devote their cognitive capacity to higher-level activities.

2. Use Modular Design to Make it Easier to Explain the Behavior of an AI System

Breaking complex business processes into smaller modules that can be supported by narrow and well-defined AI applications can make it easier to control and explain the outputs of the applications. For example, designing an AI application to operate a specific function within a process, rather than making it responsible for the entire process, helps to guarantee that it does not—and indeed cannot—obtain data from environments that it should not touch. This means that the developers and users of such AI applications have a high degree of control over

the application's functionality throughout the development and deployment process. They can have greater confidence in the application's outputs and are well placed to detect deviations early on and diagnose any problems that might occur. In essence, modular design of AI systems is akin to a divide-and-conquer approach: rather than try to create an entire explainable system, it is easier to start with multiple explainable pieces that together constitute a bigger AI system. Jason described the DBA's approach as *"feeding the dragon one little biscuit at a time, so we can design models that can be brought into production."*

3. Avoid Online Learning if the Need for Explanation is a Priority

Online learning is appealing for AI applications that have high needs to adapt to environmental changes, but it makes it more difficult to monitor and explain how such an application functions. Online learning therefore results in a reduced level of control and may even prove dangerous in some high-stakes applications. An AI application that learns while operating poses a risk of introducing bias that is not evident from the original design of the system, and that could be challenging to detect and rectify. Difficulty in testing and understanding the behavior of AI systems that use online learning makes it harder to explain how they produce their outputs.

The DBA opted to train its AI applications in a controlled, stepwise manner. This approach protects the applications from the unintended "overfitting"[23] and bias that less controlled learning mechanisms could easily introduce. Note, however, that there is a clear tradeoff between the adaptiveness of the learning mechanism and improved explanation capabilities resulting from offline training. Without subsequent online learning, AI applications trained via offline data may not remain current:

> *"... control departments would rather say, 'We have seen one case that looked like this. Dear machine, find me cases that are exactly the same.' And we have tried to tell them that 'that's fine—we had a case years ago where there were a lot of bakeries that committed a lot of fraud, but now it doesn't make sense to look for bakeries anymore, because now those bakeries are selling flowers or making computers or something different.'"* Daniel, user of the DBA's Company Registration AI application

4. Facilitate Continuous Open Discussion Between Stakeholders

The first three recommendations raise important questions concerned with ethics and responsibility, such as how to determine what is considered biased and who should have the final say in this. We therefore recommend that organizations involve various stakeholders, with distinct perspectives and expertise in the development of AI applications. Beware, though, that involving stakeholders with different backgrounds, approaches and work roles may create obstacles to their ability to communicate with each other. Mark, a data scientist at the DBA Machine Learning Lab, explained:

> *"I think the difficult part has been to get the dialogue with the case workers, who see the world in a different way. ... What exactly is it we should feed the model for getting good predictions, and how do we get the information from the case workers?"*

Communication barriers can be overcome by facilitating further discussion through workshops that involve multiple stakeholders. For example, a data scientist's ability to explain the relevant AI algorithm to domain experts serves as a Litmus test for the ease with which the workings of an AI system can be explained. The DBA's efforts to facilitate dialogue between data scientists and domain experts increased understanding on both sides. The data scientists were able to incorporate important domain-specific factors into the design of AI applications, and the domain experts simultaneously became more informed about the structure of the applications and their operational boundaries.

In addition to focusing on the expected effects on internal stakeholders, the discussions should also consider the implications of using AI systems for the wider business community, economy and society in general. At the DBA, mechanisms such as steering committee reviews

23 Overfitting is where a model accurately describes random errors in the current data to an extent that results in poor fit with future input data.

improve the management of critical ethics-related repercussions that tend to accompany the introduction of AI technologies.

Concluding Comments

There is a compelling need to be able to explain how AI systems operate, and much of the current research on the challenges organizations face in implementing AI is focused on this area. Many recent media reports attest to the disruptive, trust-eroding effects that irresponsible AI implementation can have on organizations and on society. At the same time, advances in AI technologies make it increasingly difficult to develop cutting-edge AI applications whose algorithm-driven decision making can be easily explained.

The Danish Business Authority case study reported in this article provides fresh insights for organizations that want to responsibly deploy complex AI systems in their operations. Some elements of the DBA's approach to making AI systems more explainable are visible in various other organizations, at least tacitly: paying greater attention to the quality of training data and using human oversight to control outputs are now common practices. Our analysis of the DBA's approach shows that taking account of the six dimensions of our framework for explaining the behavior of black-box AI systems can facilitate the successful introduction of AI. Because the DBA is a public-sector organization, it has especially high transparency requirements and has therefore developed tools and management procedures for explaining how its AI applications reach their decisions.

Private-sector businesses may not feel they have as compelling a need to make their AI systems explainable, so—at least at present—they may find less-comprehensive approaches than the DBA's sufficient. Nevertheless, our four recommendations for explaining the behavior of black-box AI systems are equally applicable to public- and private-sector organizations. All organizations, whether public or private, are under mounting pressure to deploy AI-based applications to improve their efficiency and effectiveness while simultaneously demonstrating accountability and responsibility to stakeholders through their ability to explain the algorithm-driven decision making of their AI applications.

In today's business environment, all organizations face constant changes in legislation, norms, codes of ethics, technologies, strategic goals, and the data they generate and use. The controls, choices and boundaries for AI systems are therefore determined by the circumstances that exist when they are set and must be managed if they are to retain their effectiveness over time. To ensure that suitable resources are available for this task, issues relating to explainable AI must be considered when preparing an AI application for production use and throughout its life. The DBA has adopted just such a practice: at set intervals, there is a review of the activities related to each AI application, and the associated costs are factored in from the implementation phase onward. This practice involves collecting feedback from application users and from data scientists on the algorithms' operation, with the functionality being adjusted accordingly.

Organizations should therefore plan to keep their tools and strategies for explaining the workings of their AI systems current through constantly evaluating and retraining their AI systems. We believe the four recommendations we have provided for using the framework for explaining the behavior of black-box AI systems will help organizations effectively address the caveats of such systems while still reaping their significant performance benefits, both now and into the future.

Appendix A: The Danish Business Authority Case Study

Between August 2018 and January 2020, we collected interview and observation data at the DBA. The data was obtained and analyzed through an iterative four-phase process (see the table below), with the phases overlapping and earlier phases informing subsequent ones. We sought to interview a wide range of employees and managers, at several levels in the DBA and with a wide range of tenure, to ensure the data was not biased by the views of long-term or more recent employees.

Phase 1 was largely exploratory and established research collaboration and identified research questions. Phase 2 focused on obtaining in-depth knowledge of the DBA's AI projects and the actors involved. Phase 3 focused specifically

The Four Data-Collection Phases

Phase No. and Data-Collection Theme	Method	Duration (minutes)	Interviewees' Pseudonyms and Roles	Focus of Outcomes
1. Machine Learning Lab Projects Overall	Group interview	105	James (team leader/chief data scientist); Mary (chief consultant, in Annual Reports)	Responsibilities of the DBA; organization structure
2. Machine Learning Lab Functions	Personal interview	90	James	The role of explainability in AI projects; allocation of tasks among stakeholders (the Machine Learning Lab, implementation unit and case workers)
	Group interview	83	David and John (both Early Warning Europe case workers)	
	Personal interview	70	Daniel (an internal case worker in Company Registration)	
	Personal interview	59	Steven (a data scientist)	
	Personal interview	51	Mary	
	Personal interview	116	James	
3. Explainability in AI Projects	Personal interview	51	Steven	Practical means to address explainability issues; the sociotechnical environment of model development
	Personal interview	54	Thomas (a data scientist)	
	Personal interview	50	Linda (a data scientist)	
	Personal interview	48	Michael (a data scientist)	
	Personal interview	52	Mark (a data scientist)	
	Personal interview	53	Joseph (a data scientist)	
	Personal interview	54	Jason (a team leader)	
	Personal interview	48	Susan (a data scientist)	
	Personal interview	62	William (an internal case worker in Company Registration)	
	Personal interview	54	Daniel	
4. Verification of Interpretations from Analysis	Personal interview	55	Jason	Validation of interpretations via interviews and an assessment exercise involving project template mapping
	Assessment exercise	N/A	Steven; Mary; Thomas; Linda; Michael; Mark; Joseph; Jason; Susan	

on explainability and involved all the Machine Learning Lab's employees and two case workers. Finally, Phase 4 focused on validating the interpretations from the analysis of the data collected and gaining fuller insights into the technical infrastructure supporting the lab. Data scientists participated in an assessment exercise with the authors by mapping a descriptive framework for every project conducted by the lab.

The interviews were recorded and then transcribed into 153,195 words of text. The interview data was supplemented with observations carried out by the authors and by document analysis. One of the authors, who has previously worked at the DBA, kept a field diary,

Appendix B: AI Projects at the DBA

Project Name	Project Description
Auditor's Statement	The Auditor's Statement algorithm speeds up verification that the valuations of company assets given in an auditor's statement are correct and that the statement does not include violations. The algorithm is used by internal DBA case workers.
Bankruptcy	The Bankruptcy algorithm predicts company distress and insolvency. It ties in with the Early Warning Europe (EWE) initiative. The algorithm is used by external consultants in the EWE community in Denmark and elsewhere in the European Union. The DBA is not responsible for actions and consequences related to this tool.
Company Registration	The Company Registration algorithm aims to detect fraud-indicative behavior among newly registered Danish companies. The algorithm is used internally by DBA case workers.
Land and Buildings	The Land and Buildings algorithm predicts violations of accounting policies related to property holdings and long term investments. The algorithm is used by internal DBA domain experts.
Passport	The Passport algorithm expedites processing of the documents submitted, supplying a text string from the machine-readable portion of a passport and comparing it with input data from the user. The algorithm is used by internal DBA case workers.
Recommendation	The Recommendation algorithm improves the user experience of the DBA's Virk portal by focusing on personalized content and optimized interfaces. The algorithm improves the portal's usability for external customers.
Sector Code	The Sector Code algorithm speeds up verification of a company's industry-sector code. As of the third quarter of 2020, 25% of company codes were incorrect. The algorithm is used by internal DBA case workers.
Signature	The Signature algorithm, in combination with the associated document filter, speeds up verification of whether company founding documents are signed. The algorithm is used by internal DBA case workers and returns three probabilities: whether the document is physically signed, whether it is digitally signed and whether the signature is missing.

recording observations and taking notes from informal conversations and meetings. This diary dates back to September 2017, when most of the projects were just beginning.

About the Authors

Aleksandre Asatiani

Dr. Aleksandre Asatiani (aleksandre.asatiani@ait.gu.se) is an assistant professor in information systems in the Department of Applied Information Technology, University of Gothenburg, Sweden. He is also affiliated with the Swedish Center for Digital Innovation (SCDI). His research focuses on artificial intelligence, robotic process automation, virtual organizations and IS sourcing. His work has been published in leading IS journals such as *Information Systems Journal* and *Journal of Information Technology*.

Pekka Malo

Pekka Malo (pekka.malo@aalto.fi) is a tenured associate professor of statistics at Aalto University School of Business, Finland. His research has been published in leading journals in operations research, information science and artificial intelligence. Pekka is considered as one of the pioneers in the development of evolutionary optimization algorithms for solving challenging bilevel programming problems. His research interests include business analytics, computational statistics, machine learning, optimization and evolutionary computation, and their applications to marketing, finance and healthcare.

Per Rådberg Nagbøl

Per Rådberg Nagbøl (pena@itu.dk) is a Ph.D. fellow at the IT University of Copenhagen doing a collaborative Ph.D. with the Danish Business

Authority within the field of information systems. He uses action design research to design systems and processes for quality assurance and evaluation of machine learning, focusing on accurate, transparent and responsible use in the public sector from a risk management perspective.

Esko Penttinen

Dr. Esko Penttinen (esko.penttinen@aalto.fi) is a professor of practice in information systems at Aalto University School of Business, Finland. He studies the organizational implementation of artificial intelligence, interplay between humans and machines, and governance issues related to outsourcing and virtual organizing. His main practical expertise lies in the assimilation and economic implications of interorganizational information systems, focusing on application areas such as electronic financial systems, government reporting and electronic invoicing. His research has been published in leading IS journals such as *MIS Quarterly, Information Systems Journal, Journal of Information Technology, International Journal of Electronic Commerce* and *Electronic Markets*.

Tapani Rinta-Kahila

Dr. Tapani Rinta-Kahila (t.rintakahila@uq.edu.au) is a postdoctoral research fellow at the UQ Business School and Australian Institute for Business and Economics, University of Queensland, Australia. His Ph.D. in information systems science was awarded by the Aalto University School of Business. His research addresses issues related to the decommissioning of IT systems, organizational implementation of artificial intelligence and automation, and the dark side of IS.

Antti Salovaara

Dr. Antti Salovaara (antti.salovaara@aalto.fi) is a senior lecturer at Aalto University, Department of Design, Finland, and an adjunct professor in the Department of Computer Science, University of Helsinki. He studies human-AI collaboration and online trolling, and the methodology of user studies. His research has been published in conference proceedings such as CHI, and in leading journals, including *Human Computer Interaction, International Journal of Human-Computer Studies, MIS Quarterly* and *European Journal of Information Systems*.

EGOV-CeDEM-ePart

X-RAI: A Framework for the Transparent, Responsible, and Accurate Use of Machine Learning in the Public Sector

Per Rådberg Nagbøl*, Oliver Müller**

*IT University of Copenhagen, Denmark, pena@itu.dk
**University of Paderborn, Germany, oliver.mueller@uni-paderborn.de

Abstract: This paper reports on an Action Design Research project taking place in the Danish Business Authority focusing on quality assurance and evaluation of machine learning models in production. The design artifact is a Framework (X-RAI) which stands for Transparency (X-Ray), Responsible(R), and explainable (X-AI). X-RAI consist of four sub-frameworks: the Model Impact and Clarification Framework, Evaluation Plan Framework, Evaluation Support Framework, and Retraining Execution Framework for machine learning that builds upon the theory of interpretable AI and practical experiences tested on nine different machine learning models used by the Danish Business Authority.

