Toward Sociotechnical AI

Mapping Vulnerabilities for Machine Learning in Context

Dobbe, Roel; Wolters, Anouk

DOI
10.1007/s11023-024-09668-y

Publication date
2024

Document Version
Final published version

Published in
Minds and Machines

Citation (APA)
Dobbe, R., & Wolters, A. (2024). Toward Sociotechnical AI: Mapping Vulnerabilities for Machine Learning in Context. *Minds and Machines*, *34*(2), Article 12. https://doi.org/10.1007/s11023-024-09668-y

Minds and Machines (2024) 34:12
https://doi.org/10.1007/s11023-024-09668-y

Toward Sociotechnical AI: Mapping Vulnerabilities for Machine Learning in Context

Roel Dobbe[1] · **Anouk Wolters**[1,2]

Received: 18 March 2023 / Accepted: 7 January 2024

Abstract
This paper provides an empirical and conceptual account on seeing machine learning models as part of a sociotechnical system to identify relevant vulnerabilities emerging in the context of use. As ML is increasingly adopted in socially sensitive and safety-critical domains, many ML applications end up not delivering on their promises, and contributing to new forms of algorithmic harm. There is still a lack of empirical insights as well as conceptual tools and frameworks to properly understand and design for the impact of ML models in their sociotechnical context. In this paper, we follow a design science research approach to work towards such insights and tools. We center our study in the financial industry, where we first empirically map recently emerging MLOps practices to govern ML applications, and corroborate our insights with recent literature. We then perform an integrative literature research to identify a long list of vulnerabilities that emerge in the sociotechnical context of ML applications, and we theorize these along eight dimensions. We then perform semi-structured interviews in two real-world use cases and across a broad set of relevant actors and organizations, to validate the conceptual dimensions and identify challenges to address sociotechnical vulnerabilities in the design and governance of ML-based systems. The paper proposes a set of guidelines to proactively and integrally address both the dimensions of sociotechnical vulnerability, as well as the challenges identified in the empirical use case research, in the organization of MLOps practices.

Keywords Machine learning · Artificial intelligence · MLOps · Design science research · Sociotechnical systems · Vulnerabilities · System safety

✉ Roel Dobbe
r.i.j.dobbe@tudelft.nl

[1] Technology, Policy and Management, Delft University of Technology, Jaffalaan 5, 2628 BX Delft, The Netherlands

[2] Deeploy, Oudegracht 91A, 3511 AD Utrecht, The Netherlands

Published online: 17 May 2024

1 Introduction

Following promises for economic and societal benefits across industries and public domains (European Commission, 2021), artificial intelligence (AI) tools and functions are rapidly adopted in high stakes social domains, reshaping many public, professional, and personal practices (Whittaker et al., 2018). While AI tools often have the potential to increase efficiency and improve decision-making, these can also lead to harms and violations of fundamental rights related to non-discrimination or privacy (Balayn & Gürses, 2021). Other emerging harms include physical dangers related to new robotic systems such as autonomous vehicles, and digital welfare systems leading to grave financial and mental harm (Dobbe et al., 2021).

In response, many efforts have emerged about to anticipate and address the implications of AI through appropriate governance strategies. These included a first wave of ethical principles and guidelines (Jobin et al., 2019), as well as technical tools for addressing issues of bias, fairness, accountability and transparency (Whittaker et al., 2018). While these guidelines and tools helped develop broader awareness of the governance challenges, there is still little known about how to situate and operationalize these principles and tools in the practice of developing, using and governing AI systems. At the contrary, critical scholars have argued that these instruments are often pushed as forms of self-regulation by industry to prevent more stringent forms of regulation (Wagner, 2018; Whittaker et al., 2018).

In technical fields, harms imposed by AI systems are primarily characterised as 'bias' or 'safety' flaws that can be adressed in the design of the technical system, leading to a focus on technical solutions (Balayn & Gürses, 2021). This way, the broader social and normative complexity of harms and the relation to design choices are naively narrowed down to a problem in the technical design of AI systems, and thus in the hands of technology companies or internal developers thereby foregoing normative deliberation and accountability (Green, 2021; Nouws et al., 2022). However, problems such as discrimination cannot be tackled only by technology specialists, but require a more holistic specification and evaluation of AI systems in their sociotechnical context (Dobbe et al., 2021).

Based on a structured literature review of the scholarly literature on AI and public governance, Zuiderwijk et al. (2021) list various knowledge gaps motivating a more the need for a comprehensive *sociotechnical system perspective* for the governance of AI systems. Firstly, AI is mostly addressed generically, and there is great need for more domain-specific studies. In every domain there are different actors, legacy practices and infrastructures that an AI system operates in. Understanding the broader system that an AI technology operates in requires a mix of methods that can capture complex interactions across stakeholders and technological features (Ackerman, 2000). A sociotechnical system lens can comprehensively describe such complexity and allow for meta-analysis and cross-domain comparison (de Bruijn & Herder, 2009). Furthermore, there is little empirical testing of AI systems in practice: "[a]s AI implementations start to bear

fruit (or cause harm) [...], there is an urgent need to pursue explanatory research designs that adopt expanded empirical methods to generate operational definitions, extract meanings, and explain outcomes specifically within public governance contexts" (Zuiderwijk et al., 2021)

In this paper we pursue empirical research to understand the extent to which existing design, use and governance practices for machine learning (ML) models are able to address the sociotechnical vulnerabilities of ML applications through which safety hazards may emerge. To map these vulnerabilities we perform an integrative literature review on sources of vulnerability of sociotechnical nature, based on recent literature on ML as well as lessons from system safety and other sociotechnical systems engineering disciplines that have dealt with sociotechnical vulnerabilities in software-based automation for a long time (de Bruijn & Herder, 2009; Dobbe, 2022; Leveson, 2012). The resulting conceptual dimensions for sociotechnical vulnerability are empirically grounded in interview-based case study research, also producing a set of challenges that emerge in addressing the dimensions in developing and governing ML applications in sociotechnical context. The key aim is to empower developers and other stakeholders who care about building safer and more just algorithmic systems to develop a shared language to address these vulnerabilities more effectively. Before we explain our contributions and methods in more detail in Sect. 1.2, we first cover related work. It is relevant to note that this study was performed in late 2021 and early 2022, and hence precedes the quick rise of generative AI tools since the end of 2022. Nonetheless, the findings in this paper largely apply, but may be nuanced or extended for more recent AI applications.

1.1 Related Work

The efforts to develop new practices for responsible AI system development are broad. Here, we put particular focus on efforts and critiques that explicitly mention and are informed by sociotechnical systems theory and engineering. The key findings in this related work echo two gaps identified by Zuiderwijk et al. (2021), namely a lack of empirical as well as conceptual accounts to better understand and describe the sociotechnical complexity of AI systems in practice and bridge technology, ethics and policy. In the following, we cover the most relevant papers, highlighting their affordances and limitations towards this aim.

Selbst et al. (2019) introduced the notion of sociotechnical systems in the discourse on fairness in machine learning. They mainly point out how fairness-aware research up till then abstracted away most context that surrounds the machine learning model, conceptualizing five traps that may contribute to undesirable narrowing of abstraction. They provide some high level takeaways but do not offer empirical grounding or engage with the ontological challenges inherent in defining ML as a sociotechnical system. Green (2021) critiques the tech ethics landscape pointing out the need for sociotechnical systems thinking to overcome the false assumptions on technology's neutrality, solutionism and determinism, without further elaborating how to ground such thinking in design practice. Winby and Mohrman (2018) develop a high-level organizational design approach for digital systems

incorporating sociotechnical analysis, but do not address the technical and sociotechnical dimensions of ML models themselves. Behymer and Flach (2016) point out the flaws in the dominant thinking of building autonomous systems separate from their human and social environment, proposing an alternative lens where the goal of design is a seamless integration of human and technological capabilities into a well-functioning sociotechnical system, based on Rasmussen's Skills, Rules, Knowledge (SRK) framework. While SRK leans on decades of experience in safety engineering, it lacks conceptual and empirical depth to address the particular nature of ML models and their interactions with context. Similarly, Oosthuizen and Van't Wout (2019) perform a study based on Cognitive Work Analysis (CWA) to understand the impact of AI technologies on users. This lens is relevant but mostly focused on the human agent with an eye for adoption, rather than understanding more integrally what kinds of vulnerabilities emerge in the broader human-AI system. Makarius et al. (2020) adopt an organizational approach, arguing that employees need to be socialized to develop so-called "sociotechnical capital" working with AI. Such a lens overlooks the risks and emergent hazards of adopting AI at a large scale. Martin et al. (2020) redefine machine learning applications as complex adaptive systems (CAS) thereby extending their boundaries of abstraction. The paper addresses the need to engage affected communities and incorporate diverse mental models of the designed system, which forms a welcome lens to introduce safety-related approaches that require deliberation and dissent (Dobbe, 2022), however it does not detail the ontological nature of machine learning technologies and their interactions in a broader CAS frame. Orr and Davis (2020) provide a much-needed empirical account showing the dynamic nature of AI systems being reshaped through various user interactions, however it does not engage with the technical workings of AI. Van de Poel (2020) introduces AI systems as fundamentally sociotechnical systems consisting of technical artifacts, human agents and institutions (rules to be followed by agents). The central insight in this account is that in order to embody and respect values, one has to design beyond the technical AI artefact, particularly addressing institutional norms, however no empirical grounding is provided. Dobbe et al. (2021) introduce a lexicon to redefine AI systems as fundamentally sociotechnical, with an emphasis on exploring the emergent normative nature of AI systems and their situated dynamics in the context of use and development practice. They propose cybernetic practices to define safety requirements through ongoing stakeholder deliberation and feedback channels for dissent, but do not provide empirical grounding. Salwei and Carayon (2022) use a healthcare use case to motivate that a sociotechnical systems approach requires that "the entire work system as well as clinical workflow must be systematically considered throughout the design of AI technology", thereby offering a tangible framework to situate the capability in the work setting. More conceptual work is needed to understand the various interactions between workers and AI capabilities. Mohamed et al. (2020) introduce post-colonial and decolonial critical theories to inform a sociotechnical foresight lens to examine issues of values, culture and power at play between stakeholders and technological artefacts. Such a lens allows seeing situated issues as instantiations of broader, often global, forms of power assymetry and violence, and enables inclusive dialogue between stakeholders in AI development, "particularly those in which marginalised groups have meaningful avenues to

influence the decision-making process, avoiding the potential for predatory inclusion, and continued algorithmic oppression, exploitation, and dispossession." That said, more empirical and practice-oriented research is needed to make this lens amenable to design.

1.2 Approach and Contributions

The central aim of this reseach is to aid practitioners by developing actionable guidelines to address sociotechnical dimensions of AI systems, particularly in the design, development, use and governance of machine learning models in their situated context. The research objectives are: (1) to conceptualize machine learning models in their situated context as sociotechnical systems; (2) to identify associated sociotechnical dimensions of the system through which vulnerabilities emerge that can lead to hazards; (3) to empirically understand when/where these vulnerabilities emerge in the practices of designing, developing and maintaining ML models; and (4) to conceptualize guidelines for anticipating and preventing harmful outcomes from lessons in system safety engineering.

The central unit of analysis of this paper are *practices* within organizations building or deploying machine learning (ML) models for their operational processes. We particularly focus on the *MLOps* process, which is a set of practices that is increasingly introduced in organizations to manage issues related to the technical performance of ML models in operational and design processes (Ruf et al., 2021). The MLOps process is further introduced in Sect. 2.

The research is both descriptive and design-oriented. On the one hand we want to understand and contribute to existing practices that are understudied. On the other hand, we do see valuable lessons in traditional sociotechnical systems disciplines, in particular in system safety in Dobbe (2022) and Leveson (2012). To combine descriptive empirical research with theoretical and conceptual design research for more sociotechnical practices, we adopt the Design Science Research (DSR) approach (Hevner et al., 2010). This framework was first developed to situate the design of IT artefacts (such as software tools or AI models) in their context and allow the evaluation and study of how such an artefact is used and changes behaviors and processes. DSR is an appropriate methodology in cases when requirements of an information system are not known yet and/or there is a high probability that the context will change as a result of the IT artefact in ways that are challenging to anticipate. Apart from technical design artefacts, DSR can also be used to work towards other relevant outputs, such as evaluation methods, organizational guidelines, policies, and work practices (Offermann et al., 2010). In this project, we follow the DSR cycles and work towards artefacts in the following way. In Sect. 2, we empirically map and analyze existing MLOps practices and corroborate these with literature. In Sect. 3, we report the outcome of an integrative literature study covering vulnerabilities of ML in context, culminating in eight sociotechnical dimensions in which such vulnerabilities emerge. In Sects. 4 and 5 we provide the key insights from an interview study with stakeholders, both actors engaged in the MLOps process and external actors involved in auditing or advocacy about the implications of

ML applications. This results in an empirical validation of the sociotechnical dimensions found in our integrative literature study, as well as an inductive analysis yielding seven key challenges for addressing sociotechnical vulnerabilities in the design and specification of ML applications in context. Lastly, in Sect. 6 we provide a set of guidelines for specification of ML in sociotechnical systems, integrated in the MLOps practices, based on lessons from system safety. These guidelines were not empirically validated (part of future work), and as such are only covered briefly in this paper, pointing interested readers to a longer treatment in the associated thesis (Wolters, 2022).

