Journal of the Association for Information Systems (2021) **22**(2), xxx-xxx
doi: 10.17705/1jais.00491

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).

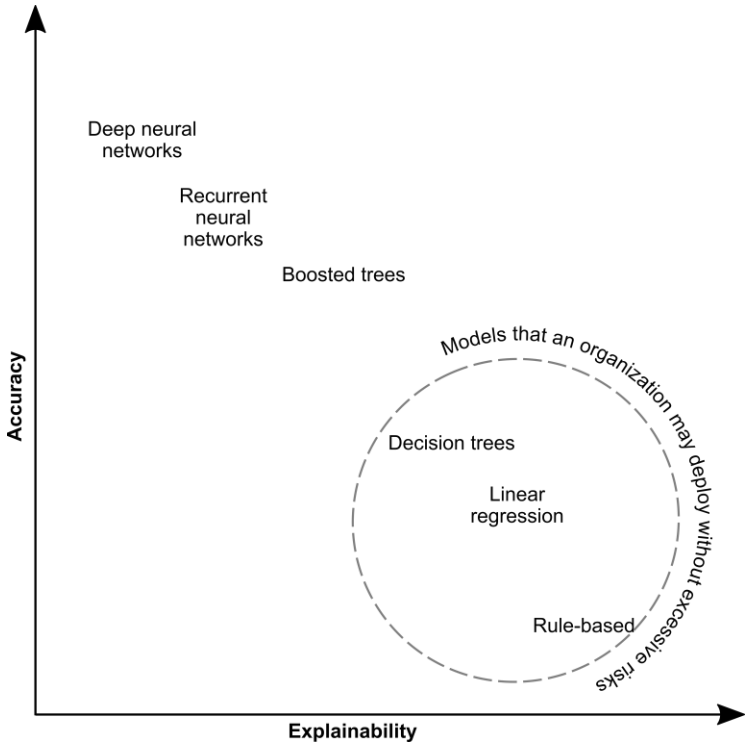


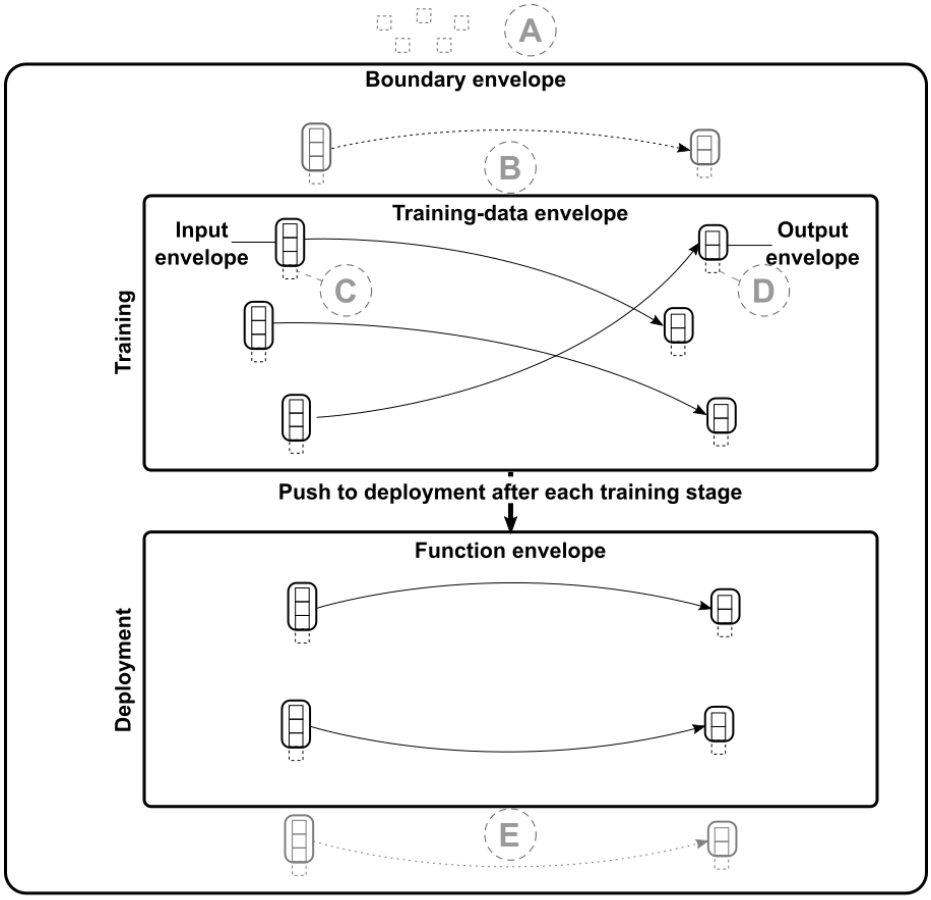
Figure 1. The Explainability-Accuracy Tradeoff

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 various 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.

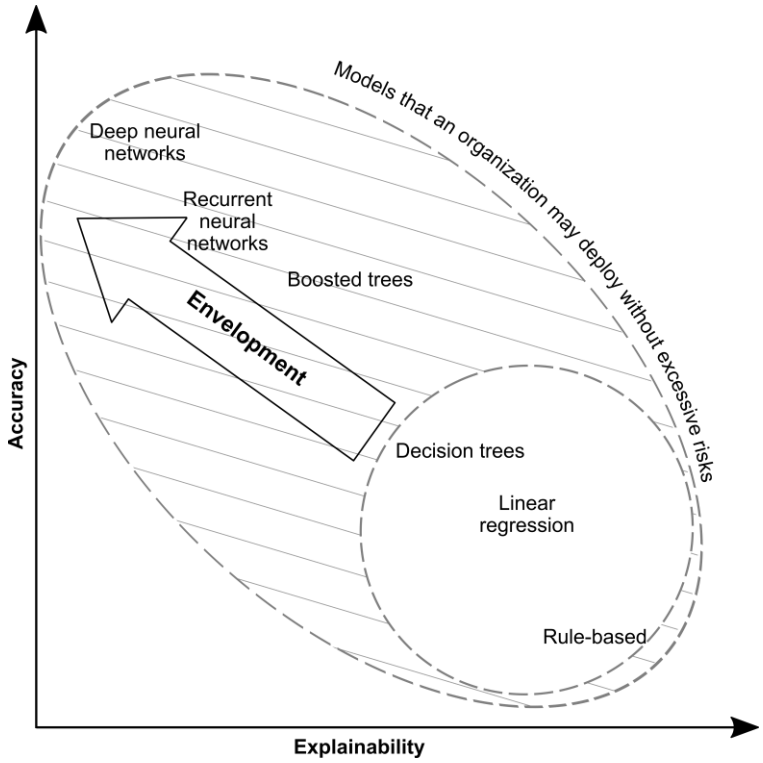


Figure 3. How Envelopment Expands the Set of Models an Organization May Adopt Without Excessive Risks

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*.