Keywords: Machine Learning Evaluation, Government, Interpretability

Acknowledgement: Thanks to all the involved employees at the Danish Business Authority

1. Introduction

Recent years have seen breakthroughs in the field of AI, both in terms of basic research and development as well as in applying AI to real-world tasks. The AI Index 2019 Annual Report of the Stanford Institute for Human-Centered Artificial Intelligence (Perraul et al., 2019), which summarizes the technical progress in specialized tasks across computer vision and natural language processing, attests that AI is now on par or has even exceeded human performance in tasks such as object classification, speech recognition, translation, and textual and visual question answering. However, augmenting and automating tasks previously performed by humans can also lead to serious problems. Research studies and real-world incidents have shown that AI systems–or better the machine learning models they are based on-– can err, encode societal biases, and discriminate against minorities. These issues are amplified by the fact that many modern machine learning algorithms are complex black boxes whose behavior and predictions are almost impossible to comprehend, even for experts. Hence, more and more researchers and politicians are calling for legal and ethical frameworks for designing and auditing these systems (Guszcza et al. 2018). Against this background, the government of Denmark released a national strategy for AI in 2019. The strategy

covers a broad array of initiatives related to AI in the private and public sectors, including an initiative concerning the transparent application of AI in the public sector. As part of this initiative, common guidelines and methods will be created to enforce the legislation's requirements for transparency. As one of the first steps, the government launched a pilot project to develop and test methods for ensuring a responsible and transparent use of AI for supporting decision making processes (Regeringen, 2019). The pilot project takes place at the Danish Business Authority (DBA) in collaboration with the Danish Agency of Digitization. In this paper, we report on the first results of an Action Design Research (ADR) project accompanying the pilot project. The overall ADR project is driven by the following research question: How do we ensure that machine learning (ML) models meet and maintain quality standards regarding interpretability and responsibility in a governmental setting? To answer this question, the project draws on literature and theory on interpretability of machine learning models and practical testing on machine learning models in the DBA.

2. Explainable AI Through Interpretable Machine Learning Models

Modern machine learning algorithms, especially deep neural networks, possess remarkable predictive power. However, they also have their limitations and drawbacks. One of the most significant challenges is their lack of transparency. Complex neural networks are opaque functions often containing tens of millions of parameters that jointly define how input data (e.g., a picture of a person) is mapped into output data (e.g., the predicted gender or age of the person in the picture). Hence, it is virtually impossible for end users, and even technical experts, to comprehend the general logic of these models and explain how they make specific predictions. As long as one is only interested in the predictions of a black box model and these predictions are correct, this lack of transparency is not necessarily a problem. Broadly speaking, there are two alternative approaches to open up the black box of modern machine learning models (in the following see Lipton, 2018, Molnar, 2019, Du et al., 2020). First, instead of using black box deep learning models, one can use less complex but transparent models, like rule-based systems or statistical learning models (e.g. linear regression, decision trees). These systems are intrinsically interpretable, but the interpretability often comes at the cost of sacrificing some predictive accuracy. The transparency of these systems works on three levels: Simulatability concerns the entirety of the model and requires models to be rather simple and ideally human computable. Decomposability addresses interpretability of the components of the model, such as, inputs, parameters, and calculations.

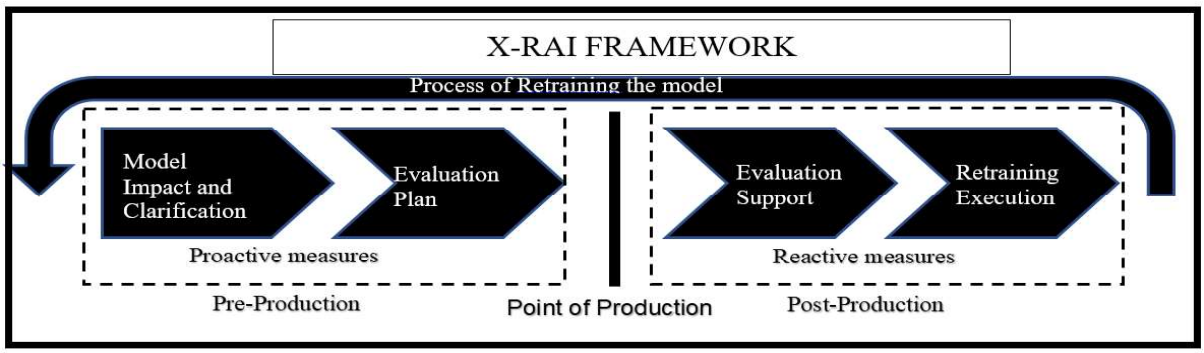
Consequently, decomposability requires interpretable model inputs and disallows highly engineered or anonymous features. Algorithmic transparency concerns the training/learning algorithm. A linear model's behavior on unseen data is provable, which is not the case with deep learning methods with unclear inner workings. Second, instead of using transparent and inherently interpretable models, one can develop a second model that tries to provide explanations for an existing black box model. This strategy tries to combine the predictive accuracy of modern machine learning algorithms with the interpretability of statistical models. These so-called post-hoc examinability techniques can be further divided into techniques for local and global explanations. Local explanations are explanations for particular predictions, while global explanations are explanations that provide a global understanding of the input-output relationships learned by the

trained model. In other words, a local explanation would explain why a concrete person on a picture has been predicted to be female, while global explanations would explain what general visual features differentiate females from other genders. Different types of post-hoc explanations exist. Text explanations use an approach similar to how humans explain choices by having a model generating explanations as a supplement to a model delivering predictions. Visualizations generate explanations from a learned model through a qualitative assessment of the visualization. Explanations by example let the model provide examples showing the decisions the model predicts to be most similar (Lipton, 2016). Local Explanations for particular predictions (Doshi-Velez & Kim, 2017) such as Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) and SHAP for explaining feature importance (Lundberg, S & Lee, S, 2017). Focusing on the local dependence of a model helpful when working with neural networks being too incomprehensible to explain the full mapping learned satisfactorily (Lipton, 2016). When choosing which approach and technique to use in order to create an explainable AI system, it is worth to consider *why* there is a need for explanation (e.g., to justify decisions, enhance trust, show correctness, ensure fairness, and comply with ethical or legal standards), *who* the target audience is (e.g., a regular user, an expert user, or an external entity), *what* interpretations are derivable to satisfy the need, *when* is the need for information (before, during, or after the task), and *how* can objective and subjective measures evaluate the system (Rosenfeld, A & Richardson, A, 2019).

3. The X-RAI Framework as a Design Artifact

The X-RAI framework is an ensemble consisting of four artifacts (Fig. 1). First, the Model Impact and Clarification (MIC) Framework, which ensures that a ML model fulfills requirements regarding transparency and responsibility. Second, the Evaluation Plan (EP) Framework, which plans resource requirements and the evaluation of ML models. Third, the Evaluation Support (ES) Framework that facilitates the actual empirical evaluation of ML models and supports the decision whether a ML model shall continue in production, be retrained or shut down. Fourth, the Retraining Execution (RE) Framework, which initiates the process of sending an ML model back to the Machine Learning Lab (ML Lab) for retraining.

Figure 12: The X-RAI Framework



The first two artifacts are part of the decisive foundation for a steering committee regarding launching the ML model into production (pre-production). The last two artifacts support the

continuous evaluation and improvement of the ML model after it goes live (post-production). The design artifacts in ADR are solutions to problems experienced in practice and with theory ingrained. The problems must be generalizable outside the context of the project (Sein et al., 2011). X-RAI is a solution to problems experienced in the context of the Danish Business Authority where government officials are the intended end users. The government officials are, in our case, educated within the sciences of law, business, and politics as well as data scientists with plural backgrounds. Their expertise varies according to the governmental institution. X-RAI must be capable of involving and utilizing stakeholders with variating expertise without excluding some by setting an unachievable technological barrier of entry.

3.1. Model Impact and Clarification Framework

The MIC Framework has been applied and tested on four ML models--three times in its initial version and one time in its current version. The MIC is a questionnaire that enables the questionee to describe and elaborate on issues related to different aspects of ML related to transparency, explainability, responsible conduct, business objectives, data, and technical issues. The primary purpose of the MIC Framework is to improve, clarify, and guide communication between various stakeholders, such as developers with technical expertise, caseworkers with expertise in the ML models decision space and management. The idea of the MIC Framework derives from an analysis of the Canadian Algorithmic Impact Assessment (AIA)[1] tool that was found to have a strong link to the Canadian directive on automated decision-making[2]. MIC differs from AIA since it is grounded in theory and business needs instead of legislation. The algorithmic information in Box 1 contains information about the ML model. Box 2 is filled out by the future owner of the system enabling them to state their needs concerning the use, explainability, transparency, users, and accountable actors. Box 3 builds directly on Lipton's descriptions of transparency with the following three sub-levels: simulatability, decomposability, algorithmic transparency. In addition, it builds on types of post-hoc interpretability with the following approaches: text explanations, visualization, local Explanations, and explanation by example (Lipton, 2016). These are supplemented with three concrete explainability methods, Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) and SHAP (Lundberg, S & Lee, S, 2017). The output verification is bound to the fact that ML models in the DBA are decision-supportive, not decision-making, which reduces the need for an explanation if the end-user can validate the truthfulness of the model output instantly. Box 4 focuses on the data dimensions of the ML model including the relation to data sources and other ML models. Box 5 explains every feature to avoid opaque ML models due to highly engineered or anonymous features (Lipton, 2016) and supplements methods such as SHAP (Lundberg, S & Lee, S, 2017). Box 6 draws on the special categories from the 2016 European Union's General Data Protection Regulation[3] and the 2018 Danish Data Protection Act[4], repeating the questions on other data

1 See https://canada-ca.github.io/aia-eia-js/ and https://github.com/canada-ca/digital-playbook-guide-numerique/tree/master/en

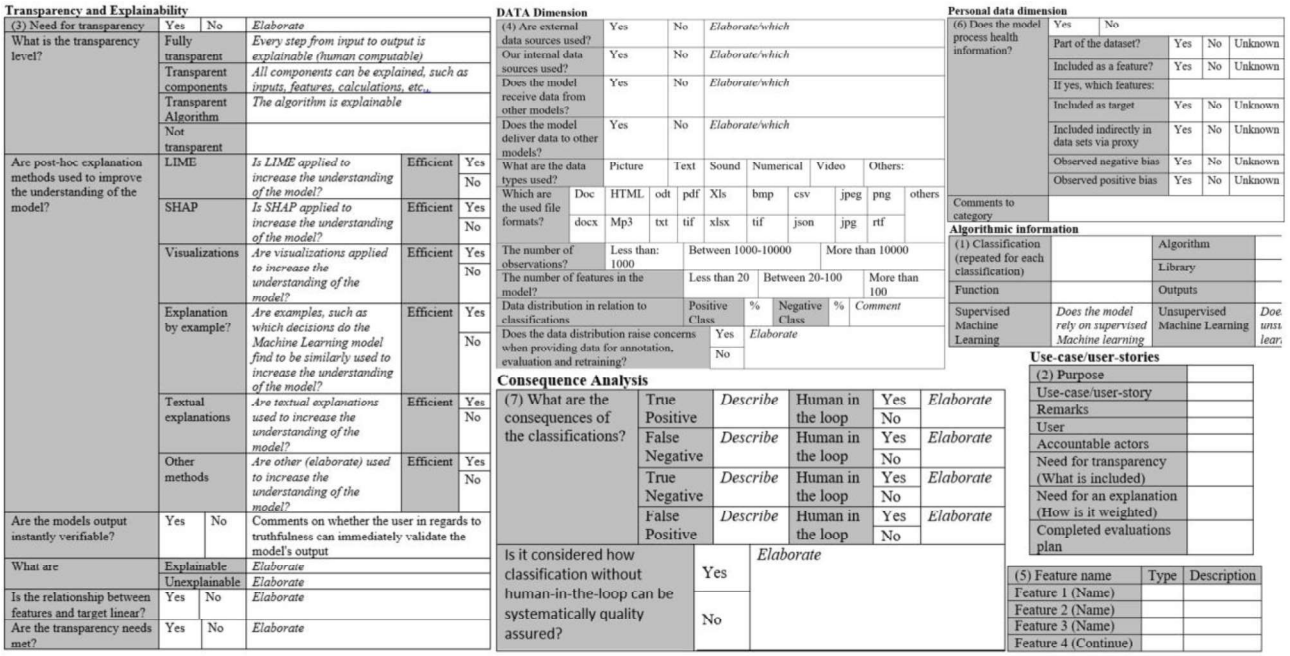
2 See https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

3 See https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02016R0679-20160504&from=EN

4 See https://www.retsinformation.dk/Forms/r0710.aspx?id=201319 (all links last checked 01/06/20)

categories to avoid discrimination. Box 7 focuses on the consequences of the output, mitigation of consequences, and ensuring the responsible application of ML models. It takes inspiration from the confusion matrix enabling an easy estimate of the frequency of each outfall.