Hence, the contributions of the research are three-fold: First, the research expands the literature on sociotechnical ML development and operations, providing insights in vulnerabilities that emerge when ML is applied in context, and how it can synthesised in dimensions for designing a ML application as part of a sociotechnical system. Second, empirical data is collected by conducting interviews with experts in the field to ground, validate and augment the synthesised dimensions with insights from practice. Third, the research informs a practical method for organisations to develop and/or use ML applications as sociotechnical systems, or a way to re-evaluate ML models already used in practice.

2 Mapping Existing Practices for MLOps

In this section we investigate currently existing practices for managing ML applications, under the embrella of *MLOps*. MLOps is an approach that aims to ensure reliable and efficient ML development, deployment and operations (Ruf et al., 2021). MLOps is a combination of ML, DevOps and Data Engineering. It is a practice to automate, manage and speed up the operationalisation of ML models (build, test, and release), by integrating DevOps practices into ML (Ruf et al., 2021). DevOps is a development methodology for software aimed at bridging the gap between Development and Operations practices, emphasizing communication and collaboration, continuous integration of sofware updates, quality assurance of software systems and delivery with automated deployment utilizing a set of development practices (Jabbari et al., 2016). At its core, MLOps is the standardisation and streamlining of ML lifecycle management, and the general desire in MLOps is to automate the ML lifecycle as far as possible to speed up the deployment and operations processes (Treveil, 2020).

2.1 Empirical Mapping of the Machine Learning Lifecycle

As MLOps practices may vary from organization to organization, we sketch the processes we mapped in the empirical research performed for various use cases in the financial industry. All these case studies rely on one external company for developing MLOps practices, which allows us to work towards one mapping.

The mapping was based on insights drawn from a series of interviews with stakeholders active in the design and management of ML models, both within

the organizations as well as for a vendor company offering services in the design and deployment of the ML models. All actors were asked to draw the key steps in MLOps and their interdependencies. This resulted in the map presented in Fig. 1.

The process of bringing a ML model into practice is generally conceptualised as a ML lifecycle. The ML Lifecycle can be divided into three stages: experimental stage, deployment stage and operations stage. The experimental stage involves all steps that lead to the construction of a ML model as well as activities to improve, correct or enrich an existing ML model deployment. The deployment stage includes the steps to integrate the model in an organisation's operational processes and infrastructure, so that it can be used to make predictions that then form an input to various business processes. The operations stage comprises the monitoring of the model and application and may trigger reasons to revisit the design and training of a model based on certain performance indicators.

2.2 Lack of Context in MLOps

The ML lifecycle presented above reflects the dominant view for how ML applications are developed. This view focuses on hardware, software, algorithms, mechanical linkages and inputs/outputs. This view is narrowly focused on primarily technical components and factors. However, the ML application is embeeded in an operational process and part of a broader sociotechnical system, which also include other technical systems, stakeholders, decision-making logics, institutions and the final outcomes, as presented in Fig. 3 and which we will further elaborate in Sect. 3.2.

A sociotechnical system consists of technological, social, and institutional elements and is mostly defined by the interactions between these elements (Van de Poel, 2020). The primary emphasis on the ML model and various technical metrics in MLOps practices leaves us with a gap between the technical conceptual framework, the ML lifecycle, and the needed sociotechnical conceptualisation of the

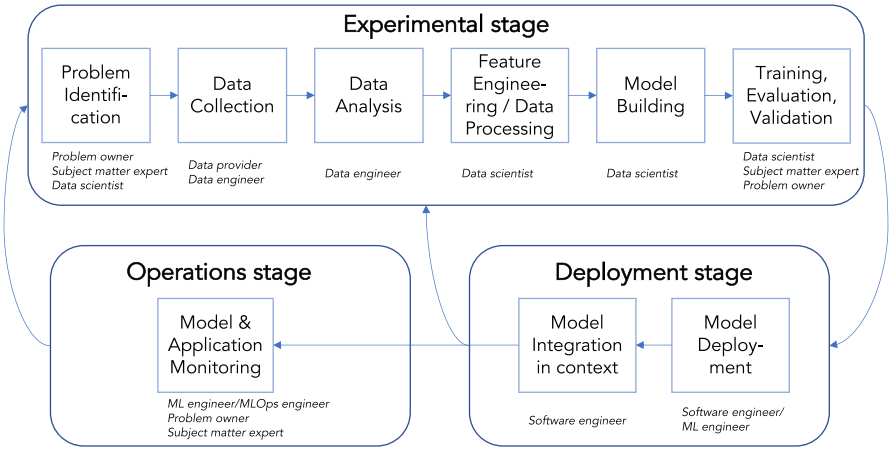

Fig. 1 MLOps practices mapped through interviews, sometimes referred to as the *ML lifecycle*. Under each activity, we list the roles of the professionals interviewed for the use cases in Sect. 4

ML-based applications in context (Alter, 2010), which we need in order to identify and address vulnerabilities that emerge in context (Leveson, 2012). For example, MLOps practices to not offer a lens to understand how in the construction of ML models and applications various dimensions of social context are abstracted away, such as critiqued in the context of fairness in ML (Selbst et al., 2019) or debiasing practices (Balayn & Gürses, 2021).

We can draw this point futher by looking at the key practices in MLOps. Continuous Integration (CI) enables automated model validation, after which the Continuous Delivery (CD) pipeline automatically delivers the model to be deployed. By definition, i.e. by virtue of their automatic nature, CI and CD practices do not consider the validation of the model's interactions with its sociotechnical context, including users and other technical systems that it depends on or which depend on the model's outputs. As such, a validated new model version in the CI pipeline could be valid from a technical perspective, but could be not meeting requirements of stakeholders in the sociotechnical system. This way, new model versions are released that could require changes in e.g. the design of decision-making process or interpretation of the model output by end-users.

If left unaddressed, these changes can cause new hazards to emerge at the level of sociotechnical interactions. In system safety, a field that has grappled with hazards in software-based automation for many decades, it is known that significant changes in the design of an automatic function require a *management of change procedure* to catch any such emergent hazards and prevent them from seeping into the operational process, which is a central procedure of safety management systems (Leveson, 2012).

The history of safety in software-based automation also tells us that vulnerabilities emerge from the interactions between social, technical and institutional components of the broader sociotechnical systems, including aspects like maintenance, oversight and management. Therefore, the identified technocentric view in MLOps, while being able to address technical or mathematical vulnerabilities in the ML application itself, is not sufficient to understand a variety of related and additional vulnerabilities emerging when using ML in context. In the next section, we adopt a sociotechnical systems lens to identify vulnerabilities that may emerge in the development and use of ML models.

3 A Sociotechnical Systems Lens for Machine Learning in Context

3.1 Scoping the Study

In this paper, we make a first step towards a sociotechnical specification approach for ML applications that can do justice to the emergent nature of key values as safety or fairness. We particularly focus on understanding what kind of vulnerabilities emerge in the development and use of ML applications in their social and institutional context that may impact such values. The aim is to bring relevant vulnerabilities into view, to inform actors involved in specifying and designing ML applications to take these into account and work towards safer and better functioning system design.

This way of designing is inspired by system safety (Dobbe, 2022; Leveson, 2012), in which safe behavior of a system is expressed in terms of satisfying key safety constraints, expressed across a sociotechnical system's social, technical and institutional components and their interactions. We work towards a canonical set of dimensions of sociotechnical vulnerability which can inform such safety-guided specification and design in future efforts.

As a result, we extend the boundary of analysis for ML to include essential system component interactions in the context of development and use. Figure 2 shows the scope of our study. It includes looking at the ML model, its integration in an application and MLOps process for design, deployment and operations, which we term the *ML application*. We also look at the interactions between the ML application and the most immediate context of use, which we look at as a *sociotechnical system*. Our lens includes the key decision-making process and outcomes that the ML application contributes to, as well as any directly applicable institutions (established laws, practice, or customs that typically impose a norm or standard), dependencies on computational and broader digital infrastructure, and the different users and stakeholders interacting with the ML application and the operational process in which it is used.

In this study, we do draw a line in how far our scope reaches in addressing broader organizational and institutional structures surrounding the main decision-making process and context of use. We note that broader structures can have a pivotal role in safeguarding software-based automation and, as such, provide additional sites of vulnerabilities (Leveson, 2012). These aspects include the various mechanisms and procedures for supervision and oversight. We do interview some actors responsible for such supervision, but we do not address the conditions for effective oversight itself. Furthermore, we do not exhaustively analyze the cultural dimensions of leadership and management. While these are known to be crucial for upholding safe automation systems (Dekker, 2016), and as such are discussed in some of our interviews, the intricacies of analyzing culture within the case studies itself are beyond the scope of this study. Lastly, the implications of ML applications are often of a political nature, and adopting an application system may be motivated by political motives, which may include the reification of historical power assymetries and

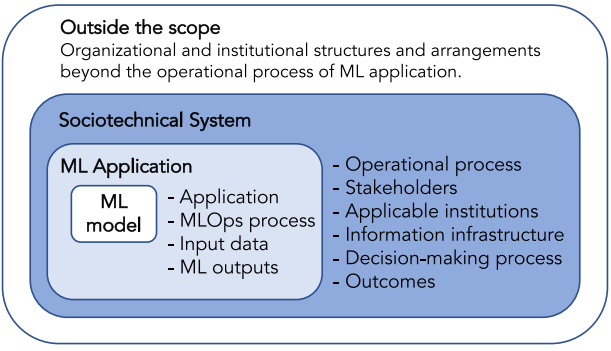

Fig. 2 Scope of the sociotechnical analysis

forms of discrimination (Mohamed et al., 2020). While we are aware of these complexities, the work does not explicitly analyze these dimensions.

3.2 Dimensions of Sociotechnical Vulnerability

To be able to guide practitioners in the sociotechnical specification, an understanding of the vulnerabilities that can emerge in the sociotechnical context of ML applications must be reached. Vulnerabilities are potential shortcomings in the ML application and its and sociotechnical context which may turn into hazardous situation from which harm to stakeholders can emerge. Hazards are defined as "a system state or set of conditions that, together with a particular set of worst-case environmental conditions, will lead to an accident (loss)" (Leveson, 2012, p. 184). Hazards are thus situations in the operational process in which the ML model plays a role that should occur as little as possible. Ideally, hazards are designed out of the system, or, otherwise, their occurrence or impact should be minimized.

While there are general techniques to map and analyze sociotechnical vulnerabilities and their associated hazards, each vulnerability tends to be contextual and require analysis of situated complexities. Nevertheless, for ML applications and across the history of system safety, we have now access to many empirical and theoretical accounts of vulnerabilities. Currently, there is no conceptual framework that covers these vulnerabilities in a way that is amenable to analysis of ML applications in their sociotechnical system context. Therefore, we construct a conceptual categorization of known vulnerability archetypes in the literature, based on an integrative literature review on vulnerabilities that emerge in sociotechnical systems, both in the machine learning literature and the historic literature on system safety.

Vulnerabilities were first identified by means of an integrative literature review. A total of twenty-four unique vulnerabilities were identified. These vulnerabilities are listed in Appendix 1. Next, these vulnerabilities were synthesized and conceptualized in eight *dimensions* of sociotechnical systems in which these occur or emerge. To arrive at these dimensions, a Grounded Theory building method was used. This method follows an inductive approach in order to generate or discover theory (Torraco, 2002). The theory evolves through continuous interplay between analysis and data collection. In this research, the data collected is data on vulnerabilities that emerge in the sociotechnical system context, as identified in scientific literature. The analysis comprehends the interpretation of these vulnerabilities by combining them with knowledge of the ML lifecycle and sociotechnical specification, as mapped and corroborated in the earlier sections. Figure 3 provides a conceptual overview and visualization of the dimensions. It is important to note that there is no one-to-one mapping from vulnerabilities (as listed in Appendix 1) to the dimensions. Instead, the dimensions serve as a meta-categorization with which each vulnerability can be characterized to emerge through different dimensions. For example, the occurrence of a false positive that was not detected by a human user or interpreter of the machine learning model output, is an expression of both misinterpretation and error. Or, a model that does

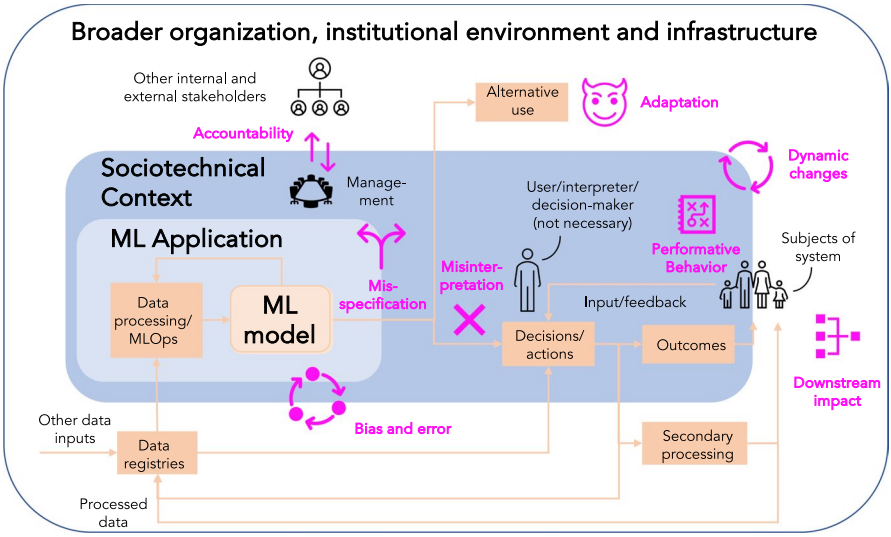

Fig. 3 Resulting dimensions of possible vulnerabilities emerging in the sociotechnical context of machine learning models and applications, based on the scope laid out in Fig. 2. The identified dimensions are represented by the bright pink color icons. The boxes and arrows denote the typical flows of digital information, in the form of data and model outputs, and their impact on decisions, other uses and outcomes

not consider important elements of the context and therefore makes mistakes, is a combination of the dimensions of misspecification and bias and error. In the following subsections, we discuss each of the eight resulting dimension.