Figure 13: Model Impact and Clarification Framework



3.2. Evaluation Plan

The Evaluation Plan (EP) was applied and tested on eight ML models in three incrementally different versions. The EP structures the ongoing evaluation of a ML model throughout its lifetime and thereby illuminates the necessary resources for maintenance. The Evaluation Plan clarifies uncertainties such as time and frequency for the evaluation meetings, involved actors including roles and obligations, data foundation, and meeting preparation. The goal is to ensure that all ML models fulfill the defined quality requirements from the cradle to the grave. The theory is ingrained indirectly in the EP through the MIC framework. The choices made when using the MIC framework influences how the ML model can be evaluated. The ML model's degrees of transparency and explainability influences the possibilities of the evaluations. The evaluation detects data drift in a procedure similar to the application-grounded evaluation where the ML model is evaluated accordingly to domain experts performance on the task (Doshi-Velez & Kim, 2017). The EP encourages the first evaluation to be as early as possible due to the difficulties in predicting complex methods such as neural network on unseen data (Lipton, 2016).

Figure 14: Evaluation Plan Framework

(1) The name of the model and version number
(2) Participants for an example the application manager, caseworkers, ML lab etc.
(3) When is the first evaluation meeting?
(4) Expected evaluation meeting frequency: (How often are we expected to meet? And are there peak periods which we need to take into consideration?)
(5) Foundation for evaluation: For an example logging data or annotated data (Annotated data is here data where the domain experts classification is compared to the machine)
(6) Resources: (who can create the evaluation/training data, internal vs. external creation of training data, what is the quantity needed for evaluation, time/money)
(7) Estimated resource requirement for training, training frequency, and complications degree (procedure regarding regular bad performance)
(8) The Role of the Model: Is it visible or invisible for external users.
(9) Is the models output input for another/is the models input an output from another model.
(10) What are the criteria of success and failure (When does a model perform good/bad. How many percent?)
(11) Is there future legislation that will impact the model performance? (Including: bias, introduction of new requirements/legal claims, abolition of requirements/legal claims, bias, etc..
(12) When does the model need to be retrained?
(13) When should the model be mutet?

3.3. Evaluation Support

The Evaluation Support (ES) framework was applied five times on three different ML models in three incrementally changed editions.

Figure 15: Evaluation Support Framework

(1) The name of the model and version number
(2) Date of evaluation
(3) When was the last evaluation of the model?
(4) What was the result of the last evaluation?
(5) Participants in the evaluation meeting
(6) Who is doing the current evaluations?
(7) How many cases/documents has been processed in the evaluation (find minimum)
(8) Was the data used for the evaluation satisfying?
(9) Was is the result of the evaluation
(10) Has the performance of the model decreased?
(11) Has the performance of the model increased?
(12) What is the threshold set at?
(13) What is the history of the threshold setting?
(14) Should the threshold level be changed?
(15) Why is the threshold setting changed?
(16) Does the model still satisfy a business need? If not should the model then be shut down?
(17) Is there future legislation that will impact the model performance? (Including: bias, introduction of new requirements/legal claims, abolition of requirements/legal claims, bias, etc..
(18) Should the model be retrained based on the evaluation?

A fourth edition is ready for testing. The ES facilitates the evaluation of the ML model at the evaluation meetings. The domain specialist responsible for the ML model answers relevant fields in the framework before the meeting. The stakeholders complete the remaining framework collaboratively at the meeting and decide if the ML model shall continue in production, be retrained, or shut down. The ES strives to evaluate the ML model accordingly to the task as described in the

applications-grounded evaluation (Doshi-Velez & Kim, 2017). In our case, we let the caseworker that normally would do the task of the ML model evaluate the classifications and report it in the ES framework. The ES primarily focuses on fulfillments of performance requirements while it lets transparency and explainability be subcomponents of interpreting the reason for ML model performance. The reason is important if the model needs retraining.

3.4. Retraining Execution Framework

The Retraining Execution (RE) Framework was applied and tested two times on two different ML models in two incrementally changed versions. The RE initiates the process of sending a ML model back to the machine-learning lab for retraining. The retraining occurs when the ML model needs to improve performance and will continue to provide value. The RE framework focuses on the reusability of evaluation data and old training data for retraining, the occurrence of new technological possibilities, the detection and elimination of bias, changes in data types and legislation, the urgency for retraining, and if the input and output are related to other models. Transparency and explainability of the ML model become relevant when explaining a root cause for the need for retraining.

Figure 16: Retraining Execution Framework

(1) The name of the model and version number
(2) What is the reason for having the model retrained?
(3) What is the result of the last evaluation?
(4) Own suggestion of root cause, why does the model need retraining? (changes in document type, legislation, tenders etc..)
(5) Is new training data available for retraining (including estimation of required resources)
(6) How important is it to have the model retrained?
(7) Is the model dependent on other models? Yes/no – what is the status on them?
(8) What is the status of training data in the current situation? (Changes in document form, legislation, tenders, etc..)
(9) Can new data be added to the existing data or is there a need for a whole new training dataset? (What old training data is reusable?)
(10) Observed suspicion (bias against industry, gender, business type, etc.) Is it a problem? Yes/No
(11) Is the models output input for other models? Yes/no – status on them
(12) Is there developed algorithms that can solve the problem better since the model was put in production?
(13) "concluding text felt" Is there taken a decision regarding the model need to be retrained? (Has all stakeholder agreed on that the model has to be retrained?)

Data distribution becomes relevant if the data are skewed and slows down and thereby increases the cost in a data annotation process with the focus on providing training examples for the minority class. The use of the retraining execution framework restarts the X-RAI process by leading to the use of the MIC framework.

4. Conclusion and Outlook

The X-RAI framework was successfully developed, applied, and tested on nine different ML models used in the Danish Business Authority accordingly to the ADR principle of authentic and concurrent

evaluation (Sein et al.. 2011). The iterations have let to incremental changes in the frameworks. The frameworks are currently standard procedures and mandatory for all ML models developed by the ML Lab in the Danish Business Authority, which we conclude to be successful in the aspect of organizational adoption of artifacts and procedures. Artifacts must have theory ingrained accordingly to ADR (Sein et al.. 2011). Interpretability theory, including the subcategories of transparency and explanation, is ingrained into the frameworks. The lens provides a strong foundation for informing how the ML models work. Future work will focus on analyzing the evaluation data and using it to design IT artifacts and integrate them into the Danish Business Authority's IT-ecosystem. An additional theoretical lens will be ingrained in the artifacts to create a theoretical foundation for responsible conduct in the design.

References

Perraul, R. & Shoham, Y. & Brynjolfsson, E. & Clark, J. & Etchemendy, J. & Grosz, B. & Lyons, T. & Manyika, J. & Mishra, S. & Niebles, J. C. (2019). The AI Index 2019 Annual Report. AI Index Steering Committee, Human-Centered AI Institute, Stanford University.

Guszcza, J. & Rahwan, I. & Bible, W. & Cebrian, C. & Katyal, V. (2018) Why We Need to Audit Algorithms. https://hbr.org/2018/11/why-we-need-to-audit-algorithms.

Regeringen (2019) Finansministeriet og Erhvervsministeriet: National strategi for kunstig intelligens

Molnar. C. (2020): Interpretable Machine Learning A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/

Doshi-Velez, F. & Kim, B. (2017) Towards a rigorous science of interpretable machine learning. https://arxiv.org/abs/1702.08608v2

Du, M. & Liu, N. & Hu, X. (2020). Techniques for Interpretable Machine Learning. Communications of the ACM. Volume 63. Issue 1. https://dl.acm.org/doi/10.1145/3359786

Rosenfeld, A. & Richardson, A (2019). Explainability in Human-Agent Systems. arXiv:1904.08123v1

Sein, M.K. & Henfridsson, O. & Purao, S. & Rossi, M. & Lindgren, R. (2011) ACTION DESIGN RESEARCH. MIS Quarterly, Volume 35, Issue 1, page 37-56

Lipton, Z (2016) The Mythos of interpretability. Presented at 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY. last revised 6 Mar 2017. arXiv:1606.03490v3

Lipton, Z (2018). The Mythos of Model Interpretability. ACM QUEUE. Volume 16, issue 3 https://queue.acm.org/detail.cfm?id=3241340

Lundberg, S & Lee, S (2017). A Unified Approach to Interpreting Model Predictions. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

About the Authors

Per Rådberg Nagbøl

Per Rådberg Nagbøl is employed as a Ph.D. fellow at The IT University of Copenhagen and does a collaborative Ph.D. in collaboration with the Danish Business Authority.

Oliver Müller

Oliver Müller is Professor of Management Information Systems and Data Analytics at Paderborn University.

Designing a Risk Assessment Tool for Artificial Intelligence Systems

Per Rådberg Nagbøl[1], Oliver Müller[2], and Oliver Krancher[1]
1 IT University of Copenhagen, Rued Langgaards Vej 7, 2300 Copenhagen, Denmark
{pena,olik}@itu.dk
2 Paderborn University, Warburger Str. 100, 33098 Paderborn, Germany
oliver.mueller@upb.de

PREPRINT
The final authenticated version is available online at https://doi.org/10.1007/978-3-030-82405-1_32

Abstract. Notwithstanding its potential benefits, organizational AI use can lead to unintended consequences like opaque decision-making processes or biased decisions. Hence, a key challenge for organizations these days is to implement procedures that can be used to assess and mitigate the risks of organizational AI use. Although public awareness of AI-related risks is growing, the extant literature provides limited guidance to organizations on how to assess and manage AI risks. Against this background, we conducted an Action Design Research project in collaboration with a government agency with a pioneering AI practice to iteratively build, implement, and evaluate the Artificial Intelligence Risk Assessment (AIRA) tool. Besides the theory-ingrained and empirically evaluated AIRA tool, our key contribution is a set of five design principles for instantiating further instances of this class of artifacts. In comparison to existing AI risk assessment tools, our work emphasizes communication between stakeholders of diverse expertise, estimating the expected real-world positive and negative consequences of AI use, and incorporating performance metrics beyond predictive accuracy, including thus assessments of privacy, fairness, and interpretability.

Keywords: AI · Risk assessment · Risk management · Interpretability · Envelopment

1 Introduction

Artificial Intelligence (AI) technologies such as machine learning (ML) allow an increasing number of organizations to improve decision-making and automate processes [1]. Notwithstanding these potential benefits, organizational AI use can lead to undesired outcomes, including lack of accountability, unstable decision quality, discrimination, and the resulting breaches of the law [2]. For instance, media and academia have revealed cases of algorithmic discrimination concerning facial recognition [3], crime prediction [4], online ad delivery [5], and skin cancer detection [6].

Drawing on the risk management literature [7, 8], we refer to such potential undesired outcomes as risks. Given the increasing adoption of AI, a key challenge for organizations these days is implementing procedures that prevent or mitigate risks from organizational AI use. A critical task in this regard is to assess (i.e., identify, analyze, and prioritize) [8] the risks associated with a new AI system (i.e., a software system based on AI) before its go-live. Risk assessment is critical for responsible organizational AI use because it allows organizations to make informed decisions grounded in a thorough understanding of the risks and benefits of using a specific AI system and because risk assessment is the foundation for risk control [8] after go-live.

The risk management literature, governmental frameworks, and the AI literature provide some foundations for understanding how organizations should assess risks from organizational AI use. Two key insights from the risk management literature are that risk management is a knowledge integration process involving business and technical stakeholders [9, 10] and that risk management operates within a tension between template-based deliberate analysis and expert intuition [8, 11]. Governmental frameworks, such as Canada's Directive on Automated Decision-Making [12], provide blueprints for risk assessment templates. The AI literature provides methods for data and model documentation [13, 14], for improving the interpretability of ML models [15], and for identifying biases [16, 17]. The AI literature has recently also advanced the concept of envelopment [18–20] to explain how organizations can address risks by limiting the agentic properties of AI technologies [21].

Although these foundations are valuable, the existing literature provides limited guidance to organizations on assessing AI risks because of two fundamental limitations. First, there is little research that explicitly takes a risk management perspective on AI. While most AI research does not explicitly draw on risk management theory [13, 14], the risk management literature does not focus on AI, examining instead risks associated with information system (IS) projects [8, 9] or with traditional software and hardware [22]. However, AI systems differ from these two in that AI systems are software (unlike IS projects) with agentic qualities (unlike traditional hardware and software) [21]. Second, given the conceptual nature of most work [20], there is a lack of empirical research that is grounded in the experience of real organizations in assessing AI-related risks. Given these gaps, our paper addresses the following research question: *How should procedures be designed to assess the risks associated with a new AI system?*

We address this research question through an Action Design Research (ADR) study [23]. We worked together with a governmental agency with a pioneering AI practice to iteratively build, implement, and evaluate the AI Risk Assessment (AIRA) tool. Our key contributions are theory-engrained and empirically validated design principles for assessing risks associated with new AI systems.

2 Literature Background

2.1 The AI Literature

There is a rapidly growing body of research from computer science and IS on AI, defined as *"systems that display intelligent behavior by analyzing their environment and taking actions – with some degree of autonomy – to achieve specific goals."* [24]. Although AI research has rarely paid explicit attention to risk assessment of new AI systems, three streams within AI research provide important perspectives on this issue: research on interpretability, on envelopment, and on dataset and model documentation.

Interpretability. The main argument why we grounded our artifact in the literature on interpretable AI is that insights into the process of algorithmic decision making enable the early detection of unintended outcomes and side-effects, hence lowering overall risk. We rely on Lipton's [15] conceptualization of interpretability with the subcategories transparency and post-hoc interpretability. Transparency refers to AI systems that are inherently understandable for humans, such as linear models and decision trees. It comprises the criteria simulatability of the model as a whole (e.g., whether a human can trace how the model transforms inputs into outputs), decomposability of its individual components (e.g., the decision rules and parameters of a model), and transparency of the learning algorithm (e.g., how a model learns its decision rules or parameters) [15]. Posthoc interpretability is an alternative to inherent transparency. For complex and opaque AI systems, it might be possible to construct a faithful abstraction of the

original black-box model that is understandable for humans (e.g., a visualization, an example-based explanation) [15]. Such post-hoc explanations can focus on an individual prediction (local explanations) or on the general patterns the model has learned (global explanations) [15].

Envelopment. Envelopment theory provides conceptual guidance for enhancing the safety of AI systems in production environments. Envelopment—a term borrowed from the field of robotics—describes how micro-environments are enveloped around robots' three-dimensional space enabling them to achieve their purpose successfully while preventing damaging people or material [20, 25, 26]. Although the concept is originally
from the physical space, Robbins suggested that the areas to be addressed by an AI system can also be enveloped into a confined virtual space. These areas are training data (its suitability for production environments), boundaries (expected scenarios and possible inputs including data types), input (how all sensed data are combined), function (the purpose of the AI), and output (the AI's production utilized to fulfill its function) [20]. For instance, an organization may envelop training data by stipulating that the model needs to be retrained with new training data if significant environmental changes question the suitability of the training data for the current production environment [20].

Model Documentation. Datasheets for datasets guides the communication between dataset creators and dataset consumers to enhance transparency and accountability. Datasets are accompanied by a datasheet documenting key aspects such as composition, collection, and cleaning [13]. Model cards for model reporting has been developed to supplement datasheets for datasets and follows a similar logic. Model cards are documentations that accompany trained ML models. The model cards contain information related to the application domain [14]. Reactive approaches are developed to audit the performance of facial recognition classifiers performance across different genders and skin colors [16, 17].

2.2 Risk Management

We draw on the risk management literature as one foundation for understanding how organizations can assess potential undesired outcomes of using an AI system. Risk management is frequently conceptualized as a process that starts with risk assessment, consisting of risk identification, risk analysis, and risk prioritization, followed by risk control [7, 8]. Our paper focuses on risk assessment. Although most of the IS risk management literature focuses on risks associated with IS projects, the literature offers two key ideas that are potentially relevant for the risk assessment of AI systems.

First, risk management is a knowledge integration process involving business and technical stakeholders. Wallace et al. [10] showed that problems in IS projects often have their origin in social-subsystem risks (e.g., unstable environments, user resistance), which translate into technical risks and project management risks. In line with these ideas, it has been shown that knowledge integration between technical and business stakeholders is key for addressing risks in IS projects [9]. Although IS projects are different from organizational AI use, organizational AI use is, like an IS project, a sociotechnical system in which users delegate their work to AI systems and the development of these AI systems to developers and data scientists [21], presenting thus a need for knowledge integration between users and data scientists.

Second, risk management operates within a tension between template-based deliberate analysis and expert intuition. The bulk of academic risk management research suggests that deliberate efforts to identify, analyze, and prioritize risks are beneficial because they help to capture a wider range of risks [8] efficiently. For instance, risk managers were shown to capture a wider range of risks when they performed a deliberate risk analysis based on templates [27].

However, another strand of the risk management literature emphasizes the key role of expert intuition for mindfully identifying and focusing on relevant risks [28], suggesting that risk assessment often requires a balance between document-based and expertise-based approaches.

3 The Action Design Research Project

The Action Design Research (ADR) project described in this paper is a university government collaboration between the Danish Business Authority (DBA) and the IT University of Copenhagen. The DBA is a Danish government agency with approximately 700 employees. The DBA offers services like the cross-governmental platform virk.dk, Covid-19 compensation, the central business register, and annual reporting to Danish and foreign businesses. It has deployed 22 AI systems to support employees in operational decision making and automation of routine tasks. The DBA presented an ideal setting for our study given its intensive use of AI, the high level of digitization in Denmark [29,30], and the strategic priority of ensuring responsible AI use in the Danish public sector [31].

The artifact developed in this ADR project was the AI Risk Assessment (AIRA) tool. The AIRA is designed to be the first out of four artifacts in the X-RAI framework [32]. Its key purpose was to assess the risks associated with a new AI system. We developed the AIRA tool between April 2019 and March 2021 through three iterations of building, evaluating, and testing (see Table 1) [23]. During this time, the first author of this paper spent approximately every other week at the DBA. Everyday interactions and meetings with DBA employees, especially around 30 meetings, including 12 one-on-one sessions with the ML lab team leader, shaped its design. These interactions have led to a rich empirical base consisting of transcripts, field notes, documents, and artifacts.

Table 1. Overview for application, test and evaluation of AIRA on AI systems

AI systems	Test approach (artifact version)
Business document compliance validator	Framework (v1) filled out at the meeting
Document preprocesing filter	Framework (v1) filled out at the meeting
Identification check	Framework (v1) filled during two meetings
Compensation	Framework (v2.1.1) filled out during two recorded Microsoft Teams interviews
Fraud	Framework (v2.1.3) filled out at the meeting
Industry code selector	Framework (v3.0.1. ML part) filled out pre meeting and evaluated at the meeting
Identification check	Framework (v3.0.1. ML part) filled out
Bankruptcy report	Frameworks (v3.0.1. Business part and v3.0.1. ML part) filled out before the meeting for discussion and evaluated at the meeting (recorded)
Fixed costs compensation	Frameworks (v3.0.2. Business part, v3.0.4. ML part, and v3.0.3. Facilitator part) filled out before the meeting and discussed at the meeting
Salary compensation	Frameworks (v3.0.2. Business part, v3.0.4. ML part, and v3.0.3. Facilitator part) filled out before the meeting and discussed at the meeting
Self-employed compensation	Frameworks (v3.0.2. Business part and v3.0.4. ML part) filled out before the meeting and discussed at the meeting

Iteration #1: The initial design of the AIRA tool was inspired by the Algorithmic Impact Assessment (AIA) tool of the Canadian government. Although the AIA tool served as a blueprint, key stakeholder at the DBA found that the AIA tool did not focus enough on algorithms and data, lacked clear roles and responsibilities, and was tailored to Canadian law. Hence, using the AIA tool as a source of inspiration, the ADR team *built* an initial alpha version of the AIRA tool consisting of ten questions. The questions addressed areas such as algorithms (e.g., underlying learning algorithms and used libraries), training data (e.g., types and sources of data), predictive performance (e.g., a confusion matrix incl. description of the consequences of each cell, the existence of ground truth), interpretability (e.g., use of post-hoc explainability methods), and decision making (e.g., is there a human-in-the-loop?). The organizational *intervention* occurred by applying the tool on three AI systems in collaboration with data scientists from the DBA. The evaluation happened in the form of feedback from the team leader of the DBA's ML Lab. The *evaluation* found that the general idea was likely to work in the context of the DBA and that the tool should be expanded to include user stories from a business perspective and data privacy. In addition, the desire to calculate a risk score, just like in the Canadian AIA tool, was articulated.

Iteration #2: The second iteration focused on expanding the contents of the tool. The *building* phase concentrated on identifying further relevant areas which need to be covered for risk assessment (e.g., a more detailed description of the purpose of the AI system from a business perspective). In addition, the level of detail for assessing the training data aspect was increased considerably. The *intervention* occurred by applying the artifact to two additional AI systems. The concurrent *evaluation* yielded two key findings. First, it was important to acknowledge the knowledge differences between different people and roles involved. Data scientists had problems answering questions related to business objectives and the business need for model interpretability, as one data scientist formulated it: *"…The need for transparency is defined by the business unit. I just try to build the best model for a given need of transparency. It is business who needs to define the requirements for transparency and how these requirements need to be understood."* (Data scientist 1). Second, it was found that going through the questionnaire from start to end was too time-consuming and that different stakeholders should contribute to different parts. Henceforth, the artifact should be filled out before the meeting and discussed at the meeting. The ADR team also realized that the original idea of automatically calculating a risk score, like in the Canadian AIA tool, was complicated by numerous context dependencies and interdependencies between questions.

Iteration #3: Based on the feedback from the previous iteration, we focused the *building* phase on restructuring the questionnaire into self-contained modules for distinct stakeholders and improving the overall user experience in terms of required time and knowledge. The first module initiated the assessment process and is to be filled out by a future user of the AI system (i.e., the business unit). The second module was filled out by those building the model (i.e., data scientists). The third module was filled out in collaboration between the user (domain experts) and data scientists in a physical meeting moderated by a facilitator. The *intervention* phase included applying the tool to six AI systems. The *evaluation* suggested potential for improvement regarding the readability of some questions and the preparation time required for participants.

4 The Artificial Intelligence Risk Assessment Tool

Figure 1a provides a schematic overview of the final version of the AIRA tool. The tool contains three modules, each targeted at a different audience. We will now describe the structure and contents of these modules in more detail.

The first module is targeted at the business unit that will use the AI system and focuses on eliciting requirements from a business perspective. Amongst others, the module contains a consequences matrix showing potential positive and negative consequences of deploying the AI system (see Fig. 1b for an example). Inspired by the concept of a confusion matrix, it asks domain experts for a qualitative description of the consequences of these four types of outcomes. Following the idea of expected utility theory [33, 34] the combination of this information with quantitative data from a classical confusion matrix (which is included in the second module of the tool, see Fig. 1) allows assessing the chances and risks of deploying the AI system. The assessment is complemented by information describing if a human receives the output of the AI system and if a human can instantly verify the truthfulness of the output.

The second module is meant to be filled out by the data scientist responsible for developing the AI system. The main themes covered in this module are the predictive performance, training data, interpretability of the model and its outputs, and its interfaces and boundaries. The interpretability part is based on the concepts and categorizations proposed by Lipton. With regards to transparency, the data scientist is, for instance, asked whether they are able to describe how the algorithm discovers decision rules (algorithmic transparency) and how these rules are later used to make predictions for specific cases (simulatability). If the AI system is based on a black box algorithm, questions regarding local and global post-explainability are asked. Another important part of the module is related to the processing of personal data. Drawing on the EU GDPR, it is checked whether the AI system processes protected personal attributes (e.g., gender, ethnicity, age) and if the model has been checked for potential biases and discrimination against these groups. At this, six types of biases (historical, representation, measuring, aggregation, evaluation, and implementation) [35] and metrics for their detection (e.g., Equal Opportunity Difference, Disparate Outcomes) are considered. Finally, the interface of the AI system to other downstream models (e.g., to discover potential chain reactions if the model fails) and potential boundary conditions (e.g., In which situations should be the model not be used?) are documented.

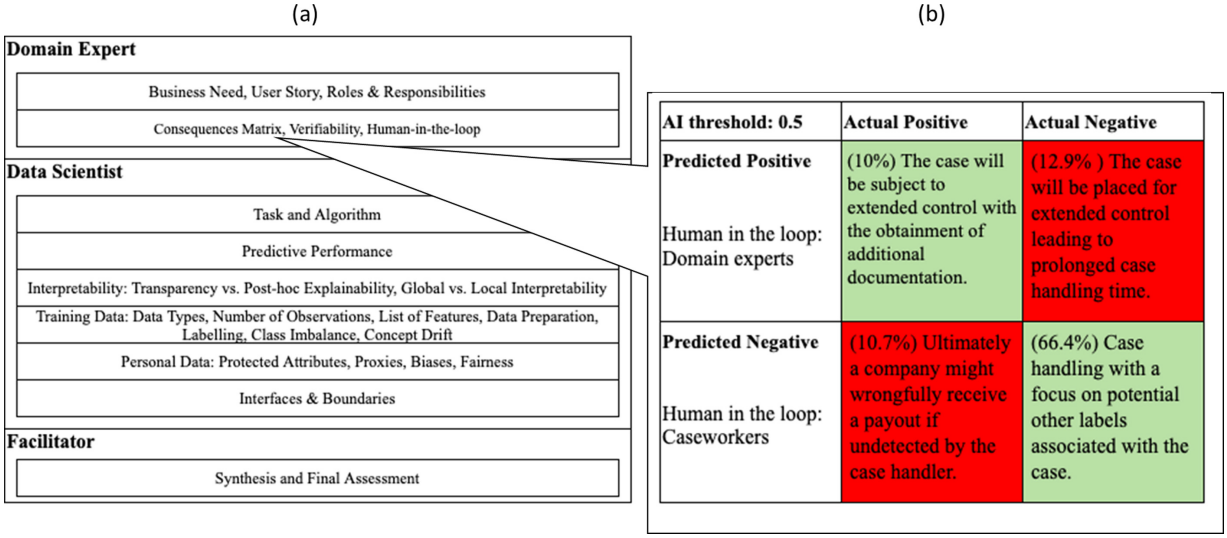


Fig. 1. (a) Schematic overview of the artificial intelligence risk assessment tool with (b) an Example of a consequence matrix

The third module comprises a synthesis and final assessment of the business and technical perspectives. This qualitative assessment, which should be conducted collaboratively

by domain experts, data scientists, and a facilitator, replaces the original idea of a quantitative risk score (like in the Canadian AIA tool). Exemplary questions include "Does the model solve the business need?", "Is the model interpretable enough?", or "Is the model free from discriminating biases?". The AI system cannot be put into production before every question in this section is answered with a yes.

5 Reflection, Learning, and Formalization of Design Principles

Going beyond the concrete and situated IT artifact described in Sect. 4, we also derived more general theoretical statements from our ADR project and formalized them in the form of design principles (see Table 2). These prescriptive statements should enable others to build instances of the here presented class of IT artifacts (i.e., AI Risk Assessment tools). According to the idea of ADR, these design principles constitute the main scientific contribution of our work. We describe design principles using a recently proposed schema [1][36].