3.2.1 Misspecification

Vulnerabilities can be caused by misspecification, which entails mistakes or gaps in the specification of the broader sociotechnical system. As ML models are an integrative part of larger sociotechnical systems, it should be acknowledged that the specification should be done at the level of the operational process in which the ML model is deployed. Facilitating the interaction between ML model and other technical, social and institutional components should be part of such specification. The absence of such specification may easily lead to ML applications that do not serve the needs of users and other stakeholders impacted, do not comply with various regulations, or cause new emergent forms of error, hazard and harm. Here we distinguish between cases of misspecification that arises because there is a lack of consensus on the outcomes and behavior of the intended system, and cases for which there is consensus, but the resulting specification is incorrect. For the former part, there is the need to address the possible normative complexity that may arise from different stakeholders holding different values or interests leading to conflicts in the specification, which may or may not lead to vulnerabilities in the eventual design and operation of the system (Dobbe et al., 2021; Van de Poel, 2015).

3.2.2 Bias and Error

Machine errors occur in the ML model resulting in errors in the model output, with impact on people, processes and organisations. It is important to be aware of the types of machine error that could occur in the development of ML applications or in the ML model output and what the potential impact could be and to whom. Machine error can be divided into two categories: machine incorrectness and machine bias. Machine incorrectness refers to a false negative or false positive output. Machine bias refers to forms of disproportionate differences in the error rates of a model across different groups or attributes of people subject to the system's outputs. Both errors as well as the resulting biases may be pre-existing in the data and social practices from which these are derived, or they may be encoded in the technical design of the ML application or operating process, or they may emerge in the operation and maintenance of the system (Dobbe et al., 2018).

3.2.3 Interpretation

The interpretation of model output by a human decision-maker is a source of vulnerabilities. The model output can be misinterpreted due to a lack of understanding how the model output is generated. Further, humans can over-rely on the model output without the consideration of other factors, by which errors in the model are adopted in the final decision, a phenomenon called *automation bias*. Human-ML interactions may also cause new forms of bias to emerge through so-called *disparate interactions* (Green & Chen, 2019a).

Additionally, human decision-making brings about error that can be reduced or enforced by the introduction of a ML model in the decision-making process. Human error can be divided into two categories: bias and noise. Noise is an unwanted variability in professional judgement across different decision-makers, whereas bias is the systematic error that is made by humans in the judgement of certain situations (Kahneman et al., 2016). ML models may be an aid in reducing noise and bias when designed and used properly, but can also confirm or even reinforce existing or create new forms of noise and bias.

ML models can be hard to interpret for humans, due to their opaque or complex nature. Different approaches to deal with the interpretability of models can infer vulnerabilities. Some models are so complex that it is impossible to understand for humans how decisions are made, which are often referred to as black-box models. In system safety, the growing complexity of models for control and decision-making leads to the *curse of flexibility*, which refers to the tendency of software-based automation to grow in complexity, hence running into the inherent constraints on people's cognitive and intellectual abilities to keep understanding its workings, effects and limitations (Leveson, 2012).

An approach is to improve the interpretability of models is to make ML models explainable. The EU requires a 'right to explanation' of ML models (Rudin, 2019). However, explanation is a highly contextual concept that depends on the situation and the actors involved in giving and receiving an explanation, making this an ambiguous requirement (Miller, 2019). Lastly, effort could be put in making ML

models inherently interpretable. The way human decision-makers interpret model outputs can thus affect the ultimate decisions made using the ML application. This raises the question of how ML applications should be incorporated into decision-making processes (Green & Chen, 2019a).

3.2.4 Performative Behaviour

When a ML application is integrated into a social context, the ML model's outputs will interact with a pre-existing social system, consisting of actors in that system (Selbst et al., 2019). It is well-known in economics and policy fields that predictions may influence the outcome they aim to predict due to *performativity*. The behaviour of actors can be affected by the introduction of an ML application, because model's logics may generate incentives to behave in a certain way to influence the model's outputs. Those actors could be human decision-makers using the ML application, or actors that are affected by the decisions made with help of the ML application (Milli et al., 2019). However, performativity is largely ignored in supervised learning literature and practices (Perdomo et al., 2020). To truly understand what the outcome of using a ML application is, it is vital to measure and validate how the distributions of predictions and outcomes shift over time. If human behaviour is delineated too much or wholly ignored a priori in the specification of the sociotechnical system, the risk emerges that if humans do not behave as expected, the sociotechnical system leads to unwanted outcomes. To prevent this, systems could accommodate the autonomy of users in contexts where this is wished or aim to actively account for performativity in designing ML-based applications.

3.2.5 Adaptation

In the specification of the system, it is not entirely possible to predict how users of the ML application will be using the system in terms of rational work processes (Rasmussen, 2000). Developers of the ML application should be aware that operators or users of the system could deviate from the specified use of the system to address the complexity of the environment. A work environment in which a ML application is integrated is often complex, and humans that work in that environment have the ability to cope with this complexity by inventing clever strategies that do not match with what the system developers consider 'rational behaviour' of the user (Rasmussen, 2000). System developers should be aware that people can adapt in order to use a system for an unintended purpose, possibly outside of the assumed sociotechnical context, which may include harmful dual use. Therefore, assuming them to be rational and behave in a certain way is problematic.

3.2.6 Dynamic Change

Once a ML function is put in production, vulnerabilities could also emerge over time due to dynamic change of the context and broader environment. The world is continuously changing, leading to data shifts, concept drifts, changes in regulations or organisational strategies and priorities (Selbst et al., 2019). Dynamic change can

have direct effect on the performance of the ML model outputs, for example through a decrease in accuracy. To address this decrease in performance, a ML model can be updated with new data and retrained. However, if done without care, this may lead to runaway feedback-loop that reinforce bias over time (Ensign et al., 2018). Further, dynamic change might necessitate revisions in the broader process and sociotechnical system, for example to adhere to new regulation. To conclude, it is key to anticipate dynamic change in the environment of the ML application and its possible impacts on the outputs of the model and outcomes of the associated process.

3.2.7 Downstream Impact

The introduction of ML models in decision-making processes, can have large downstream impact throughout the organisation and beyond. First, decisions made with ML in one process, can be stored in data registers and have impact on other processes (Peeters & Widlak, 2018). Therefore, it is important for stakeholders to consider the downstream impact of decisions made in the ML lifecycle. Further, machine error in the model could work its way to secondary processes or systems in which the model output is used. Additionally, it could be human error in the final decision made using the model output that is used in secondary processes or systems. Moreover, if bias is inferred in the ML lifecycle, this could have downstream impact on the people that are ultimately affected by the decisions made using a ML application, or are affected by secondary processes or systems that are influenced by the ML application.

3.2.8 Accountability

The processes and stakeholders that depend on the outputs of a ML model and decisions made using a ML application should be held accountable in appropriate ways for the resulting outcomes. However, it is often difficult to determine who carries the accountability of what part of the system, leading to various accountability and responsibility gaps (Loi & Spielkamp, 2021; Raji et al., 2020; Santoni de Sio & Mecacci, 2021; Whittaker et al., 2018). Undefined or unclear assignment of accountability is a primary source of vulnerabilities and system hazards (Leveson, 2012). However, an overemphasis on accountability and blame may also backfire, as it could deteriorate the willingness of professionals to open up about possible vulnerabilities in fear of retribution. Therefore, accountability is a core challenge in any safety-critical or socially sensitive context, requiring a careful balancing with safety concerns (Dekker, 2016).

4 Empirical Validation

The second phase of the project entails the validation of the eight dimensions of sociotechnical vulnerability in various application environments. The goal is to get a empirically grounded and validated understanding of the findings resulting from the literature review and theory development.

We first introduce the case studies. We then present our empirical analysis based on interviews with relevant stakeholders in these case study contexts as well as with actors from external organizations. The empirical analysis validates the vulnerability dimensions and presents challenges found across different stakeholders engaged in the design, development, use and governance the addressed ML applications.

4.1 Use Cases

The application environment for this research is the use of ML in the financial domain. In order to gather data about the application environment, two use cases within this application environment are studied, one on the use of ML for financial crime detection and one on the use in email marketing.

4.1.1 Selection of Use Cases

The second phase of the research aims to create an in-depth understanding of the practice of specification of ML applications and the addressing of the sociotechnical dimensions synthesised in Sect. 3.2. To do so, the financial domain has been chosen as the application environment to dive into. There is great variety within the financial domain in organisations that use ML, the maturity of the organisations regarding ML and the context in which ML models are used. To get insight in the differences and similarities, two ML use cases with different characteristics have been selected for analysis. A use case is a specific situation in which ML is developed and used.

In the case studies, the development of the ML application as well as the use of the ML application in practice are studied. The use cases researched are situated within different banks, serve a different goal, the type of decision-making process varies and the ML models were either developed completely in house or by partnering with and external firm. To gain a rich understanding of the use cases, a variety of stakeholders involved are interviewed.

4.1.2 Use Case Descriptions

4.1.2.1 Use Case 1: Financial Crime Detection Banks in The Netherlands are obliged by law [Law for the prevention of money laundering and terrorist financing (Wwft)] to report unusual transactions to the Financial Intelligence Unit Nederland Financial Intelligence Unit (xxx). Also, Dutch banks have to do customer investigations AFM (xxx). As such, banks have a responsibility to detect suspicious behaviour in order to mitigate the risk of money laundering and financing terrorism. We consider a case study at a Dutch bank (hereafter referred to as 'Bank A'). Bank A installed a rule-based system that sends out alerts on suspicious transactions to the transaction monitoring analysts within the bank. Transaction monitoring analysts then start an investigation of the customer(s) involved. They assesses whether the customer is indeed performing suspicious activities, after which the customer may be further reviewed by another department within the bank. Finally, the bank reports such customers to the Financial Intelligence Unit. However, the transaction monitoring ana-

lysts encountered the problem that the rule-based system provided many alerts that were in the end not assessed as suspicious activity (i.e. a high rate of false-positive alerts). As the bank wants to detect as many suspicious activities as possible, but in an effective and efficient manner, this was problematic.

To improve this process, Bank A decided to develop ML models. The first ML model was developed to decrease the number of false-positives from the rule-based system, by predicting whether an alert is actually a true-positive alert or a false-positive alert. The predicted true-positive alerts are then pushed forward to the transaction monitoring analysts to investigate. The transaction monitoring analysts investigate the alert to detect potential financial crime. As such, they ultimately assess whether the alerts are actual true-positive alerts or not. The second model is a model that detects new suspicious behaviour, that is not detected by the rule-based system, to be pushed forward to the transaction monitoring analysts to investigate. The transaction monitoring analysts use the model output as a source of information for their investigations.

4.1.2.2 Use Case 2: Email Marketing At a different Dutch bank (hereafter called 'Bank B'), there was a wish to better inform private customers about investing opportunities, through personalization, in order to ultimately increase the conversion rate from customers without an investing account towards investing. Before a ML application was developed, the bank did a marketing campaign for investing a few times a year, in which every private customer without an investing account received an email with information about investing. To increase the conversion rate, the bank wanted to move towards a more year-round way of marketing for investing.

In order to do so, the Marketing Intelligence team of the bank had partnered with an external ML engineering/consulting firm to develop a ML application. The ML model was developed to predict the probability of conversion to investing for customers at a certain moment in time. If the probability is higher than a set threshold, this customer will be presented in the model output. The Marketing Intelligence Analyst then receives this output and pushes the customers to an emailing system, from which these customers receive an offer. The Marketing Intelligence Analyst performs some additional checks before the emails are sent out.

4.1.3 Selection of Interviewees

In total, 18 professionals have been interviewed by means of semi-structured interviews. Eleven professionals were directly or indirectly involved in the use cases within the banks and the external ML engineering/consulting firm. They have been interviewed to gather insights into the application environment. The interviewees were selected as follows. The key roles identified in the core practices in MLOps, as indicated in Fig. 1, were selected as a first step to find interviewees. Furthermore, the sociotechnical systems lens that is taken throughout this project provided directions for selecting additional interviewees. This way, interviewees were selected that were involved in the specification of the ML application, the development of the ML application, the use of the model in a decision-making process, and in compliance. The interviewees sometimes fulfil more than one role as presented in Fig. 1. Two

interviewees within bank A, are not directly involved in the development of this use case, but provided overarching insights on integration of use cases and responsible AI within the bank in general. Further, in the email marketing use case, no compliance officer has been interviewed, due to unavailability. The names of the interviewees and the names of the organisations they work at are not disclosed due to privacy reasons.

Besides stakeholders involved in the practical use cases, seven interviews have been performed with stakeholders not directly involved in the use cases and outside the banks. These interviews were performed to get a broader understanding of the implications of ML applications for customers of the banks and the larger public, and to understand how these relate to the sociotechnical dimensions identified in Sect. 3.2. Representatives of several civil society organisations and regulatory bodies were interviewed, including from the organisations Waag, Amnesty International, Privacy First, Platform Bescherming Burgerrechten, Bits of Freedom, The Dutch Data Protection Authority (Autoriteit Persoonsgegevens) and the Dutch National Bank (De Nederlandsche Bank, or DNB). The names of the representatives were not disclosed due to privacy reasons. Table 1 presents an overview of all interviewees, including their function, stakeholder role(s), use cases involvment, and organisation. The detailed protocols used for the interviews can be found in (Wolters, 2022).

5 Interview Analysis

To synthesize the empirical insights gathered in the expert interviews, two types of analyses are carried out: a deductive analysis and an inductive analysis. For the deductive analysis, the coding frame has been developed at the beginning of the analysis process (Friese et al., 2018). This frame contained the dimensions defined in Sect. 3.2 as codes. During the analysis, the code frame has been enriched with additional codes to cover the content of the entire data set (Friese et al., 2018). This results in analysis on the level of the dimensions, which describes to what extent the dimensions are addressed in the use case. Afterwards, an inductive analysis is performed to identify the main challenges in ML practice. An inductive analysis is data-driven, and the researcher is not trying to fit the data into a pre-existing coding frame (Friese et al., 2018). This analysis allows us to analyse along the dimensions and additional interview data to identify the main challenges in ML practice from a sociotechnical systems view.

5.1 Results of a Deductive Analysis of the Sociotechnical Dimensions in Practice

This section presents the results of a deductive analysis, in which is investigated to what extent the dimensions described in Sect. 3.2 are considered by stakeholders involved in the ML lifecycle of the use case, and whether vulnerabilities related to these dimensions actually emerged. Additionally, the perceptions of the civil society organisations and regulatory bodies broaden the view on the sociotechnical

Table 1 Interviewee information

Interviewee function	Stakeholder role(s)	Use Case	Organisation
Manager data science	Product owner	Financial Crime Detection	Bank A
Lead data scientist	Data scientist	Financial Crime Detection	Bank A
Transaction monitoring analyst	End-user, subject matter expert	Financial Crime Detection	Bank A
Data Engineer	Data engineer	Financial Crime Detection	Bank A
Privacy Officer	Compliance officer	Financial Crime Detection	Bank A
ML engineer	ML engineer, MLOps engineer	Use case overarching	Bank A
Enterprise data science advisor	Responsible AI integration	Use case overarching	Bank A
Manager marketing intelligence	Product owner	Email marketing	Bank B
Marketing intelligence Analyst	End-user, subject matter expert	Email marketing	Bank B
Marketeer	subject matter expert	Email marketing	Bank B
ML engineer/project leader	Data scientist, ML engineer, MLOps engineer	Email marketing	External ML firm
Representative	Reflection of technology	–	Waag
Representative	Addressing human rights	–	Amnesty International
Representative	Addressing privacy protection	–	Privacy First
Representative	Addressing legal protection on civil rights	–	Platform Bescherming Burgerrechten
Representative	Addressing human rights	–	Bits of Freedom
Representative	Regulatory body	–	Dutch Data Protection Authority
Representative	Regulatory body	–	DNB

dimensions and provide a sense of relevance of each dimension. This leads to answering the sub-question: *To what extent are sociotechnical dimensions addressed in practice, based on use case specific and general insights?*

5.1.1 Misspecification

In both use cases, there exists awareness among the data science stakeholders that it is important to consider the larger sociotechnical system when developing ML applications. Nevertheless, examples of misspecification can be identified in both use cases. The misspecification dimension has been found relevant in practice among civil society organisations and regulatory bodies, which point out that technology is not the solution to every problem. Besides that, not specifying the larger sociotechnical system of which the ML model is part can lead to harmful consequences for organisations and individuals.

5.1.1.1 Misspecification in the Use Cases In the financial crime detection use case, the manager of data science does realise that it is vital to take the role of the end-user into account during the development of the ML application, as the value of the system would be zero if the end-user would not understand what comes out of the system. While the realisation is there, an example of misspecification can be identified in this use case. The description of the features to be used in the ML application were specified by the data science team, while the transaction monitoring analysts, being the end-users of the ML model output, have to use these descriptions for their analysis. This led to overly technical descriptions, that were difficult to understand by the transaction monitoring analysts. In the specification of the features, the end-user component of the sociotechnical system was thus not sufficiently considered, leading to disfunction. The transaction monitoring analysts had to make a translation document retrospectively to make the descriptions and thus the ML model output understandable for the analyst. This could have been prevented, by involving the end-user in the specification of the features and their descriptions.

Comparing this example of misspecification with the email marketing use case sheds light on another approach to deal with the specification of features, to prevent the misspecification described above. In this use case, the ML engineering/consulting firm intentionally chose not to use the most complex model and features. This choice was made to allow for the marketing intelligence analysts to understand the features and to propagate this understanding to the rest of the organisation. The email marketing use case does have another example of misspecification. The ML engineering/consulting firm initially automatically scheduled the model runs and transfers of the output files to the analysis system. However, the marketing intelligence analyst highlighted the importance of the checking role he has. Subsequently, it happened regularly that the transfer had failed, which led to a change in the process, where the marketing intelligence analyst now checks the model output before transferring to the analysis program. This example shows that a specification meant to lead to an efficient process (no human checks) is not necessarily desirable or effective.

Moreover, in both use cases, the systems are specified for internal organisational objectives, not necessarily for fair outcomes for people affected by the systems. In the financial crime detection use case, the objective of one of the models is to reduce the false-positive rate of alerts being investigated. This reduction is seen as valuable for optimizing the alert investigation workflow. However, a false-positive alert means that a customer of the bank is investigated for financial crime, which was not mentioned as a problem. In the email marketing use case, the system is specified to increase the conversion rate towards investing products. This way, the system is optimized to select the most promising customers of the bank to consider investments. This selection might lead to unfair outcomes, such as privileging those who are wealthy already to become even wealthier.

5.1.1.2 General Insights on Misspecification Besides the identification of misspecification within the two use cases, the broader selection of interviews shed light on how misspecification can be present in ML use cases in general. First, several representatives of civil society organisations and regulators pointed out that technology is chosen as solution to problems, whereas it is not always the best solution. As the representative from Privacy First said: "What we see is a kind of love for technology, where goals are achieved with technological solutions, while solutions may need to be found in another area" (Personal Communication, January 22, 2022). Second, every model is a simplification of reality, and it is vital to understand that the world is more complex than the information ultimately present in a ML model, according to the representative of Bits of Freedom (Personal Communication, January 17, 2022). The fact that ML models are simplifications does not always penetrate the sociotechnical system, making ML be seen as a sort of holy grail, while it only captures something very specific, and is not a replacement for critical thinking in an organisation (Personal Communication representative Waag, January 19, 2022). Therefore, it should be debated whether developing a ML application is an appropriate solution to a problem, to prevent misspecification. Furthermore, there are risks of misspecification seen by several several representatives of civil society organisations and regulators in the usage of data. Data is often seen as something factual and objective, while in reality it is a translation of what is seen in the world. Data is just used without really thinking and reflecting on the data and where it comes from, and what social problems have influenced that data (Personal Communication representative Bits of Freedom, January 17, 2022). If data is collected in one context, and later used for another context, this could lead to harmful consequences. As an example a situation was shared in which there were two data sets about hours worked and invoiced by workers, collected by two different organisations, which were subsequently combined into one data set by a third party to seek for fraudulent activity. While in one data set the invoiced hours only represented the on-site working hours, the other data set contained the on-site hours as well as the preparation time. This way a lack of proper specification in the use of the combined data contributed to a ML model that made grave errors unjustly accusing people of fraud (Personal Communication representative Platform Bescherming Burgerrechten, January 18, 2022). Another often mentioned set of consequences of misspecification is when a model is developed and put into production, but not used in practice. The reasons mentioned relate to mis-

specification of the sociotechnical context in which the model needs to operate. For example, the model was not needed, is not trusted by the end-user or adjustments in the way of working of the end-user are not properly foreseen (Personal Communication Data Science Advisor, December 22, 2021).

5.1.2 Machine Error

Machine error can be roughly divided into two categories; machine incorrectness and machine bias, as explained in Sect. 3.2.2. The impact of decisions made based on the models in the two use cases is fairly different. This seems influential to how is dealt with the potential machine error in the models. In the financial crime detection use case, the performance of the models' internal workings are always evaluated before pushing the output to the analysts, whereas in the email marketing use case the model is trusted to be working as expected. Machine error is recognised by the civil society organisations and regulatory bodies as a relevant dimension.

5.1.2.1 Machine Incorrectness in the Use Cases In the financial crime detection use case, the models are used in a quite sensitive context, which makes it important to prevent incorrectness in models. The impact of incorrectness could be that customers of the bank are unjustly investigated by the transaction monitoring analysts, or that customers that should be detected by the model are not detected, which could lead to money laundering or terrorist financing being undetected. Because the latter is considered to be important to prevent, there is an acceptance of a less accurate model, which leads to more false positives, in order to be able to find the true positives. Further, there are multiple mechanism in place to prevent incorrectness in this use case. First, data scientists work with a four-eyes principle, by which every piece of code is checked by another data scientist during development. Second, there is a dedicated independent model validation team that validates a model including all code before it is put to production. Lastly, there is monthly performance monitoring in place for the models, that run monthly, to check whether the model performance is comparable to its performance during training and whether the features' distributions have changed to detect potential machine incorrectness. After completion of the performance monitoring, the model output is directed to the transaction monitoring analysts. Although the bank has these mechanisms in place, the interviewed transaction monitoring analyst pointed out that in a third model, of which the first version is currently live, the testing of the model was not performed in the case management system the analysts are using. Once the first real output of the model in the case management system appeared, there was a lot of incorrectness; for example, generated alerts that did not contain transactions and features that were not visible in a customers' account. As a result, the analysts are dealing with model output that has a lot of incorrectness, and a new version is still not live at the time the interview was conducted (Personal Communication, January 14, 2022).

On the contrary, every interviewee in the email marketing use case pointed out that the impact of machine error is relatively low. The model is not part of a mission-critical activity, and the worst case impact to customers is that customers are accidentally repeatedly being emailed by the bank, or are being emailed

while they had opted-out for certain emails. Although this relatively low impact, the marketing intelligence manager mentions the importance of being transparent. For example, if customers that had opted-out for emails would have been emailed (Personal Communication, January 11, 2022). In the first live version of this model, there appeared to be overfitting on a certain feature. This was only detected once the model was already live, because the test set did not contain the needed exceptional cases. There was only 1.5 years of data so most of the data had to be used for training, but if more data had been available for testing this overfitting could have been detected earlier, before going live (Personal Communication external ML engineer/project manager, January 13, 2022). To prevent emails being sent based on incorrect model output, there is a human in the loop, the marketing intelligence analyst, who does manual checks on the model output to check for example if customers had opted-out. That said, he only checks proposed customers on a narrow set of characteristics, otherwise trusting the model to be working as expected (Personal Communication marketing intelligence analyst, January, 21, 2022).

5.1.2.2 Machine Biases in the Use Cases To prevent biases in the models, in both use cases it was chosen not to use certain data. In the email marketing case, the external ML engineer/project manager pointed out that the bank thinks it is very important to use data wisely, well within the lines of the GDPR (General Data Protection Regulation). Therefore, the choice was made not to use gender data and postal codes (because those may be a proxy for ethnicity). The choices of what could not be included were made based on intuition. In the financial crime detection use case, the data science department wanted to detect potential bias in the ML models caused by proxies for gender, age, ethnicity and origin. However, the privacy officer was prohibited to do this for ethnicity and origin, because those are sensitive personal data. This presents a notable paradox: To be able to detect a bias on a particular attribute, analysts require data on this particular attribute, for example ethnicity. However, the interviewees indicate that the GDPR prohibits this, as a ML model should adhere to *privacy by design*, which means data on ethnicity and ethnicity cannot be processed without a clear and proportional motivation. Interpreting this as a limitation precludes the ability to check for biases.