The first three design principles are grounded in risk management theory and focus on eliciting input and feedback from a diverse group of motivated stakeholders. More specifically, the risk assessment should involve both ML designers and users in the assessment process (DP #1). Support for this principle comes both from the risk management literature [9, 10] and from the issues encountered in the second integration when we used one document that did not cater for the needs of specific stakeholders. We also made the experience that it can be difficult to involve experts in the risk assessment, which they may perceive as a formality with little business value [8]. To not burden experts with too many forms and rules and allow for advances in technology and domain-specific approaches, we decided not to prescribe precisely which methods and metrics to use during the assessment but instead to rely on their expertise in choosing the right tools (DP #2). The predictions made by the AI systems deployed at the DBA can have critical real-world consequences for businesses and citizens. Hence, in line with the focus on both probability and impact in risk management [7], it is not sufficient to evaluate their performance purely in terms of statistical measures (e.g., accuracy, precision, or

Table 2. Design principles for an artificial intelligence risk assessment tool

Principle of…	Aim, implementer, and user	Mechanism	Rationale
1: Multi-perspective expert assessment	To perform a multi-perspective risk assessment (aim), organizations using AI should…	… ensure that the AI system is jointly assessed by users (domain experts) and developers (data scientists)	Risk assessment in socio-technical systems implies integrating knowledge from business and technical perspectives [9, 10]
2: Structured intuition	To motivate and engage diverse stakeholders to participate in risk assessment (aim), organizations using AI (implementers) should…	… prescribe aspects that need to be assessed, but not the specific methods or tools to be used for that assessment	Risk assessment needs to strike a balance between deliberate analysis and structure to ensure motivation and coverage of key risks [8]

[1] As the Context element did not vary between our design principles ("In organization with values similar to the European Union where AI is used to aid or make decisions.") we decided to omit it from the table. We also omitted the optional Decomposition element.

3: Expected consequences	To make risk assessments based on expected real-world consequences instead of lab results (aim), organizations using AI (implementers) should…	… combine probabilities of outcomes of algorithmic decisions (e.g., true positive/negative rate) with their respective costs and benefits	Considering both risk probabilities and their impacts is a common practice in risk management [7, 8]. Drawing on expected utility theory [33], we extend this idea to also take positive outcomes into consideration
4: Beyond accuracy	To account for risks beyond "false predictions" (aim), organizations using AI (implementers) should…	… evaluate AI systems not only in terms of predictive accuracy but also in terms of dimensions like interpretability, privacy, or fairness	We draw on Lipton's [15] desiderata of interpretable ML (trust, causality, transferability, informativeness, and fair and ethical decision making) and the accompanying properties of interpretable models in terms of transparency and post-hoc explainability. The principle is further backed up by the EU GDPR
5: Envelopment of black boxes	To leverage the superior predictive power of complex "black box" AI systems with minimal risks, organizations using AI (implementers should…	… envelop the training data, inputs, functions, outputs, and boundaries of their AI systems	In robotics, envelopes are three-dimensional cages built around industrial robots to make them achieve their purpose without harming human workers or destroying physical things [25]. The idea has recently been transferred to ML by Robbins [20] and Asatiani et al. [19]

recall). Instead, decision-makers should assess the expected consequences in terms of the probabilities of correct and erroneous decisions and their costs and benefits in the downstream business processes (DP #3).

The last two design principles are grounded in the literature on interpretable and safe ML. In line with the previous principle, a purely technical evaluation in terms of predictive accuracy will not capture all possible risks stemming from the use of AI in governmental contexts. Algorithmic decisions must be precise and interpretable for audiences with varying levels of ML knowledge (e.g., citizens, caseworkers, lawyers, politicians) and comply with a country's legal frameworks and ethical values (DP #4).

Finally, we realized that in some situations, it might not be possible to use inherently transparent AI systems (e.g., because a deep neural network offers drastically superior predictive performance on text or image data over a simple statistical model). Adopting the idea of envelopment from the field of robotics, we propose to build virtual envelopes acting as safety nets around parts of an AI system to detect and mitigate risks (DP #5). Examples include

putting a human in the loop to check the outputs of an AI system or to monitor if the distribution of input data at production time is still compatible with the data the model was trained on.

6 Discussion

In this paper, we asked the research question: *How should procedures be designed to assess the risks associated with a new AI system?* We addressed this research question through an ADR project where we built, implemented, and evaluated the AIRA tool at a public sector organization with pioneering AI use. Our key outcomes are an artifact—the AIRA tool—and five design principles for AI risk management.

Although there is little research on the specific topic of AI risk management, the closest research is work on AI model documentation, including the Canadian AIA tool, Datasheets for datasets [13], Model cards for model reporting [14], and auditorial approaches [16, 17]. Our work goes beyond this existing research in four important ways. First, our work puts greater emphasis on guiding the communication between stakeholders
of diverse expertise, focusing on the interaction between AI systems builders and users. This emphasis manifests in questionnaires for three distinct user groups (domain expert, data scientist, facilitator) and in design principle #1. Second, the AIRA tool goes beyond existing approaches by its greater focus on establishing a joint understanding of the consequences of AI use among involved stakeholders, helping the participants to assess risks relative to the benefits of the AI system. This manifests in design principle #3. Third, the AIRA tool emphasizes incorporating model performance metrics beyond accuracy, including assessments of bias, fairness, and interpretability. This balanced assessment is important because the interpretability of AI is essential for preproduction risk identification and for postproduction risk monitoring. Fourth, we contribute to a stronger theoretical grounding of literature on AI documentation and assessment by discussing how the broader risk management literature and envelopment theory can inform AI documentation and assessment efforts.

Our research is not without limitations. First, the artifact has not been subject to summative evaluation. It was not possible to compare the undesired outcomes when using the AIRA tool to undesired outcomes when not using the tool. Second, the AIRA tool might not transfer without adjustments to other countries and the private sector. Third, the AIRA tool is a proactive measure, helping ensure that compliance requirements are
met when implementing a new AI system; but it does not address the changing nature of society, including AI systems impact on own environment. A false sense of security can occur if the AIRA tool is applied with a once-and-for-all mindset due to e.g., data drift issues that can impact the model performance and responsibility when running in production. Given that the focus of the AIRA tool is on risk assessment and not on risk
response planning, the AIRA tool would need to be complemented by proactive measures such as an evaluation plan before production and reactive measures in production such as evaluation and retraining [32].

References

1. Benbya, H., Davenport, T., Pachidi, S.: Special issue editorial: artificial intelligence in organizations: current state and future opportunities. MIS Q. Executive 19, ix–xxi (2020)
2. Mayer, A.-S., Strich, F., Fiedler, M.: Unintended consequences of introducing ai systems for decision making. MIS Q. Executive 19, 239–257 (2020)

3. Hill, K.:Wrongfully Accused by an Algorithm. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html (2020)
4. Angwin, J., Larson, J.,Mattu, S., Kirchner, L.:Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing?token=1B8jKuq-H9G4ZEq4_95FZ7ZaZ9a3rKDs. Accessed 11 Oct 2020
5. Sweeney, L.: Discrimination in online ad delivery: google ads, black names and white names, racial discrimination, and click advertising. Queue 11, 10–29 (2013). https://doi.org/10.1145/2460276.2460278
6. Lashbrook, A.: AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind. https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/. Accessed 12 Oct 2020
7. Boehm, B.W.: Software risk management: principles and practices. IEEE Softw. 8, 32–41 (1991). https://doi.org/10.1109/52.62930
8. Moeini, M., Rivard, S.: Sublating tensions in the IT project risk management literature: a model of the relative performance of intuition and deliberate analysis for risk assessment. J. Assoc. Inf. Syst. 20 (2019). https://doi.org/10.17705/1jais.00535.
9. Barki, H., Rivard, S., Talbot, J.: An integrative contingency model of software project risk management. J. Manag. Inf. Syst. 17, 37–69 (2001)
10. Wallace, L., Keil, M., Rai, A.: Understanding software project risk: a cluster analysis. Inf. Manage. 42, 115–125 (2004). https://doi.org/10.1016/j.im.2003.12.007
11. Baskerville, R.L., Stage, J.: Controlling prototype development through risk analysis. MIS Q. 20, 481–504 (1996). https://doi.org/10.2307/249565
12. Treasury Board of Canada Secretariat: Directive on Automated Decision-Making. https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592. Accessed 17 Oct 2020
13. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H., Crawford, K.: Datasheets for Datasets. arXiv:1803.09010 [cs] (2020)
14. Mitchell, M., et al.: Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* 2019, pp. 220–229 (2019). https://doi.org/10.1145/3287560.3287596
15. Lipton, Z.C.: The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. Queue 16, 31–57 (2018). https://doi.org/10.1145/3236386.3241340
16. Buolamwini, J., Gebru, T.: Gender shades: intersectional accuracy disparities in commercial gender classification. In: Proceedings of Machine Learning Research, vol. 81:1–15, p. 15 (2018)
17. Raji, I.D., Buolamwini, J.: Actionable auditing: investigating the impact of publicly naming biased performance results of commercial AI products. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 429–435. ACM, Honolulu HI USA (2019). https://doi.org/10.1145/3306618.3314244
18. Asatiani,A., Malo, P., Nagbøl, P.R., Penttinen, E., Rinta-Kahila, T., Salovaara, A.: Challenges of explaining the behavior of black-box AI systems. MIS Q. Executive 19, 259–278 (2020)
19. Asatiani, A., Malo, P., Nagbøl, P.R., Penttinen, E., Rinta-Kahila, T., Salovaara, A.: Sociotechnical envelopment of artificial intelligence: an approach to organizational deployment of inscrutable artificial intelligence systems. J. Assoc. Inf. Syst. 22, 325–352 (2021). https://doi.org/10.17705/1jais.00664
20. Robbins, S.: AI and the path to envelopment: knowledge as a first step towards the responsible regulation and use of AI-powered machines. AI Soc. 35(2), 391–400 (2019). https://doi.org/10.1007/s00146-019-00891-1
21. Baird, A.,Maruping, L.M.: The next generation of research on IS use: a theoretical framework of delegation to and from agentic IS artifacts.Manage. Inf. Syst.Q. 45, 315–341 (2021). https://doi.org/10.25300/MISQ/2021/15882
22. Badenhorst,K., Eloff, J.:Computer security methodology: risk analysis and project definition. Comput. Secur. 9, 339–346 (1990)
23. Sein, M., Henfridsson, O., Purao, S., Rossi, M., Lindgren, R.: Action design research.Manag.Inf. Syst. Q. 35, 37–56 (2011)

24. European Commission: Communication from the commission to the european parliament, the European council, the council, the European economic and social committee and the committee of the regionS Artificial Intelligence for Europe, Brussels (2018)
25. Floridi, L.: Children of the fourth revolution. Philos. Technol. 24, 227–232 (2011). https://doi.org/10.1007/s13347-011-0042-7
26. Floridi, L.: Enveloping the world: the constraining success of smart technologies. In: CEPE. 2011: Crossing Boundaries Ethics in Interdisciplinary and Intercultural Relations, p. 6. INSEIT (2011), Milwaukee Wisconsin (2011)
27. Keil, M., Li, L., Mathiassen, L., Zheng, G.: The influence of checklists and roles on software practitioner risk perception and decision-making. J. Syst. Softw. 81, 908–919 (2008). https://doi.org/10.1016/j.jss.2007.07.035
28. Bannerman, P.L.: Risk and risk management in software projects: a reassessment. J. Syst. Softw. 81, 2118–2133 (2008). https://doi.org/10.1016/j.jss.2008.03.059
29. United Nations: United Nations E-Government Survey 2018. United Nations (2018)
30. United Nations: Department of Economic and Social Affairs: United Nations e-government survey 2020: digital government in the decade of action for sustainable development. United Nations, Department of Economic and Social Affairs, New York (2020)
31. The Danish Government: National Strategy for Artificial Intelligence. Ministry of Finance and Ministry of Industry, Business and Financial Affairs (2019)
32. Nagbøl, P.R., Müller, O.: X-RAI: a framework for the transparent, responsible, and accurate use of machine learning in the public sector. In: Proceedings of Ongoing Research, Practitioners, Workshops, Posters, and Projects of the International Conference EGOV-CeDEM-ePart 2020, p. 9 (2020)
33. Morgenstern, O., Von Neumann, J.: Theory of Games and Economic Behavior. Princeton University Press (1944)
34. Briggs, R.: Normative Theories of Rational Choice: Expected Utility. https://plato.stanford.edu/entries/rationality-normative-utility/ (2014)
35. Suresh, H., Guttag, J.V.: A Framework for Understanding Unintended Consequences of Machine Learning. arXiv:1901.10002 [cs, stat] (2020)
36. Gregor, S., Kruse, L.C., Seidel, S.: Research perspectives: the anatomy of a design principle. J. Assoc. Inf. Syst. 21,1622–1652 (2020). https://doi.org/10.17705/1jais.00649

Challenges and Practices in the Evaluation of AI Systems in the Public Sector

Per Rådberg Nagbøl (IT University of Copenhagen), Oliver Krancher (IT University of Copenhagen), Oliver Müller (Paderborn University). 2022.

Submitted for review to the forthcoming Research Handbook on Public Management and Artificial Intelligence.

Abstract

While research on the development and adoption of AI systems is growing, organizations can harness benefits and avoid harm from AI systems only if AI systems maintain high performance after they are developed and adopted. A key activity in this regard is the evaluation of productive AI system. In this Action Design Research Study, we built, implemented, and evaluated an infrastructure for evaluating productive AI systems at the Danish Business Authority, and we examined the challenges that such an infrastructure needs to address. We found that key challenge revolve around tedious work, resource availability, maintaining an overview, ensuring sufficient priority, and timing of evaluations. We propose that these challenges can be addressed by a digitized evaluation infrastructure that automatically stops systems not evaluated and supports managers and evaluators to choose strategies for the timing of evaluations, for making evaluation work meaningful, and for leveraging synergies between evaluation and other activities.