The handling of potential bias in the ML applications as surfaced by the interviewees in the use cases, i.e. not using certain features and detecting bias for certain factors by proxies, is known not to be a sufficient strategy for attaining fair outcomes. The found practices do not account for the diverse fairness requirements and needs stakeholders involved in or affected by the sociotechnical system (Balayn & Gürses, 2021).

5.1.2.3 General Insights on Machine Error Machine error was recognized as a relevant dimension by the representatives of the civil society organisations and regulators. To illustrate, biases are seen as one of the largest problems in ML applications, in particular in the context of predicting social behaviour or crime. A large risk is the use of indicators that are actually proxies for protected grounds,

as was discussed within the use cases as well (Personal Communication, January 25 representative Amnesty International, 2022). The representative of the AP also points out to guard for unwanted indirect inferences based on proxies, although it is increasingly difficult to completely prevent this from happening in contexts where ever-growing numbers of variables are used (Personal Communication, January 24, 2022). There is a lot of attention on the impact of machine error on people in the proposed EU AI act (Personal Communication representative DNB, January 18, 2022). Society does not have a high acceptance of machine error, which disincentivizes organisations to be open about flaws in their systems. This non-acceptance reflects in a statement made by the representative of Amnesty International, who stated that ML models with high error rates should not be used for any consequential decision-making (Personal Communication, January 25, 2022).

5.1.3 Interpretation

Interpretation of a ML application by human decision-makers is recognised by many civil society organisations as a potential source of risks for the quality of the ultimate decisions. At the same time, human intervention is mandatory if a decision affects a person to a significant degree in the GDPR. In the use cases, stakeholders seem to be less aware of the vulnerabilities of human intervention. In the marketing use case, human intervention is seen as a means to reduce risks by providing an extra checking and controlling mechanism, instead of a step that can impose additional vulnerabilities. There is an increased risk of human noise in the financial crime detection use case, as the more complex system output increases the potential for deviation in interpretations among analysts. Further, the potential for human bias is not adequately considered in both use cases and the human interpretation step is not monitored or evaluated.

5.1.3.1 Interpretation in the Use Cases In the financial crime detection use case, the analysts that use the model output in their work encountered a challenge in the beginning to use the model output. The number of factors that are taken into account using the ML models instead of the previous rule-based system was largely expanded, which made it difficult to understand how to interpret the output, where to look at and what to think about at all (Personal Communication transaction monitoring analyst, January 14, 2022). Meanwhile, the transaction monitoring analyst thinks the model output is well interpretable, in combination with the translation document on the model features. The analysts are supported in the model interpretation by means of explainability methods, consisting of highlighting the three most important features that contributed to the model output, as well as the most important transactions (Personal Communication Lead Data Scientist, January 12, 2022). Further, analysts that start working with the ML models' output follow a training and receive a working instruction document. Nevertheless, the transaction monitoring analyst pointed out that the introduction of the ML models increases the risk of deviations in interpretation among analysts compared to the previous rule-based system, because the output is a lot more complex (Personal Communication, January 14, 2022). The data science advisor within the bank does see that bias could emerge in the decision-making

process besides the model, but that it is difficult to quantify this (Personal Communication, December 22, 2021). The manager data science points out that there are a lot of checks on the model itself, but potential human bias is not well considered, which could be improved. At the same time, the human intervention is an important mitigation measure for automated decision-making, and completely automating the decision-making is not desired for these important decisions (Personal Communication manager data science, January 12, 2022).

In the marketing case, a human-in-the-loop, the marketing intelligence analyst, has a controlling and checking responsibility for the model output, after which the proposed customers by the model are being sent an email. The model output only presents the customer IDs that are proposed, which makes it difficult to understand why the model chooses certain customers (Personal Communication marketing intelligence analyst, January 21, 2022). The marketing intelligence analyst feels that the checking function he has is very important, however, he does not have enough tools and guidance to perform this function adequately. He would like to have better reports on the customers that are selected, to get a picture of the customers (Personal Communication marketing intelligence analyst, January 21, 2022). The ML engineer/project manager does not see a great risk for human bias to emerge, as the marketing intelligence analysts mainly has a checking role (Personal Communication, January 13, 2022). Additionally, the marketeer thinks the role of the marketing intelligence analyst rather decreases the risk on potential mistakes, than imposing new potential mistakes (Personal Communication, January 14, 2022).

5.1.3.2 General Insights on Interpretation Both representatives of Bits of Freedom and Amnesty International mention that a human-in-the-loop is seen by many organisations as a means to take away concerns about ML, while it is not the solution to the problem and much more is needed to prevent harmful outcomes (Personal Communication, January 17 and 25, 2022). Automation bias and limited time for the job can make humans overly rely on the model output, and a model can be used by humans as confirmation to their own bias (Personal Communication representatives Platform Bescherming Burgerrechten and Amnesty International, January 18 and 25, 2022). At the same time, human intervention in automated decision-making that can affect a person to a significant degree is mandatory by the GDPR (Personal Communication AP, January 24, 2022). The ML engineer within bank A points out that the chain of activities in a decision-making process around a ML model is more important for the quality of the ultimate decisions than the model itself (Personal Communication, January 24, 2022). It really depends on the mindset of the human decision-makers that are in the chain between the model output and the final decision. If the decision-making process is badly designed or not thought through, a very good model can lead to terrible outcomes (Personal Communication ML engineer bank A, January 24, 2022).

5.1.4 Performative Behaviour

Performative behaviour as a dimension did not show up frequently in the interview data about the use cases, which could be caused by unawareness of the dimension

or performative behaviour being of less relevance in these use cases. The civil society organisations and regulatory bodies did recognise performative behaviour as a relevant dimension. The representative of DNB pointed out that it is very relevant, yet the point of concern does not get much attention in the field, which could be an explanation of the dimension not being extensively elaborated on in the use cases.

5.1.4.1 Performative Behaviour in the Use Cases Considerations on performative behaviour have not been mentioned in the email marketing use case. In the financial crime detection use case, there was one comment on behaviour. The ML models required changes in the way of working of the analysts, to which they reacted that they could not or did not want to work with the model output (Personal Communication Manager Data Science, January 12, 2022).

5.1.4.2 General Insights on Performative Behaviour The general insights on performative behaviour consist of insights on the change of behaviour among people about whom a ML model takes a decision, as well as change of behaviour among human decision-makers that use the model output to take a final decision. The representative of DNB argued that a trade-off exists between transparency and black-box models. For example, if criminals get insights in how a ML model that is used to detect money laundering works, it becomes easier for them to circumvent being detected as a money launderer. This point of concern gets little attention and does not play a big role in the discussion around explainability (Personal Communication, January 18, 2022). The representative of Amnesty International sheds a different light on this trade-off, arguing that it is not required to make the whole code public, but people should be informed when personal characteristics as nationality, age, postal code or salary are used. The potential for circumventing the system is not at stake, as people cannot change these characteristics, but should be able to know based on what a decision is made (Personal Communication, January 25, 2022). Behavioural change among human decision-makers that use model output is also mentioned by interviewees. The introduction of a ML model decreases the level of ownership a human decision-maker has, compared to a situation without a ML model in place, which can influence the outcome (Personal Communication representative of Waag, January 19, 2022). Lastly, if a human decision-maker has to adhere to a predefined target in terms of validating and rejecting model output, this has influence on the behaviour and the final decisions as such (Personal Communication AP, January 17, 2022).

5.1.5 Adaptation

Adaptation as a dimension can be recognised in the use cases, although it has not been widely discussed in the interviews. In the financial crime detection, the adjustments in the way of working were very challenging, which was not an issue in the email marketing use case. In that use case, the marketing intelligence analyst has actively been given the possibility to adjust configurations towards the environment's needs. Only one representative of a civil society organisation had encountered examples of the adaptation dimension in practice. Whereas it seems that

adaptation is not on top of mind among the external stakeholders, ML applications were developed and not being used in practice due to the adaptation of the designated end-users of ML applications in Bank A.

5.1.5.1 Adaptation in the Use Cases In the financial crime detection use case, the transaction monitoring analyst pointed out that it had been a big challenge to start working with the ML models, as it required a whole new way of working (Personal Communication, January 14, 2022). For the transaction monitoring analysts to understand what the models' output meant in the context of actual increased risk of money laundering or terrorism financing was a struggle in the beginning of using the models in practice. To address the difficulties of using the models' output in practice, there have been intensive feedback sessions between the data science team and the transaction monitoring analysts (Personal Communication transaction monitoring analyst, January 14, 2022).

In the email marketing use case, the marketing intelligence analyst has been given the possibility to adjust some configurations of the model. For example, they can adjust the threshold of the model output, by which a change leads to more or less customers to be emailed based on the model output (Personal Communication external ML engineer/project manager, January 13, 2022). While these possibilities give the marketing intelligence analyst the possibility to adjust the system to the needs of the environment, he has been given instructions not to change configurations too often, because it makes the ML application hard to evaluate (Personal Communication external ML engineer/project manager, January 13, 2022).

5.1.5.2 General Insights on Adaptation As the introduction of ML models often require a change in the way of working among the human decision-makers that are part of the decision-making process, this should be accommodated for in the development and integration of the ML application in its context. In bank A, there have been models that were developed and put into production, but were not used very much in practice, because this accommodation was lacking or there was no trust in the model (Personal Communication data science advisor, December 22, 2022). Another phenomenon within the adaptation dimension is function creep: It happens a lot that a ML application that is designed and developed for one purpose is over time used for other purposes as well (Personal Communication representative Amnesty International, January 25, 2022). Moreover, it can happen that the initial design of the system does not show risks on human right violation, but the way in which the system is used in practice does, for example leading to groups of people being treated differently than intended in design (Personal Communication, January 25, 2022).

5.1.6 Dynamic Change

Dynamic change is not widely recognised by the civil society organisations. Most representatives did not encounter an example of dynamic change in practice, but can image the relevance of it. On the other hand, stakeholders involved in the use cases do recognise this dimension, as external factors such

as regulatory changes or internal changes in data could have large impact on the working of the models over time, while alignment of changes within the bank's different departments involved is a challenge.

5.1.6.1 Dynamic Change in the Use Cases In both use cases, the department in which the models are developed and ran in production are not the owners of the underlying data sources. At the same time, changes in underlying data can have direct impact on the models. In the financial crime detection use case, the Manager Data Science pointed out that if a new version of a ML model is developed, this requires various governance checks, while such checks are not in place for changes in the data sources, owned and triggered by the IT department (Personal Communication, January 12, 2022). Therefore, the data science team needs to continuously engage with the IT department about data changes. Additionally, they monitor potential changes in data or output that can be indicative of problems in the models, monitoring metrics such as feature distributions, alert volume, and false positive rates (Personal Communication Lead Data Scientist, January 12, 2022).

In the email marketing use case, changes in underlying data are seen as a relevant dimension by all stakeholders involved. Changes are not always adequately communicated to the marketing intelligence department, while significant changes could result in the model stop performing properly (Personal Communication Marketing Intelligence Analyst, January 21, 2022). Other developments within the bank, such as updated privacy guidelines or new products, could require adjustments in the model as well. Currently, the department is depending on the external ML engineering/consulting firm to make adjustments in the model (Personal Communication Marketing Intelligence Analyst, January 21, 2022). Lastly, external factors could really impact the model performance. For example, the stock exchange dip in March 2020 due to the start of the Covid-19 pandemic in The Netherlands, was a scenario the model had not been trained on and which led to unreliable outputs that could not be used (Personal Communication external ML engineer/project manager, January 13, 2022).

5.1.6.2 General Insights on Dynamic Change Representatives of Waag, Bits of Freedom, Amnesty International and Platform Bescherming Burgerrechten have not encountered dynamic change as a dimension in practice, but can image that it is a source for vulnerabilities. On the other hand, the representative of DNB does recognise the problems for supervision that dynamic change could cause. The proposal of the new AI act does require a conformity assessment for new high risk applications of ML, but does not take the dynamic change dimension into account (Personal Communication representative DNB, January 18, 2022). Especially in the case of self-learning ML models, an investigation on one day could lead to the outcome that the model is compliant, but the next week it could not be compliant any more (Personal Communication representative DNB, January 18, 2022). This is a realistic problem that is challenging to assess for supervising bodies.

5.1.7 Downstream Impact

Among representatives of civil society organisations and regulatory bodies, downstream impact is seen as an important dimension, in which especially data quality and data selection can have severe impact on the outcomes of ML models. At the same time, the subject receives little attention in practice compared to the attention to the development of the ML models themselves, which has been noticed in the interviews with stakeholders involved in the use cases, in which the concerns of downstream impact seemed to not play a large role.