Keywords: Artificial Intelligence, Evaluation, Government

Introduction

Governmental organizations and business alike are making increasing use of Artificial Intelligence (AI) systems to automate and support various tasks across different domains (Berente et al., 2021; Sun and Medaglia, 2019). While empirical research to date has focused on the development, adoption, and implementation of AI systems (Asatiani et al., 2021; Sun and Medaglia, 2019; van den Broek et al., 2021), less attention has been paid to the maintenance phase, i.e., the part of an AI system's lifecycle that starts after the system has been implemented in an organization and ends with its decommissioning. Given the high costs associated with building AI systems, the maintenance phase is critical because the longer an AI system can be productively used, the more likely it is that the initial cost will be recovered. Moreover, a focus on the maintenance phase is important given that productive AI systems (i.e., AI systems in the maintenance phase) may cause harm, such as by making decisions that discriminate against particular social groups (Hill, 2020; Mayer et al., 2020), and that preventing such harm is important throughout the entire lifecycle of an AI system.

A key activity during the maintenance phase is the *evaluation* of AI systems. Evaluation has been defined as the cybernetic process of assessing the performance of a system in relation to performance expectations (Doshi-Velez and Kim, 2017; Eisenhardt, 1985; Kirsch, 2004). In the context of AI systems, evaluation involves thus an assessment of the performance characteristics of an AI system such as its accuracy, fairness, and transparency (Lipton, 2018; Russell and Norvig, 2002) in relation to stakeholders' performance expectations. Evaluation during the maintenance of an AI system is not only an opportunity to discover performance issues not found during development. It is also critical to prevent decrease in performance (e.g., in accuracy or fairness) over time. Performance may decrease due to changes in the environment that lead production data to drift away from the AI system's training data. For example, the performance of an AI systems that is trained to recognize signatures may

decrease if the technologies through which citizens sign applications changes. If such environmental changes are not detected, the organization may be unaware of running a productive AI system that makes poor decisions. Performance may also decrease because of changes in behaviors, standards, and laws, which might cause the AI systems to enforce an old and incorrect version of the law.

Despite the recent surge of interest in AI systems, relatively little research has focused on evaluating the performance of AI systems during maintenance. Socio-technical AI research has focused on issues such as top management involvement (Li et al., 2021), collective learning (Fügener et al., 2021; van den Broek et al., 2021), delegation and augmentation (Baird and Maruping, 2021; Teodorescu et al., 2021), pre-production risk assessment and mitigation (Asatiani et al., 2021, 2020; Nagbøl et al., 2021), and unexpected outcomes (Mayer et al., 2020; Strich et al., 2021) without explicit attention to evaluation during maintenance. Technical research has explored strategies for evaluating AI systems (Doshi-Velez and Kim, 2017; Hernández-Orallo, 2017), though without focusing on the issues that arise when organizations attempt to implement these strategies in organizational realities throughout the lifecycle of a system.

Although existing work does not explicitly focus on the evaluation of AI systems during maintenance, it offers a key insight that provides important background knowledge for the design of evaluation systems, namely the insight that effective use of AI requires integrating domain and AI knowledge. For example, a study on the design of pre-production risk assessment emphasizes the importance of a multi-perspective expert assessment, involving both AI experts and domain specialists, in an approach that goes beyond accuracy metrics and relies on the stakeholders' diverse experience and expertise for assessing AI systems (Nagbøl et al., 2021). An ethnographic study describes the interplay of machine learning (ML) expertise and domain expertise in human-ML hybrid practice in the domain of hiring. It

finds that developers and domain experts are in an interdependent relationship where domain experts contribute to defining, evaluating, and complementing machine input and output while developers contribute novel ML-based insights from the data (van den Broek et al., 2021). Based on archival data on drug development, Lou and Wu (2021) make a similar claim that the development and use of AI systems requires integrating the knowledge of AI and medical experts. Lebovitz et al. warn against treating the ground truth as objective when based on uncertain knowledge, pointing to a tension between how domain experts evaluate their work according to know-how and how AI systems are evaluated accordingly to quality measures of know-what and ground truth measures. They recommend that humans should make the final judgment in areas of high uncertainty while AI systems in fields with more established knowledge claims should be trained and validated accordingly to quality measures representing the know-how and standard of expert's practical performance (Lebovitz et al., 2021). Doshi-Velez and Kim propose a three-level taxonomy of interpretability evaluation (applications-grounded evaluation, human-grounded metrics, functionally-grounded evaluation), highlighting that evaluation strategies may differ in the way in which they involve human domain expertise (Doshi-Velez and Kim, 2017).

While these studies provide important background knowledge, we know little about how organizations can ensure the effective ongoing evaluation of their AI systems in production. Therefore, against this backdrop we ask the research questions: *How can organizations ensure effective evaluation of productive AI systems?* With the two sub-questions: (1) What are challenges in planning and enforcing the evaluation of productive AI systems? (2) How can these challenges be addressed?

We have addressed these questions through an Action Design Research (ADR) study in the Danish Business Authority (DBA). ADR provides a good fit for the research project because it allows studying the planning and execution of evaluation under authentic circumstances.

The DBA provides an excellent setting by being a frontrunning organization[1] in a world-leading country in e-government (Nations, 2018; United Nations. and Department of Economic and Social Affairs, 2020), providing rare opportunities for exploring issues of evaluating AI systems in production. In the remainder of this paper, we present our ADR methods, report our findings about challenges and solution strategies in AI evaluation, and discuss these findings.

Methods

The projects methodological approach is Action design research (ADR) which creates generalizable knowledge through solving practical problems by combining action and design research (Sein et al., 2011). Key outcomes of ADR are one or more artefacts and design principles. In our case, the artefact is a method for evaluating productive AI systems, which we call Evaluation Plan (see the section on the Design Artefact below for more details). ADR proceeds along the four stages of problem formulation, building intervention and evaluation (BIE), reflection and learning, and formalization of learning.

The first stage, *problem formulation*, is initiated through the engagement with a practical problem and scoping the project (Sein et al., 2011). The stage is based on two principles. *Principle 1: Practice-Inspired Research* turns a non-unique practical problem into a knowledge creation opportunity by treating the problem as an instance of a class of problems. Through our existing collaboration with the DBA on issues of AI management, we identified the evaluation of AI systems as a key challenge in public-sector organizations relying on AI system, suggesting that artifacts and design principles developed through the research project could be of value to organizations other than the DBA (Sein et al., 2011). *Principle 2:*

[1] The DBA was nomination for Danish digitization price sammenhængsprisen for the public sector for their AI supported work with the covid-19 compensation https://offdig.dit.dk/da/Om-OffDig/Digitaliseringsprisen

Theory-ingrained artifact emphasizes that the artifact should not be purely based on the designers' creativity or practical requirements but also grounded in literature and theory (Sein et al., 2011). In line with the principle of theory-ingrained artefact, we integrated our emerging findings on challenges and solution strategies with theories that can explain and inform the challenge or the solution strategies and thus inform the artefact.

The second stage, *building, intervention, and evaluation* (BIE), describes an iterative process of building the artifact, intervening in the organization, and continuous evaluating both the problem and artifact, ultimately leading to the realized design of the artifact. It relies on three principles: Principle 3: Reciprocal Shaping, Principle 4: Mutually Influential Roles, and Principle 5: Authentic and Concurrent Evaluation. *Reciprocal shaping* focuses on the mutual influence that the two domains in form of the IT artifact and organizational context have on each other. The principle of *mutually influential roles* emphasizes the necessity of mutual learning among the participants in the design project where different actors provide different perspectives into the project. In line with this principle, data scientists, domain experts, and managers from the DBA contributes important insights into their requirements, methods, and challenges while the researchers contributed knowledge about the literature and theories on AI systems and on theories that shed light on the emerging findings. *Authentic and Concurrent Evaluation* represents the key idea the evaluation of the artefact (i.e., the evaluation of the Evaluation Plan) is not a stage in a process but an ongoing endeavor (Sein et al., 2011). Consistent with this principle, the decisions about designing, shaping, and reshaping the Evaluation Plan and implementing it into organizational work practices were accompanied by an ongoing evaluation.

The third stage, *reflection and learning,* runs in parallel to Stages 1 and 2 but focuses on the insights that result from the development of the artefact through reflections about the problem scope, the ingrained theories, and the emerging ensemble artifact and its evaluation. It relies

on *Principle 6: Guided Emergence*, which recognizes that the learnings are not only the product of the researcher but also of its organizational use, the participants' perspectives, authentic outcomes, and concurrent evaluation (Sein et al., 2011). In line with this principles, reflection and learning occurred through an ongoing dialog between the researcher and participants at the DBA, the work on and use of the Evaluation Plan, and its evaluation.

The fourth stage, *formalization of learning*, involves a conceptual move from one instance of a problem to a general solution applicable for a whole class of problems to satisfy *Principle 7: Generalized Outcomes* (Sein et al., 2011). Following this principle, we moved from our instance of the problem—the use of the Evaluation Plan at the DBA—to design principles that can help inform the evaluation of AI systems in organizations more generally.

Empirical work

The first author of this article has been working with the DBA since September 2017, spending about half of his time in the Machine Learning Lab at the DBA, taking part in everyday work-life activities. He kept a field diary with notes from observations in the organization and participation in meetings and conversations with colleagues and consultants, which was supplemented with insights from reading and writing emails and documentation on different platforms such as (e.g., Git, Teams, Jira, and Confluence).

Design Artifact

The design artifact, the Evaluation Plan, is part of a broader framework for responsible AI use X-RAI (Nagbøl and Müller, 2020). Together with the Artificial Intelligence Risk Assessment (AIRA) tool (Nagbøl et al., 2021), is intended for proactive pre-market use, creating the foundation for post-market evaluation and retraining of AI systems. The Evaluation Plan is in its supplementary nature inheriting the theory ingrained into AIRA, including principles such as multi-expert assessment and structured intuition (i.e., providing

some structure while leaving experts room for their judgment) (Nagbøl et al., 2021). In line with the principle of structured intuition, the Evaluation Plan is a questionnaire that provides some structure while leaving room for expert judgment. Table 1 shows the Evaluation Plan, as it was implemented at the DBA during iteration 3 (see below for a description of iterations).

Table 1: Evaluation Plan Artifact

Question No.	Question
Q1	Who should participate in the evaluation (e.g., application manager, relevant business unit, ML lab)?
Q2	Who owns the model/the solution (usually the business)?
Q3	When should the first evaluation meeting take place?
Q4	What is the expected meeting frequency (How often should you meet and evaluate)?
Q5	What is the current threshold setting for the AI system?
Q6	What is the basis for the evaluation (e.g., logging data, annotated evaluation data, i.e.., data where human categorization is compared with the model)?
Q7	Is data unbalanced to a degree where this must be taken into account when fabricating data for evaluation and retraining. If so, how?
Q8	What resources are needed (e.g., who can make evaluation data, evaluation data is provided internally or externally, how much needs to be evaluated, what is the cost in time / money)?
Q9	What is the expected resource need for the evaluation?
Q10	Is the model visible or invisible to external users?
Q11	Does the model receive input from other models? If so, which ones?
Q12	What are success and error criteria (eg When does a model perform good / bad, what percentage, business value, labor waste)?
Q13	Is there future legislation that will have an impact on the model's performance (e.g., introduction of new requirements, abolition of requirements or the like)?
Q14	Are there other future factors that affect the model's performance (e.g., bias, circumstances, data, standards or the like)?
Q15	When should the model be retrained?
Q16	When should the model be muted or deactivated?

BIE Iterations

The initial work with designing the artifact (the Evaluation Plan) started in February 2019 in close collaboration with stakeholders from the company registration (business unit), a product owner, and the Machine Learning lab. The Evaluation Plan was designed to accompany the Evaluation Framework and the Retraining Framework in a three-framework process. The process was expanded with a fourth framework for Artificial Intelligence Risk Assessment (AIRA) (Nagbøl et al., 2021) inspired by the Canadian Algorithmic Impact Assessment tool (Secretariat, Treasury Board of Canada, 2020) and further developed into the X-RAI (Nagbøl and Müller, 2020) method. The intervention occurred by using the Evaluation Plan on 16 different AI systems in the DBA. In three iterations, the artifact was evaluated with the three different foci: usability and content (iteration 1), behavioral impact (iteration 2), and challenges (iteration 3).

Iteration 1: Usability and Content

The Evaluation Plan was evaluated accordingly to ADR principles of authentic and concurrent evaluation. It was introduced into the organizational work practices in a word format. The evaluation focused on the user needs, usability, and content of the Evaluation Plan. The Evaluation Plan was ongoingly modified accordingly to the findings from the evaluation. The evaluation focused on the understandability of the questions and on its suitability for estimating the resource needed. The evaluations led to minor changes to the artifact until the reach of a satisfactory maturity level, where the artifact was redesigned into a YAML format for integration into an IT infrastructure.

Iteration 2: Behavioral Impact

The second evaluation focused on evaluating to which extent the Evaluation Plan fulfilled its expected behavioral impact, i.e., the impact of securing and structuring the evaluation of AI systems in the DBA. To this end, we gathered the compiled Evaluation Plans and other

relevant documentation such as Evaluation Schemas from the third framework. The compiled Evaluation Plan frameworks were then analyzed. The analysis revealed that 16 AI systems had compiled an Evaluation Plan stating the time for the first evaluation and the expected following evaluation frequency. There were, to our awareness, only three AI systems with filled-out evaluation frameworks, one of which filled out the evaluation framework only partially. The two full evaluations of the AI systems had taken place before Covid-19. The lockdown of society and the working from home situation caused by Covid-19 increased the difficulties in maintaining an overview of the status of the different AI systems. Therefore, we decided to conduct formal interviews to validate our findings from previous evaluations, discover overlooked practices, and gain a deeper understanding of causes and reasons.

Iteration 3: Challenges

The third evaluation focus was on discovering overlooked evaluation practices and gaining a deeper insight into the circumstances impacting evaluation. We decided to do seven semi-structured interviews with stakeholders named in the Evaluation Plans. The stakeholders held diverse positions related to IT development, ML, and different departments using AI systems. Interview durations varied from 44 to 80 minutes. The interviews were structured around the following themes: introduction questions and background, AI systems purpose and use, quality assurance, evaluation, accountability, risk, challenges, and trust. The interviews were transcribed and coded in Nvivo. In coding, we followed an inductive process where lower-level challenges and design principles were aggregated to a few higher-order categories, similar to data analyses approaches in case study research and grounded theory research (Charmaz, 2006; Yin, 2009). The design principles are planned to be implemented in a subsequent, digitized version of the Evaluation Plan.

Findings: challenges and solutions

Figure 1 provides an overview of our findings. Our data analysis led us to identify the five challenges shown on the left-hand side of Figure 1. These challenges can be addressed by an Evaluation Plan infrastructure that is based on five design principles shown on the right-hand side of Figure 1. The arrows show which design principles help address which challenges.

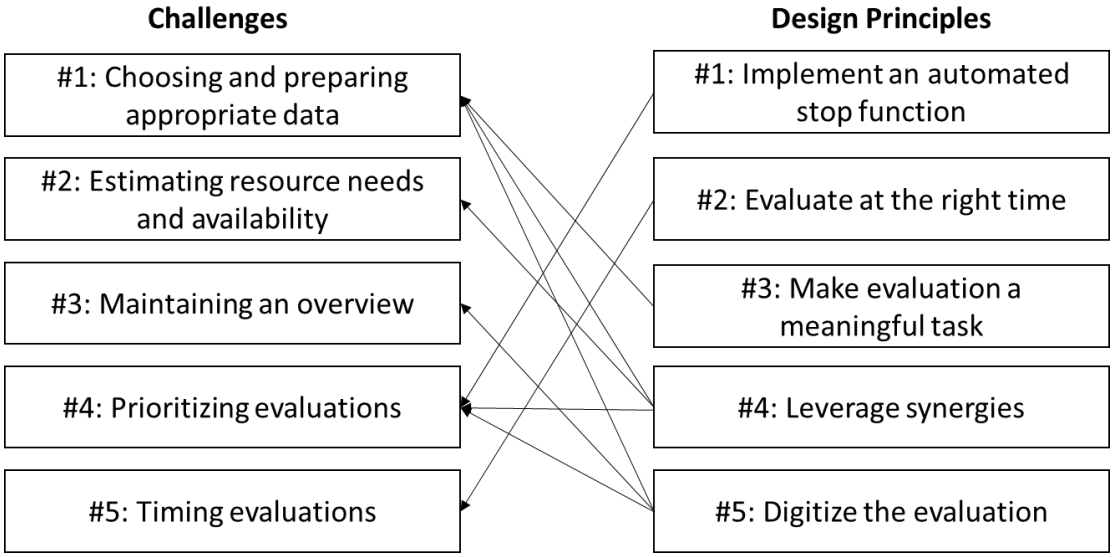


Figure 1: Challenges of and Design Principles for AI Systems Evaluation

Challenges

#1: Choosing and preparing appropriate data

Several of our informants mentioned challenges in choosing and preparing the data that is needed to evaluate productive AI system. These challenges revolved around tedious annotation work and bias in the available data.

Our participants perceived annotating data for supervised ML as a rather tedious, resource-intensive task. Much like during the initial training of an AI system, postproduction evaluation involves selecting data and providing ground truth about the data. One domain specialist highlighted that annotation activities during evaluation can be very time-

consuming, requiring domain specialists to "read … these 10,000-15,000 lines" and "put a number here and a cross there" (Daniel). This work was sometimes perceived as tedious and frustrating especially when looking for very small minority classes: "They wanted me to … look through 100 lines once a month. But almost none of the lines was related to [a specific type of record]. It might be valuable for the model, … [but] there was no [record] that fit exactly what we needed. Then you're lost." (Daniel)

Other participants mentioned difficulties related to choosing and preparing data that were related to bias in the available data. For instance, one participant mentioned a potential source of bias in data obtained from another public-sector organizations. They said employees of the other organization would likely focus their checks on those companies that were most likely to submit incorrect reports. Using these data as a basis for annotation during evaluation could result in "a massive bias" (Rikke), given that the data from companies that were likely to submit correct reports was underrepresented in the data.

#2: Estimating resource needs and availability

Our informants reported that it was often difficult to anticipate the resources available and needed for evaluation because the use of AI fundamentally changed business processes and priorities. For instance, Torben mentioned that the work in his team increased, rather than decreased, after AI was introduced. AI provided the opportunity of more effectively detecting fraudulent applications early in the process, helping employees "to take out the right ones" early in the process rather than to "waste a lot of businesses' time" nor manual checks later. Because of this more effective, AI-enabled checking process, the DBA found it economic to increase the size of the team from 4 to 18 employees, who were now responsible for following up on the alarms triggered by the system. These and other fundamental changes in the business process and the role of human labor in the business process would have made it very difficult, if not impossible, for the DBA to anticipate what amount of resources would

be available for evaluation and not bound by the new work activities that are enabled by the AI system.

#3: Maintaining an overview

The analysis of the filled-out Evaluation Plan revealed that it was increasingly difficult to maintain an overview of the status of AI systems evaluations. As the number of productive AI systems increased and as these systems evolved, the Evaluation Plans increasingly became historical documents displaying an intention to evaluate rather than a tool for monitoring the evaluation. In this situation, maintaining an overview was difficult for several reasons. First, as the use AI increased, so grew the number of AI systems, of data scientists, and of business processes supported by AI. Second, the Covid-19 crises required the DBA to direct managerial attention to urgent issues such as systems supporting the allocation of compensation packages for companies suffering from the pandemic. This drew attention away from the evaluation of AI systems. Thirdly, AI systems were not only added but also paused or retired, making it more difficult to maintain an overview. Some AI systems were periodically switched on and off: "They are not retired, just temporarily switched off … the intention is that it is periodically switched on but not permanently … it is one of the things we will have periodically switched on, for example, from April to June…" (Torben). Fourth, staff changed as described by an informant while looking at the Evaluation Plan "Liselotte on Y has left, and Harald has left ML Lab, and I named on X, and Maria is to my knowledge still here, but I have not seen her for a long time, but I believe she is still employed…." (Kim).

#4: Prioritizing Evaluations

The Evaluation Plans were initially followed until the start of the Covid-19 pandemic. The pandemic caused an exceptional situation at the DBA where enormous amounts of resources were needed to rapidly develop systems such as systems for administering the compensation

packages that the Danish Government granted to businesses suffering from lockdowns. This exceptional situation made it difficult for the DBA to allocate resources to the evaluation of existing AI systems.

It was not only the Covid-19 pandemic that bound resources; so did the development of a new digital platform that was introduced to make evaluations more effective and efficient in the future: “There has not been time… it has been flagged, but it has not been prioritized there has been put more will towards things that had to be built…we have been living with compensation (covid-19) for almost two years and besides there has been a new platform (Intelligent Control Platform) that had to be built… we were about to find a routine if you go to years back for how everything should be evaluated….” (Theo). Developing this required refactoring all the existing AI systems.

While the pandemic and the development of the new digital platform were one-off events that bound resources, our informants also described the challenges of mobilizing sufficient resources for evaluation in organizational realities. For instance, one informant said: ““…the challenge is that if the business unit if we start to be pressured on the resources on task and on time and when one can see that an evaluation of a model is going to take around 20 hours and these are hard to find then we end up not doing it...” (Torben)

#5: Timing evaluations

There was substantial uncertainty about when evaluations should best be conducted. At the beginning, a rule of thumb was that the first evaluation should take place 14 days after go-live and that subsequent evaluations should be performed every third month. Our informants agreed that deciding on the time and scope for the first evaluation was difficult. For example, are child diseases and early implementation issues something to include, or should the first evaluation only touch on matters aligned with the subsequent evaluations? Some informants argued that 14 days too early for some AI systems: “I think it is a little optimistic … there

might still be some issues and minor mistakes that must be corrected right when the model is put in production … I also think that the business unit would need some time to look at the cases…" (Rikke). Another informant pointed to the scarcity of available data when systems are evaluated too early: "We often first know our models' effect when the caseworkers have worked the cases flagged by the model…" (Theo).

Not only the decisions about timing of the first evaluation were difficult to make; so were decisions about the timing of subsequent evaluations because "the models will automatically perform worse over time" (Oscar), making it required to time ongoing evaluations before performance decreases substantially. As one informant put it: "You do not know when the fraud patterns are changing ... it can change the day after the evaluation." (Ida).

Design principles

Informed by the challenges described in the previous section, our engagement in the DBA and our analysis of interview data suggests that the challenges can be addressed by an Evaluation Plan and underlying infrastructure that are based on the design principles shown in Table 2. We describe these design principles accordingly to the schematic guidelines suggested by Gregor et al. (2020).

Table 2: Design Principles

Principle	Aim	Mechanism	Rationale
#1: Implement an automated stop function	To enforce compliance with the Evaluation Plan…	…ensure that the AI system cannot run in production without being evaluated by humans as per the Evaluation Plan.	As (semi-)autonomous systems, AI systems can cause undesired consequences. Emergency stop measures as known from other dangerous machines like power saws or lawn mowers can help to prevent some of these consequences.
#2: Evaluate at the right time	To make sure that the AI system is up	…consider event-based and frequency-based	According to representation theory (Recker et al., 2019), the basic purpose of any information

	to date when needed…	timing strategies in line with expected real-world changes.	system, including AI-based systems. is to faithfully represent certain real-world phenomena. Hence, AI systems need to be re-evaluated and, if needed, re-trained whenever the real-world phenomenon they are representing changes.
#3: Make evaluation a meaningful task	To ensure motivated evaluators…	…design the annotation task so that it is an opportunity for autonomy, competence, and relatedness.	According to self-determination theory (Ryan and Deci, 2000), satisfying the basic psychological needs for autonomy, competence, and relatedness can increase people's intrinsic motivation for a given task.
#4: Leverage synergies between AI system evaluation, human training, human work, and AI system training	To reduce costs and make evaluation work less tedious ...	… recycle data between work, evaluation, and training activities.	According to representation theory (Recker et al., 2019), information systems are representations of real-world work systems. Hence, the task of training and assessing an AI-based decision-making system (a type of information system) has important parallels to the task of training and assessing a human decision-making system, suggesting that synergies between these two can be leveraged, e.g., by reusing the products of human training efforts for AI training or assessment.
#5: Digitize the evaluation	To ensure compliance with Evaluation plans and maintain an overview …	… implement a digital platform that automatically collects data about evaluation activities and outcomes.	According to control theory (Eisenhardt, 1985), accurate information about a controlee's behavior makes it more likely that the controlee will engage in the desired behaviors. Digitizing the evaluation infrastructure helps make information about evaluation activities transparent and thus encourages evaluators (i.e., controlees) to comply with Evaluation Plans.

#1: Implement an automated stop function

A key challenge especially during the Covid-19 pandemic was in ensuring evaluations receive sufficient priority. To address this challenge, the DBA is currently implementing an

automated emergency stop function in its Intelligent Control Platform, i.e., the infrastructure that is developed and implemented to digitize the Evaluation Plan. Automated stop function is a feature that ensures a productive AI system stops running if it is not evaluated as per the evaluation plan. Such a function is similar to an automatic train stop system, which stops a train automatically if the train conductor fails to regularly push a button. As one informant told us: "[The head of the department] is going towards setting something up in the Intelligent Control Platform so that we switch off models if it is not signed off that they are evaluated" (Ida).

#2: Evaluate at the right time

Another key challenge was to decide on the timing of evaluations. The timing the evaluation is essential for ensuring that AI systems maintain their standards and perform when needed. Through our engagement in evaluating 16 AI systems, we learned that there are multiple logics for timing the evaluation of productive AI systems and that different AI systems need different logics. For example, AI systems that build on trends, such as fraud detection systems where fraudsters change behavior over time, have other needs for evaluation and retraining than the industrial classification codes system where it will not be necessary to retrain and evaluate the system before the standards change. The idea that different AI systems require different temporal evaluation logics is consistent with representation theory (Recker et al., 2019), which holds an information systems, including AI systems, are representations of certain real-world phenomena. Whether the representation of the reality (i.e., the AI system) needs to be reassessed depends clearly on the pattern of change in the real-world phenomenon.

While the timing of the first evaluation often depended on the question of when enough data would be available, the timing of subsequent evaluations followed one of the following logics: frequency-based, event-based, seasonal, and autonomous driven. *Frequency-driven*

evaluation works from the logic evaluations must be conducted as a reoccurring event with a fixed period, for example, every third month. The idea is that the fixed period should create a natural evaluation flow. There are several considerations to make when deciding on the frequency. One important question is how long it is tolerable to run on a false premise because the pattern in behavior can change the day after the evaluation. A question to ask when planning the evaluation "for how long time can we tolerate that it answers incorrectly" (Ida) to address "…to live with not discovering that there suddenly is something that we do not catch that we think we catch right..." (Ida). The impact of the AI system, the thoroughness of prior evaluations, and the amount of dynamism were mentioned as further factors that affect the needed frequency: "How big impact it has but also how the earlier assessments have looked if we have had an evaluation rather quickly and then held one after three months and everything looks fine, and this is not an area where something is going to happen, and it is probably business as usual then there is no reason we should meet again in three months, and then we can set it up to be biannually" (Torben).