5.1.7.1 Downstream Impact in Use Cases Within bank A, where the financial crime detection use case was developed, there are measures to ensure good data quality and compliance with the GDPR, as the bank mostly uses gold standard data sources for ML models, which are the most accurate and reliable of its kind (Personal Communication Privacy Officer, January 17, 2022). At the same time, the data sources are maintained and continuously improved or changed by the IT department, which can have downstream impact on the models (Personal Communication Manager Data Science, January 12, 2022). A large challenge is to keep grip on where the ML model outputs are used within the organisation (Personal Communication Privacy Officer, January 17, 2022). To keep a grip on where model outputs are used and thus limit downstream impact, employees or department need either a data sharing agreement or authorization from the data owner, who is the data scientist who developed the model, to use the model output for different purposes (Personal Communication ML engineer, January 11, 2022). It is clear that measures have been taken in bank A to prevent vulnerabilities due to downstream impact. However, the changes in data sources are important to monitor, as well as the requirements for departments to be able to use model output in secondary decision-making processes.

In the email marketing use case, the occurrence of vulnerabilities within the downstream impact dimensions seems limited, as the data quality was good in general, and the output of the ML model is not used in secondary decision-making processes. However, the marketing intelligence analyst did see the possibility for this to become the case in the future (Personal Communication, January 21, 2022). Downstream impact is thus a dimension to keep in mind in bank B.

5.1.7.2 General Insights on Downstream Impact Downstream impact is seen as an important source of vulnerabilities by civil society organisations and regulatory bodies. Both downstream impact by data issues as downstream impact by interconnected ML models have been mentioned. The representative of DNB called data management, governance and quality at least as important as the development of models, whereas it is a relatively small part of the discussion (Personal Communication, January 18, 2022). Representatives of Amnesty International and Bits of Freedom also highlighted that the data used can have severe impact on model output, which receives too little attention (Personal Communication, January 17 and 25, 2022). A ML model can in turn interact with other ML models, which can

lead to losing control on wrong model output or if one model fails, other models fail too, which can have large impact on the larger system (Personal Communication representatives DNB and Waag, January 18 and 19, 2022).

5.1.8 Accountability

Lack of accountability can lead to large issues as outcomes based on ML models can have big impact on people, is a shared point of view from the civil society organisations and regulatory bodies. Within the use cases, there can be vulnerabilities related to accountability identified. In the financial crime detection use case, remaining knowledge on the models within the bank is challenging, which is key to be able to provide accountability. In the email marketing use case, reproducibility of the customers selected and dismissed is not easy to achieve, and the responsibilities are not officially defined among the involved stakeholders.

5.1.8.1 Accountability in Use Cases In the Financial Crime Detection use case, accountability of the model, model output and outcome are divided among different stakeholders. The data science team is model owner and thus responsible for the model. The ML team carries responsibility for correct implementation of the model, and the transaction monitoring analysts are responsible for the decision whether a customer should be reviewed or not. Ultimately, the leader of the financial crime detection department carries ultimate responsibility for everything that happens in the department, and thus for the model, model output and final outcomes (Personal Communication Lead Data Scientist, January 12, 2022). To be able to provide accountability over the ML models used for certain output, reproducibility is in place for the model version, model output, features and optionally parameters, metrics and other metadata (Personal Communication ML engineer, January 24, 2022). Being able to explain an outcome to a customer is a challenge within bank A, because there is a continuous flow of employees leaving and joining the bank (Personal Communication Privacy Officer, January 17, 2022). The risk exists that at a certain point, nobody knows how a ML model works any more. To prevent this, transparency on model development is key (Personal Communication Privacy Officer, January 17, 2022). As the model output is used by the transaction monitoring analysts in the use case, they have to work uniformly to make the final outcome reproducible as well (Personal Communication Transaction Monitoring Analyst, January 14, 2022). It is vital to be able to explain how a model works and how the outcome is achieved to supervisory agents such as the Data Protection Authority (Personal Communication Privacy Officer, January 17, 2022). As the model output is part of a larger decision-making process, making the model output explainable is not enough to being able to understand the final outcome (Personal Communication ML engineer, January 24, 2022).

Being able to explain to a customer why he or she has been selected is seen as a great challenge in the use of ML among stakeholders in the email marketing use case (Personal Communication Manager marketing intelligence and Marketing Intelligence Analyst, January 11 and 21, 2022). If a customer asked this, the bank cannot fully explain why he or she receives the email (Personal Communication

Marketing Intelligence Analyst, January 21, 2022). At the same time, customers could always unsubscribe from receiving certain types of emails (Personal Communication Marketing Intelligence Analyst, January 21, 2022). Also, the model uses only seven features, so the stakeholders within the bank have insight into the data that have led to a model output, which makes the model explainable to a certain degree. The model has been developed by an external ML engineering/consulting firm, but the responsibility for using the model and the model output is carried by the bank (Personal Communication external ML engineer/project manager, January 13, 2022). Within the bank, the responsibilities of the model and model output are not officially defined (Personal Communication marketeer, January 14, 2022). Reproducibility of model output is covered by the external ML engineering/consulting firm. However, reproducibility of the final outcome, thus customers that are selected and customers that are dismissed are not easy to reproduce, because the final selection of the customers is overwritten every week. If needed, the marketing intelligence analyst could match the model output data with the customers that have been emailed to trace back the final selection (Personal Communication Marketing Intelligence Analyst, January 21, 2022).

5.1.8.2 General Insights on Accountability The lack of accountability of ML applications is seen as a big issue among civil society organisation and regulatory bodies. Issues related to accountability can be divided into internal accountability issues and external accountability issues. Internal accountability issues are issues within organisations. In the financial sector, the board level is ultimately accountable for the use of ML models within the organisation, whereas they often do not fully understand what the use of ML entails due to a lack of knowledge and skills (Personal Communication representative DNB, January 18, 2022). There are few experts within and outside of banks that understand how ML models really work. This could lead to board members not being aware of the risks and the impact of ML models on the organisation and the larger financial system (Personal Communication representative Waag, January 19, 2022). External accountability includes situations where people affected by a faulty decision made using a ML application seek justice. Here the lack of explanation over model and broader process, as well as an absence of channels to defend themselves are core issues. These problems are partly caused by the secretive way in which many of such systems are used by organisation, outside of the awareness of subjects. While the GDPR requires organisations to be transparent about the ML models that are used, the AP notices a lack of transparency and proactive communication among such organisation (Personal Communication representative AP, January 24, 2022). When people are not informed about what is going on, it is impossible for them to detect errors in the outcome (Personal Communication representative Platform Bescherming Burgerrechten, January 18, 2022). As this outcome can have large impact on individuals or groups of people, the outcomes should be explainable, and the organizations deploying the systems should be held accountable, which is often not the case in practice (Personal Communication representative Bits of Freedom, January 17, 2022).

5.2 Results of an Inductive Analysis: Challenges in ML Practice

The inductive analysis results in seven challenges. These are discussed in the following paragraphs.

5.2.1 Challenge 1: Defining the System Boundaries for Design, Analysis and Governance

In specification of the ML lifecycle, the system boundaries need to be defined. This defines which components and interactions are taken into account in designing and analyzing the ML model in its context, and determines which stakeholders should be involved. When the system boundary is defined too narrowly at the start, this has consequences for the efficiency and effectiveness of the system development and may be a source for vulnerabilities.

A ML model does not operate in isolation, but becomes part of a larger sociotechnical system. Not including this within the system boundaries may lead to issues later in the ML lifecycle. This is noticed in the financial crime detection use case. The end-users of the ML application, the transaction monitoring analysts, were not involved in the financial crime detection use case from the start. They were not involved until the model had been developed and was ready to be tested and become part of the work of the end-user. For the end-user, the introduction of the ML model required a great shift in their way of working. Moreover, they were not able to interpret the model output, as the feature descriptions they had to use were defined in technical language. Therefore, a translation of the feature descriptions had to be specified ad-hoc and working instructions needed to be created by a delegation of the transaction monitoring analysts.

As illustrated, defining the boundaries too narrow is a source of issues. On the other hand, it is not feasible to involve everyone and specify every detail within the sociotechnical system from the beginning. As such, defining adequate system boundaries represents natural tradeoffs.

5.2.2 Challenge 2: Dealing with Emergence in the Sociotechnical Context

Not every detail within the sociotechnical system can be specified in the beginning. This is especially the case because vulnerabilities may emerge over time, emerge due to interactions between different system components, and emerge due to dynamics within and beyond the sociotechnical context.

In the use case, several efforts are made to deal with emergence over time. For example, the ML models' performance is monitored over time using monitoring metrics as accuracy, prediction volume and feature distributions. Moreover, feedback from the end-users is collected to improve the ML models over time.

However, there are also blind spots identified in the use case on emergent dimensions. First, the behaviour dimension is hardly addressed in the use case. This could for example lead to human decision-makers that overly rely on ML model outputs due to time pressure. Second, the adaptation dimension is insufficiently considered, as the way of working was not part of the specification in the use case. Third, the

feedback gathered from end-users is used to improve the ML applications. However, this does not take into account that such feedback may be subject to human bias or noise, which can be incorporated in technical updates of the model and application.

As illustrated, vulnerabilities may emerge after deployment, when a ML application is used in its sociotechnical context. Awareness of the emergent dimensions is needed, and how to address them remains challenging.

5.2.3 Challenge 3: Understanding the Risks of Human-ML Interactions in Decision-Making Processes

A human intervention in a decision-making process is mandatory if automated decisions affects a person to a significant degree, as described by the GDPR. However, a human-in-the-loop brings about new potential vulnerabilities that can lead to harmful outcomes of a ML application, such as automation bias, confirmation bias, disparate interactions, and limited time for the job.

In the financial crime detection use case, human decision-makers should be able to deliberately use the model output for their final decision. To do so, they have to understand the model output to that extent. Initially, human decision-makers experienced difficulties to understand the model output and use it in their work. Eventually, this is addressed using explainability techniques, to guide the human decision-makers in using the model output. This helps them, but the introduction of the ML application in their work has increased the risk of deviations in interpretation, as the complexity of information has grown.

Furthermore, the introduction of a ML application can impose vulnerabilities in the behaviour and adaptation dimensions. On both dimensions, awareness in lacking in the use cases, as these dimensions did not come forward in the interviews.

Although the introduction of a human decision-maker is often mandatory by the GDPR in a ML decision-making process, it must not be solely seen as a means that takes away vulnerabilities. Rather, it may introduce new ones. As the current focus in the specification of use cases is mainly on the ML model itself, the specification and design of impacts in the decision-making process and its outcomes may receive less priority. As the ML engineer within the bank summarized: “you can have a very good model, but if the decision-making process around it is badly designed, a very good model can lead to terrible outcomes.”

5.2.4 Challenge 4: Recognising the Hazards Related to Data in the ML Lifecycle

As data is at the core of ML applications, it has large influence on the final output of the model and the potential for vulnerabilities to emerge. Biases in data, bad data quality, and changes in data are thus important vulnerabilities to consider.

As the representative of the National Bank of The Netherlands (DNB) pointed out: “data management, governance and quality is at least as important as the developme t of models, whereas it is a relatively small part of the discussion”. This statement is corroborrated in the use cases. In both banks, data management lies in a different department than where the ML applications are developed. The data science department has to follow strict governance processes to be able to launch a new

model or model version. However, the data sources are not covered by these strict governance processes. As such, changes in data can be made, which directly influences the ML models and output. Keeping a grip on this, requires continuous alignment between the departments. This organisational complexity can be a source of vulnerabilities when changes are not communicated within the organisations.

5.2.5 Challenge 5: Developing Knowledge and Shared Language Across Different Actors

Different types of knowledge exist and are needed within the sociotechnical context of ML applications. These have to be developed, shared and maintained. Furthermore, stakeholders with different types of knowledge need a shared vocabulary to effectively communicate in developing, operating and governing such sociotechnical systems.

Firstly, there are few experts on ML within banks and in external organisations. This has consequences that could result in vulnerabilities. For instance, one of the banks does have ML experts in-house, but ML experts often leave the bank and new ML experts are hired. This makes maintaining knowledge on models that run in production a challenge, and may even lead to models not being able to continue due to lack of knowledge of their operations and risks. To address this, extensive documentation on the models is made and updated.

Secondly, the expertise of ML experts and the expertise of end-users on the process and context in which the ML model is used may be hard to connect. In both use cases, the people that ultimately have to work with the ML application output are not ML experts. The other way around, ML experts are no operational experts. As such, to develop a ML application that can be integrated in the way of working of operational experts, communication between the different stakeholders is needed for alignment. However, this communication is often lacking in organisations. To illustrate with the use case, the operational experts were only involved when the ML application development was finished. Furthermore, the model validation team only validated the models, and did not communicate with operational experts to validate whether the models could actually be integrated in the way of working, neither did they assess possible hazards or undesirable impacts in the operating or decision-making process.

5.2.6 Challenge 6: Providing Transparency of ML Models and Process Outcomes

Transparency of ML applications is challenging on multiple levels. First, ML models are known for their opacity, also referred to as a 'black-boxes'. Therefore, it may be challenging or impossible to understand why or how a ML model arrives at a certain output. To address this opacity, explainability techniques are used in the financial crime detection use case. This provides the end-users with some guidance in understanding the model output.

Interviews with civil society organisations, regulators, and other external organisations pointed out that organizations that use ML are often not transparent about it. As a result, civil society organisations struggle to get insight in and address ML applications

that are imposing vulnerabilities that can lead to harm for citizens. Therefore, they only get insight if harm is already imposed, while it is better to prevent harm. Furthermore, the non-transparency of organisations raises the question why they would not be transparent about it, do they have something to hide? As a result, this strengthens the distrust of civil society organisations towards organisations that use ML. Governance approaches and design efforts to use ML responsibly are thus not seen and acknowledged either.

Lastly, there are reasons why organisations could be non-transparent about the use and the inner working of ML applications. Being transparent may hamper their competitive position and the possession of intellectual property. Besides that, if people know how a ML model takes decision, they could game the results, by adapting their behaviour to get a certain outcome. This may impose risks, for example if the ML application is used to detect financial crime.

5.2.7 Challenge 7: Operationalizing Regulations Applicable to ML Applications

The financial sector is highly regulated, but what it means for the use of ML and what future regulations will require from organisations is not yet crystallised.

The enforcement of legislations lies in the hands of several regulatory bodies, among which the Dutch Data Protection Authority (Autoriteit Persoonsgegevens) and the DNB. Furthermore, financial institutions have extensive responsibilities to organise internal supervision and compliance with the GDPR and other legislation. According to the interviews with the Dutch Data Protection Authority, organisations have to be develop a particular level of maturity to be able to do so effectively. A privacy officer in one bank endorsed that the bank has organised its own internal supervision, but that there is little external supervision. This lack of external supervision could lead to undetected risks in the use of ML within the financial sector. Furthermore, the daily supervision of the DNB hardly looks at ML applications at the moment. This will change in the future by the EU AI act, which gives the DNB a mandate to do so.

As illustrated, current legislation and the proposed EU AI act are considered insufficient to enforce a responsible way of developing and using ML applications. Currently, banks are left free to organise this themselves, but the privacy officer of one bank points out that guidance from external higher authorities on how to deal with ML would be very welcome.

It is still too early to understand how the final EU AI Act will look like, and how it will affect organisations. The GDPR caused a substantial increase of awareness on data protection and privacy among organisations, so the question is whether the AI act has the same effect on AI within organisations. How the AI act will be adopted in organisations will in turn influence the position regulatory bodies take.

6 Guidelines for Sociotechnical Specification

In the final stage of the Design Science Research (DSR) project, we developed an artefact to translate the validated dimensions and found challenges to an updated practice for MLOps, extending the initially mapped practice (see Fig. 1). This

practice contains (1) an iterative lifecycle that operationalizes a sociotechnical lens on ML applications across the ML lifecycle, also adding a dedicated sub-practice stage on *sociotechnical specification*, and (2) a set of ten guidelines for organizations to establish this practice.

We did not empirically test or validate the new practice and guidelines, and as such we do not cover these in great detail in this paper. Instead, we provide a list of the guidelines and a depiction of the lifecycle (in Fig. 4). The guidelines do form a pragmatic onramp for addressing vulnerabilities of ML models in their sociotechnical context with a comprehensive set of implications for organizations building, deploying and governing these systems:

- Guideline 1: Establish a multidisciplinary team at the beginning of the ML lifecycle
- Guideline 2: Define the system boundaries as multidisciplinary team
- Guideline 3: Enable the identification, addressing, and mitigation of vulnerabilities in the sociotechnical specification
- Guideline 4: Formulate an initial specification of the sociotechnical system before starting the experimental stage of the sociotechnical ML lifecycle
- Guideline 5: Create feedback channels for different stakeholders during the development and operations of the sociotechnical system
- Guideline 6: Specify monitoring and evaluation mechanisms for the sociotechnical system in operation
- Guideline 7: Verify and validate the sociotechnical system before operationalizing
- Guideline 8: Establish transparency of the sociotechnical system, about its development, design, use and governance

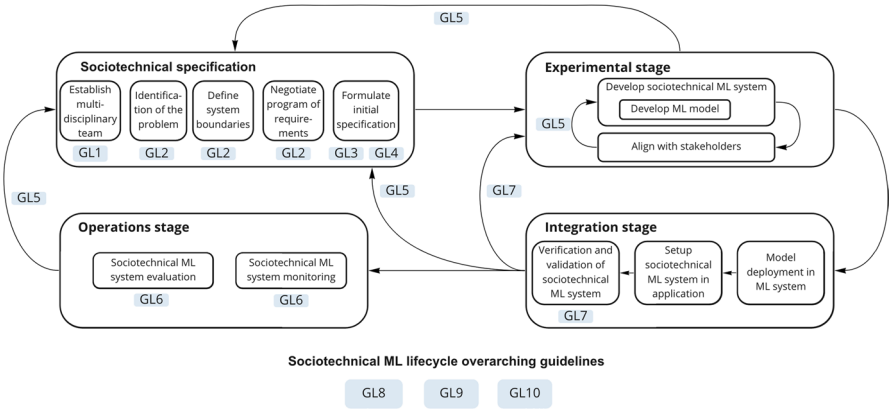

Fig. 4 Visualisation of the sociotechnical ML lifecycle with the developed guidelines. The larger boxes are the different stages of the ML lifecycle. Within each stage, the smaller boxes denote activities. The blue-colored boxed numbered GL1-GL10 denote the activities and points in the lifecycle where the guidelines apply

- Guideline 9: Create knowledge and communication between stakeholders in the sociotechnical system
- Guideline 10: Establish a safe culture and adequate management within the organisation

Interested readers are referred to Wolters (2022) for a more detailed explanation of and reflection on these guidelines.

Figure 4 presents the sociotechnical ML lifecycle, and the guidelines that relate to the different activities. The arrows present activities that follow up on each other and feedback channels that may lead to changes in the deliverables of earlier performed stages. As can be seen, the activities in the sociotechnical specification are not linear, which is why they are not connected with arrows. The only requirement is that an initial specification should be formulated before moving to the experimental stage. Besides that, the monitoring and evaluation of the sociotechnical system in the operations co-exist, thus also does not represent a linear process. Lastly, guideline 8, 9, and 10 are separately presented as they do not address a specific activity, but are guidelines that should be taken into account across the organization and throughout the sociotechnical ML lifecycle.

7 Conclusions and Future Work

This paper provides an empirically-driven conceptualization and validation of vulnerabilities in the sociotechnical context of ML applications. Furthermore, our interview analysis identified a set of challenges that need to be addressed to properly account and design for the context-specific and emergent issues related to sociotechnical complexity. The design science research methodology allowed us to funnel these insights into a set of guidelines to offer a pragmatic framework for practitioners in development, MLOps, as well as those involved in the domain of application or as policy makers or administrators/managers, so to address sociotechnical vulnerabilities in the design, development, use and governance of machine learning in sociotechnical systems context.

Beyond practitioners, the results also contribute to civil society organisations and regulatory bodies as the theoretical framework supports the articulation of vulnerabilities, which are often encountered in the field, but which are still hard to express in a language that is understood by policymakers and ML application developers. In sharing our results with these interviewees we received promising feedback that the vulnerability dimensions and guidelines may help to form a shared lexicon to build more effective bridges between different communities addressing the risks and hazards of ML-based systems.

The research also provides a scientific knowledge base, which can be used in future research on vulnerabilities in sociotechnical systems. Moreover, the empirically validated vulnerability categories provide a useful theoretical lens for research addressing sociotechnical complexity of ML in practice.

Four directions to build upon this research are recommended. First, evaluation and demonstration of the guidelines in real-life ML use cases is recommended. This

research did not involve an iterative step between design and evaluation, but concluded with a first design of guidelines. Evaluation and demonstration are recommended to research the value of using the guidelines in actual ML use cases and to iteratively improve the guidelines. Second, researching other ML use cases within different organisations within the financial sector and in other sectors would enrich the research output. As the financial sector is highly regulated and risk averse, it would be insightful to include use cases from less regulated and more risk seeking sectors, to complement this research. Third, this research widened the technical view that dominates the ML field towards a sociotechnical systems view. The next step is to widen this view further, by addressing the interactions with broader organisational and institutional mechanisms. And lastly, it is relevant to note that this study was performed in late 2021 and early 2022, and hence precedes the quick rise of generative AI tools since the end of 2022. Nonetheless, the findings in this paper largely apply, but may be nuanced or extended for more recent generative AI applications.

Appendix 1: Vulnerabilities in Sociotechnical ML Systems

To enable blind review, we are adding the list of vulnerabilities found in the integrative literature review discussed in Sect. 3.2. This list will be published separately, and may be relevant in the review process.

Vulnerabilities are potential shortcomings in the ML application and its sociotechnical context which may turn into hazards that cause harm to stakeholders. Hazards are defined as "[a sociotechnical] system state or set of conditions that, together with a particular set of worst-case environmental conditions, will lead to an accident (loss)" (Leveson, 2012, p. 184). Hazards are thus states that the broader sociotechnical system should never be in, and that have to be designed out of the system. To identify the vulnerabilities, an integrative literature review has been performed. The selected papers describe one or more vulnerabilities that can emerge for ML applications in their sociotechnical system context.

Selected Literature for Integrative Literature Review

To create a comprehensive overview, literature with a variety of research themes have been consulted. Table 2 presents the selected literature for the integrative literature review that serves as the basis for Sect. 3.2, and the main theme of every paper. The following paragraphs present an overview of the different vulnerabilities that can be present in sociotechnical ML systems.

Choosing ML as a Solution

One should recognize that in not every situation, building a ML model is the solution to a problem. Firstly, if definitions of fairness are politically contested or shifting, it might not be possible to capture the facets of how it changes in the

Table 2 Overview selected literature

Title	Reference	Main theme
AI Governance in the City of Amsterdam: Scrutinising Vulnerabilities of Public Sector AI Systems	Brom (2021)	AI governance
System safety and artificial intelligence	Dobbe (2022)	AI safety
"Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI	Sambasivan et al. (2021)	Data in ML
Explanation in artificial intelligence: Insights from the social sciences	Miller (2019)	Explainability of AI systems
Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead	Rudin (2019)	Explainability of AI systems
Will Algorithms Blind People? The Effect of Explainable AI and Decision-Makers' Experience on AI-supported Decision-Making in Government	Janssen et al. (2020)	Explainability of AI systems
Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making	Veale et al. (2018)	Fairness and Accountability of AI systems
Translation Tutorial: A Shared Lexicon for Research and Practice in Human-Centered Software Systems	Kohli et al. (2018)	Fairness and Accountability of AI systems
Beyond Debiasing: Regulating AI and its inequalities	Balayn and Gürses (2021)	Fairness of AI systems
A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics	Dobbe et al. (2018)	Fairness of AI systems
A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle	Suresh and Guttag (2021)	Fairness of AI systems
Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data	Veale and Binns (2017)	Fairness of AI systems
Fairness and Abstraction in Sociotechnical Systems	Selbst et al. (2019)	Fairness of AI systems
Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making	Kahneman et al. (2016)	Human decision-making
Complacency and bias in human use of automation: An attentional integration	Parasuraman and Manzey (2010)	Human–Machine interaction
Confirmation Bias: A Ubiquitous Phenomenon in Many Guises	Nickerson (1998)	Human–Machine interaction

Table 2 (continued)

Title	Reference	Main theme
Judgemental Forecasts of Time Series Affected by Special Events: Does Providing a Statistical Forecast Improve Accuracy	Goodwin and Fildes (1999)	Human–Machine interaction
Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments	Green and Chen (2019a)	Human-ML system interaction
Machine Learning Informed Decision-Making with Interpreted Model's Outputs: A Field Intervention	Zejnilović et al. (2021)	Human-ML system interaction
The Principles and Limits of Algorithm-in-the-Loop Decision Making	Green and Chen (2020)	Human-ML system interaction
The Problem of Concept Drift: Definitions and Related Work	Tsymbal (2004)	Machine dynamics
Demystifying MLOps and Presenting a Recipe for the Selection of Open-Source Tools	Ruf et al. (2021)	MLOps
Engineering a Safer World	Leveson (2012)	Systems engineering
Engineering problems in machine learning systems	Kuwajima et al. (2020)	Systems engineering

model (Selbst et al., 2019). Secondly, when there is not enough information about the social context, models are as likely to improve as to make the situation worse (Selbst et al., 2019). Therefore, it is essential to study what could happen when implementing the model, rather than implementing it just based on its potential to improve the situation (Selbst et al., 2019). It could also be that the attributes of a system are immeasurable, for example when those involve human psychology. In these situations, implementing a model could be not the right solution to a problem.