Event-driven evaluation is based on events that impact the AI system's performance or change the context of the AI systems so that the predictions are no longer suitable. Examples of such events include changes of technical standards and of industrial classification codes: "We also have XBRL with taxonomies they are changing all the time" (Daniel), "… revision of industrial classification code yes we know that would happen in 24 I think maybe 25… (Oscar).

Other AI systems have a *seasonal* flow with activity fluctuating depending on the time of the year or other recurrences. It is important to consider when planning the evaluation: " I will say that there should be more meetings the closer we get to the big filling period occurring from around the end of April until the end of June…" (Daniel)

Lastly, the DBA considered using *autonomous monitoring* of AI systems for detecting changing behavior of the AI system. Abnormalities or changes in distribution of, for an example, positive and negative classifications, can be an indicator of a need for evaluation "There is an alarm that looks at the probability returned by the model if they suddenly change a lot … If something is flagged … If … the model has this amount of true positives in this quarter, but in another quarter it had only so many true positives, why that? So again, create some rules for when it must be flagged…" (Theo).

While the properties of the real-world phenomenon represented in the AI system may affect the evaluation logics, our interviews also suggested other considerations. One important consideration is when resources can be freed up resources for evaluation. As one informant shared: "If there is an office there has five different (AI systems), then one would probably prefer having spread the evaluation work across the months" (Ida). Another important consideration is the interrelatedness of AI systems. For example, if the output of one AI system is the input to another AI system, these dependencies would need to be considered when scheduling the evaluation of the two systems.

#3: Make evaluation a meaningful task

As discussed above, one challenge was that choosing and preparing appropriate data for evaluation was often seen as a time-consuming and tedious task that, while being tedious, required highly skilled labor, such as an employee with legal or audit background. When reflecting on this challenge, our interviewees suggested several strategies for making evaluation work a meaningful task. These strategies can well explained by self-determination theory, which suggests that people will find work enjoyable if the work provides opportunities for autonomy, competence, and relatedness (Ryan and Deci, 2000). For instance, evaluation can be framed as an opportunity for competence development by emphasizing that the evaluator will obtain a first-hand feeling of how the AI system performs

on negative and positive classifications. Evaluation can be framed as an opportunity for autonomy by communicating that the evaluator plays the role of an educator of the AI system by ingraining their expert knowledge or by including the management of running AI systems into individuals' job descriptions. Evaluation can also be seen an opportunity for relatedness by involving multiple evaluators, which may provide opportunities for knowledge sharing and learning rom each other. While these strategies may help make evaluation work more enjoyable, other strategies focus on communicating the benefits and rewards from evaluation. For example, it may be helpful to communicate why the evaluation is essential, how the evaluation benefits the quality of everyday work, and what consequences may occur if the evaluation is not conducted.

#4: Leverage synergies between regular work, AI system evaluation, human training, and AI system training.

While evaluation work may appear tedious and may struggle to receive priority, our informants shared that a number of strategies help leverage synergies between evaluation and other activities, which may help reduce tedious elements of evaluation work and relax resource issues. Specifically, our informants recommended leveraging synergies between regular work, AI system evaluation, human training, and AI system training. The human oversight policy often a natural quality insurance and validation on one of the classification categories and for some AI systems both negative and positive classifications. Quality insurance is critical; hence, there have been different kinds of ongoing quality insurance and evaluation of the AI systems despite the lack of use of X-RAI's evaluation framework "… We have every week in the audit unit ongoing meetings about the model and our experiences with the model on both on caseworker level but also with our boss, and it is the thought we have these meetings among other things so that we can collect and deliver some back to ML (ML lab) when we get so far." (Kim). Our informants also pointed us to potential synergies between evaluation and human training. Indeed, the formalized and standardized evaluation

flow allows for acquiring, storing, and sharing knowledge and experience from the evaluation of the AI system, contributing to continuous individual and organizational learning and the development of best practices including utilizing the experiences that is already ingrained in the evaluation schema. The AI system supports human learning "We have just hired a new in the team who needs training, so we always switch them on (AI system) because they are some good case to get out about the formation" (Torben). The data annotation is an opportunity to work dedicated with one specific interpretation of, for example, a law repeatedly, thus stipulating learning. It is then relevant when working through the cases to annotate the data. Another synergy is to store and declare annotated evaluation data so that it can be recycled as training data when retraining the AI system. Hence, a retraining procedure starts with deciding which evaluation data to recycle. Integration of evaluation into the regular workflow is an option "… the best in the world would be that our case management is constructed in a way if I, for an example, had processed a case there was selected by the signature model then I could while closing my handling of the case do some evaluations of the positives" (Torben).

#5: Digitize the evaluation

Among the most important challenges related to evaluation were the difficulties of maintaining an overview of evaluation activities and difficulties of prioritizing evaluations. The DBA reacted to these difficulties by introducing the Intelligent Control Platform (ICP), a digitized infrastructure for managing AI systems evaluation. As control theory suggests (Eisenhardt, 1985), controlees (e.g., evaluators) are more likely to show the expected behaviors (e.g., evaluating AI systems when needed) if the information about the controlees' behaviors is transparent. Hence, a digitized evaluation infrastructure can be an important element for not only maintaining an overview of evaluation activities as the number of productive AI systems is increasing but also for enforcing that evaluations are conducted as

required. Interestingly, the platform also helps make evaluation and retraining become a more straightforward task because it allows the evaluator to annotate relevant data in a system that automatically stores it and makes it accessible for retraining purposes, which further helps cope with resource bottlenecks. As Daniel put it: “We also wanted to have a tool … it is because we should have created what we call a GUI (Graphical User Interface) where we got a model to be capable, where we better could annotate if it is fictitious (AI system prediction) then the other part goes in and says this correct good enough to start a case… “ (Daniel 18:48). The DBA has started building an infrastructure for evaluation and retraining AI systems. That infrastructure will allow for easier evaluation and faster adaption and deployment of AI systems to changes in their environment. The expectation is that the new platform will allow for better monitoring, evaluation, and retraining: ” … it makes it easier for us to evaluate the possibility because we are sitting with it closely now, it becomes easier also when discovered that the model starts to perform worse and then update the model thereby easier to put a new model in production. It is easier to retrain the models because everything is in one place in our repository” (Oscar).

Discussion

This chapter was motivated by the observation that little work has examined the evaluation of productive AI systems in organizational realities even though evaluation of productive AI systems is critical for avoiding harm and ensuring benefits from AI systems. Against this background, we asked: *How can organizations ensure effective evaluation of productive AI systems?* We focused on two sub-questions: *(1) What are challenges in the ongoing evaluation of AI systems, and (2) How can these challenges be addressed?* We have methodologically relied on action design research to answer our research questions. We have built, implemented, and evaluated our design artifact, the Evaluation Plan in the DBA, including conducting seven semi-structured interviews. As a result, we have identified five

key challenges: Choosing and preparing appropriate data, Estimating resource needs and availability, Maintaining an overview, Prioritizing evaluations, and Timing evaluations. Our engagement with the DBA and with the data led us to suggest five design principles that help address these challenges: Implement an automated stop function, evaluate at the right time, make evaluation a meaningful task, leverage synergies between regular work, AI system evaluation, human training, and AI system training, and digitize the evaluation.

Although there is little research on the topic of evaluation of productive AI systems, our study shows that there is foundational research such as representation theory (Recker et al., 2019) and control theory (Eisenhardt, 1985) and provides important guidance for designing evaluation infrastructure. Moreover, our findings also relate to research on collaboration and combining knowledge and insights from domain experts and AI experts in earlier stages of AI development (Lebovitz et al., 2021; Lou and Wu, 2021; Nagbøl et al., 2021; van den Broek et al., 2021). We have found a similar symbiotic relationship between domain experts in company law and audit and AI experts in the context of Government. We extend their claim by arguing that the collaboration must continue after the AI systems go live. We have designed a tool to support and structure such collaboration after the AI system go-life described challenges and suggested solutions. Our study is also related to the three-level Taxonomy of Interpretability Evaluation with Applications-grounded Evaluation, Human-grounded Metrics and functionally-grounded Evaluation (Doshi-Velez and Kim, 2017). Challenges related to conducting evaluation in the DBA with an approach similar to the Applications-grounded Evaluation where the real domain experts evaluate the ML model by doing the real tasks and solutions to those challenges.

Limitations and future research

It is important to point out that this study solely focus on evaluation in relation to the X-RAI framework as an artifact accordingly to the design principles of ADR. The quality insurance

mechanism and evaluation of governmental conduct and work in the DBA is beyond the scope of this study. Pointing towards a lack of evaluation of a given AI system must not be interpreted as no evaluation or quality insurance at all for the given AI system. Understanding how AI impacts governmental conduct is a direction for future research. We still need to investigate how to design tools to aid the evaluation of AI systems. We know that data drift occurs and altering the AI systems performance. The focus seems to have been on the impact of metrics of accuracy. It is a natural next step to research data drifts influence on bias and fairness over time.

References

Asatiani, A., Malo, P., Nagbøl, P.R., Penttinen, E., Rinta-Kahila, T., Salovaara, A., 2021. Sociotechnical Envelopment of Artificial Intelligence: An Approach to Organizational Deployment of Inscrutable Artificial Intelligence Systems. Journal of the Association for Information Systems 22, 325–352. https://doi.org/10.17705/1jais.00664

Asatiani, A., Malo, P., Nagbøl, P.R., Penttinen, E., Rinta-Kahila, T., Salovaara, A., 2020. Challenges of explaining the behavior of black-box AI systems. MIS Quarterly Executive 19, 259–278.

Baird, A., Maruping, L.M., 2021. The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts. MIS quarterly 45.

Berente, N., Gu, B., Recker, J., Santhanam, R., 2021. Managing artificial intelligence. MIS Q 45, 1433–1450.

Charmaz, K., 2006. Constructing Grounded Theory: A Practical Guide through Qualitative Analysis. Sage, Thousand Oaks, CA.

Doshi-Velez, F., Kim, B., 2017. Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608 [cs, stat].

Eisenhardt, K.M., 1985. Control: Organizational and economic approaches. Management Science 31, 134–149.

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

Gregor, S., Kruse, L.C., Seidel, S., 2020. Research Perspectives: The Anatomy of a Design Principle. Journal of the Association for Information Systems 21, 1622–1652. https://doi.org/10.17705/1jais.00649

Hernández-Orallo, J., 2017. Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. Artificial Intelligence Review 48, 397–447. https://doi.org/10.1007/s10462-016-9505-7

Hill, K., 2020. Wrongfully accused by an algorithm, in: Ethics of Data and Analytics. Auerbach Publications, pp. 138–142.

Kirsch, L.J., 2004. Deploying common systems globally: The dynamics of control. Information Systems Research 15, 374–395.

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. Management Information Systems Quarterly.

Li, J., Li, M., Wang, X., Thatcher, J.B., 2021. STRATEGIC DIRECTIONS FOR AI: THE ROLE OF CIOS AND BOARDS OF DIRECTORS. MIS Quarterly 45.

Lipton, Z.C., 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16, 31–57.

Lou, B., Wu, L., 2021. AI ON DRUGS: CAN ARTIFICIAL INTELLIGENCE ACCELERATE DRUG DEVELOPMENT? EVIDENCE FROM A LARGE-SCALE EXAMINATION OF BIO-PHARMA FIRMS. MIS Quarterly 45.

Mayer, A.-S., Strich, F., Fiedler, M., 2020. Unintended Consequences of Introducing AI Systems for Decision Making. MIS Quarterly Executive 19.

Nagbøl, P.R., Müller, O., 2020. X-RAI: A Framework for the Transparent, Responsible, and Accurate Use of Machine Learning in the Public Sector, in: Proceedings of Ongoing Research, Practitioners, Workshops, Posters, and Projects of the International Conference EGOV-CeDEM-EPart 2020. Presented at the EGOV-CeDEM-ePart 2020, p. 9.

Nagbøl, P.R., Müller, O., Krancher, O., 2021. Designing a Risk Assessment Tool for Artificial Intelligence Systems, in: Chandra Kruse, L., Seidel, S., Hausvik, G.I. (Eds.), The Next Wave of Sociotechnical Design. Springer International Publishing, Cham, pp. 328–339.

Nations, U., 2018. United Nations E-Government Survey 2018. United Nations.

Recker, J., Indulska, M., Green, P., Burton-Jones, A., Weber, R., 2019. Information Systems as Representations: A Review of the Theory and Evidence. Journal of the Association for Information Systems 20, 5.

Russell, S., Norvig, P., 2002. Artificial intelligence: a modern approach.

Ryan, R.M., Deci, E.L., 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American psychologist 55, 68.

Secretariat, Treasury Board of Canada, 2020. Algorithmic Impact Assessment (AIA). aem.

Sein, M., Henfridsson, O., Purao, S., Rossi, M., Lindgren, R., 2011. Action Design Research. Management Information Systems Quarterly 35, 37–56.

Strich, F., Mayer, A.-S., Fiedler, M., 2021. What do I do in a world of artificial intelligence? Investigating the impact of substitutive decision-making AI systems on employees' professional role identity. Journal of the Association for Information Systems 22, 9.

Sun, T.Q., Medaglia, R., 2019. Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare. Government Information Quarterly 36, 368–383.

Teodorescu, M.H., Morse, L., Awwad, Y., Kane, G.C., 2021. FAILURES OF FAIRNESS IN AUTOMATION REQUIRE A DEEPER UNDERSTANDING OF HUMAN-ML AUGMENTATION. MIS Quarterly 45.

United Nations. , Department of Economic and Social Affairs, 2020. United Nations e-government survey 2020: digital government in the decade of action for sustainable development. United Nations, Department of Economic and Social Affairs, New York.

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.

Yin, R.K., 2009. Case study research: Design and methods. Sage, Thousand Oaks, CA.