Framing of ML Applications

Abstractions are essential to ML, but setting the abstraction boundaries too small, can result in unfair decisions (Selbst et al., 2019). One may define the system boundaries as an *algorithmic frame*, which entails the evaluation of the system on the algorithmic performance, for example the accuracy of the algorithm on training data. Widening the systems boundaries results in the *data frame*, which expands the frame to not just the algorithm but also the inputs and outputs of the final model (Selbst et al., 2019). While the *data frame* facilitates fairness considerations, this frame still eliminates the larger sociotechnical context of the ML application. Contrary, a *sociotechnical frame* does recognize explicitly that the ML application is part of a larger sociotechnical system and includes the decisions made by humans and institutions within the abstraction boundary. This expansion of system boundaries is essential to be able to evaluate fairness of ML in its sociotechnical context, regarding the eventual outcomes of the process in which ML is used. If this is not recognized and the ML application is evaluated as if it is fully autonomous, while in reality it is part of a sociotechnical system with institutional structures and human decision-makers, harmful consequences may follow (Suresh & Guttag, 2021).

Function Creep

Portability is usually pursued in ML, for example by reusing code to train algorithms or to provide "fair" ML applications by defining a definition of fairness that is portable (Selbst et al., 2019). However, portability causes that assumptions made about one sociotechnical context, are also used in other sociotechnical contexts. Since framing the system should entail the sociotechnical context of a ML application, assumptions should be made specific to this sociotechnical contexts. Therefore, a system designed for one sociotechnical context, is not portable between sociotechnical contexts (Selbst et al., 2019). If systems designed for one context are used in other contexts, this is called function creep, referring to a system or technology which function is expanded beyond its original specified purposes (Koops, 2021). Applied to ML applications, an example of function creep could be a ML application that has been specified and developed for one decision-making process, to be eventually being used for other objectives as well.

Mathematical Fairness Definitions Eliminate Societal Nuances

Many efforts have been made in the Fair ML field to mathematically define fairness in order to incorporate fairness ideals into ML (Selbst et al., 2019). However, fairness in society is fundamentally vague, and limiting the notion of fairness to a mathematical formulation is problematic in two ways. First, even if there would be an appropriate mathematical definition of fairness among potential definitions, it is impossible to determine this using purely mathematical means. The social context of the ML application determines what is fair and what is not. To illustrate, in one social context the consequence of a false positive prediction in an automated CV screening would mean a little extra work for the employer because a candidate is interviewed while he should be closed out in the CV screening. In the context of criminal justice, a false positive would mean that a prisoner would be held in prison longer, instead of being released. This illustrates that normative values determine what is fair in which social context (Selbst et al., 2019). The second problem is that fairness is a complex concept, for which no definition might be a valid way to describe it. Fairness may be procedural, contextual, and politically contestable, and mathematical definitions eliminate those nuances (Selbst et al., 2019).

Fail-Safe Mechanisms and Plan B Procedures

Like all technologies, ML applications sometimes fail in their functioning. Therefore, solely relying on constraints at the level of the ML model is insufficient and dangerous (Dobbe, 2022). Instead, hazards have to be identified at the level of operations. This includes acknowledging that the ML model itself can produce errors, but also other emergent failures that relate to the interaction between ML model and the operational or broader organizational and institutional context. Once hazards are identified, it is key to understand what design choices may be available to prevent such hazards from emerging in the first place, or to provide mechnisms that ensure that the system "fails safely" in event of errors or other operational failures (Dobbe, 2022).

Wrong Assumptions About Operations

When a ML application is developed, assumptions about the system in operation are made. However, it can be that the assumptions made are not appropriate (Leveson, 2012). It could also be that the sociotechnical context changes over time, causing initially correct assumptions to become incorrect (Leveson, 2012).

Setting the Threshold for a Positive vs Negative Output

The use of ML models can result in false positive or false negative output (Janssen et al., 2020). False positive output (or Type I mistake) means that the model output indicates that a given condition is present, while it is actually absent. A false negative output (or Type II mistake) means that the model output indicates that a given

condition is absent, while it is actually present (Janssen et al., 2020). ML models can be tweaked to increase or decrease the number of false positives or false negatives. This involves trade-offs from a societal point of view about what is desirable. For example, in the case of searching for a criminal, many false positives mean that innocent people are classified as criminals, but decreasing the number of false positives can in turn increase the number of false negatives, which means that criminals might not be detected as criminals (Janssen et al., 2020). Setting the threshold for a positive versus negative outcome could have different impact for different groups. This could impose fairness issues that were not detected before deployment (Veale & Binns, 2017).

Reproduction of Past Disparities

In supervised ML models, labelled data from previous decision-making is used to train the model (Veale & Binns, 2017). ML models are supposed to discriminate between data points, but some logics of discrimination are not socially acceptable. Thus, if the historical data that the ML model is trained on reflects unwanted discrimination, it is likely that these patterns will also be shown in the model's predictions. Therefore, there is a risk of past disparities to be reproduced using a ML model (Veale & Binns, 2017). For example, the widely-debated COMPAS risk assessment tool wrongly labelled black defendants as future criminals twice as much as white defendants, due to historical discrimination embedded in the training data (Green & Chen, 2019a).

Technical Bias

Besides the aforementioned bias in the historical data, other sources of bias occur when developing and employing ML applications (Dobbe et al., 2018). Technical bias is bias that is caused throughout the development stage of turning data into a model that can make predictions. Choices made in this stage can infer bias, for example when setting the scale of model variables (e.g. ordinal or nominal), choosing the type of model that is used, or when the model is optimized to certain objectives (Dobbe et al., 2018). Making these choices requires justification and often value judgement, which are context-specific and often ethical. Overlooking these questions in practice can have harmful consequences, especially in high-stakes domains.

The Detection of Incorrect Output of Black Box Models

ML models can be black boxes, that do not explain their predictions in a by humans understandable way (Rudin, 2019). These models' learned rules are so complex and non-linear that they are practically inexplicable, even to the model developers themselves, let alone to end-users or policymakers (Zejnilović et al., 2021). The use of black box models in high-stakes domains can have severe consequences, because incorrect decisions based on an inexplicable model are hard to detect.

Explainers can be Inaccurate

Recently, there is a lot of work on explainable ML, where a second (post hoc) model is developed to explain the first black box model (Rudin, 2019). However, since a second model is developed, this model could be inaccurate as well. If an explainable model is correct 90% of the time, one cannot know whether an explanation is correct and whether to trust the explanation or the original model (Rudin, 2019).

Explainers Do Not Consider the Context

Also, current research in ML explainability is mainly focused on technical issues, such as developing measures to explain models and their outputs, and may not adequately consider the variety and complexity of the contexts where the ML applications are deployed (Zejnilović et al., 2021). Explainability tools may not necessarily provide the expected outcome when used in a real-life setting (Zejnilović et al., 2021). Therefore, ML application developers deploying explainers should consider the real-life setting including individual, technical, institutional, and political factors users are coping with. If not, explainers might lead to outcomes that deviate from the expectations (Zejnilović et al., 2021).

An Explanation is not a General Attribute

Moreover, different communities understand explanation in substantially different ways (Kohli et al., 2018). To a ML researcher, an explanation is a description of the operation of the model, which covers the mechanisms used to relate inputs to output (Kohli et al., 2018). In the social science, there is a robust notion of how explanations should behave. Explanations should be causal, explaining why an output was reached or an event occurred; they should be contrastive, explaining why event X happened over event Y; they should be selected, which means they should be based on a few key causes rather than complete descriptions of a mechanism; and they should be social, which means they are meant to transfer knowledge about the system they are explaining (Kohli et al., 2019; Miller, 2018). Therefore, explanations should be contextual, and are not just a presentation of associations and causes. While an output may have many causes, the person that requires an explanation often cares only about a small subset, relevant to the context (Miller, 2019). Therefore, ML models cannot be explained at a general level, but an explanation should be suitable for the stakeholder that requires an explanation.

Inherently Interpretable ML Models are Challenging

Another way to prevent harm by black box models is to develop inherently interpretable models, instead of trying to explain black box models (Rudin, 2019).

Interpretability is domain-specific, so it has no single definition. There is a spectrum of interpretability between fully transparent models where one can understand how all the variables are jointly related to each other, and models constrained in model form, so that it is either useful to someone, or obeys structural knowledge of the domain (Rudin, 2019). For example, models that are forced to increase as one of the variables increases or models that prefer variables that domain experts have identified as important (Rudin, 2019). There is a widespread belief that more complex models are more accurate, so that complicated black box models are necessary for top predictive performance (Rudin, 2019). However, if the data are structured, with good representation in terms of naturally meaningful features, this is often not the case (Rudin, 2019). In these cases, there is often no significant difference in performance between more complex models (e.g. deep neural networks, boosted decision trees, random forests) and much simpler models (e.g. logistic regression and decision lists) after pre-processing (Rudin, 2019). However, there are currently multiple challenges that prevent practitioners to develop inherently interpretable ML models. Firstly, companies make profits from the intellectual property that is created by a black-box model, as they charge per prediction (Rudin, 2019). Secondly, to develop interpretable models, significant effort is needed in terms of computation and domain expertise. Lastly, to uncover 'hidden patterns', which is often called in favour of black-box models, a ML researcher has to be able to both create accurate and interpretable models, which is a difficult optimization challenge (Rudin, 2019).

Deployment Bias

In DSSs, human decision-makers make the final decision, supported by the prediction made by the ML application. Even though the ML model's output could be unbiased according to certain metrics, the human decision that follows could still lead to biased and thus unfair decisions (Balayn & Gürses, 2021). Deployment bias arises when users introduce unexpected behaviour that affects the final decision (Suresh & Guttag, 2021). One example is confirmation bias, which means that human decision-makers seek for evidence or interpret evidence in ways that are in line with their existing beliefs or expectations (Nickerson, 1998). In other cases, people charged with using DSSs ignore or resist the model output (Green & Chen, 2019b). Lastly, the introduction of DSSs can prompt people to alter their behaviour, as they may overly rely on the ML application's output or focus on different goals due to incentives that are created by the introduction of the system (Green & Chen, 2019b). For example, they follow the ML application without considering contradictory information, leading to decisions that are not based on an analysis of all available information, but biased towards the model output (Green & Chen, 2019a; Parasuraman & Manzey, 2010). To incorporate the ML application's predictions, a user interface has to be developed which should be used the human decision-maker (Suresh & Guttag, 2021). Most user interfaces involve simply presenting the model output to a human decision-maker, relying on the person to interpret and incorporate that information (Green & Chen, 2019a). It is challenging to prevent

deployment bias, but designing ML applications that help users balance their faith in model predictions with other information and judgements is an important part. This could involve choosing a model that is human interpretable or developing interfaces that help users to understand model uncertainty and how predictions are to be used (Suresh & Guttag, 2021).

Noise in Professional Judgement

Besides consequences due to human bias, judgements by human decision-makers can undesirably vary from one individual to the next, which is called noise (Kahneman et al., 2016). Noise can even occur in the judgement of one individual from occasion to occasion, for example caused by irrelevant factors as the weather or mood (Kahneman et al., 2016).

Unclear Responsibilities of Decisions

Reliance on ML applications could change people's relationship to the decision-making task, by creating a 'moral buffer' between their decisions and the impact of those decisions (Green & Chen, 2019a). This can lead to human decision-makers to let go of a sense of responsibility and subsequently accountability, because they have the perception the ML application is in charge (Green & Chen, 2019a). Besides that, data scientists often express that they do not bear responsibility for the social impact of their models. Those phenomena can result in situations where both the data scientists developing the ML models and the users of the ML models think the other to be primarily responsible for the outcomes (Green & Chen, 2019a). This scenario should be avoided.

Data Shifts

Data shift are changes in the input data distribution of the model. In the experimental stage, ML models are trained and tested on training and test data. But when the ML model is used in operation, the operational data that is fed to the model as input data tend to change over time (Kuwajima et al., 2020). This phenomenon is called data shifts, which might lead to decreased accuracy and fairness of the ML application over time (Balayn & Gürses, 2021). Data shifts may arise due to several reasons. The populations on which the models are applied might change over time or the way data is captured differs between training and operation, making the data input different from the training data (Balayn & Gürses, 2021). Also, changes in the real-world lead to changes in the data distributions that represent real-world concepts (Tsymbal, 2004). This way, the input data distribution changes, which might lead to model behaviour changing in unwanted ways (Brom, 2021). This would require data shift detection.

Concept Drift

A related vulnerability is concept drift, which are changes in how well the model understands the relationship between input and output. Often, the cause of this change is hidden, and not known upfront, for example change in ways humans think and behave, which leads to changes in what the model is expected to infer (Balayn & Gürses, 2021).This would require concept drift detection, sets of techniques that can be used to automatically detect shifts in distributions potentially relevant to the model's task (Veale et al., 2018).

External Factors Change Over Time

If changes in an organisation or society, for example changes in rules and regulations are established over time, this brings challenges for the development of ML applications (Veale et al., 2018). Awareness of these changes and adequate communication and preparation for them are essential but far from straightforward (Veale et al., 2018).

Runaway Feedback-Loops

Besides pre-existing bias in data and technical bias occurring in the development of ML applications, bias can also arise when the ML application is altered using feedback from its use (Dobbe et al., 2018). For example, in predictive policing, where discovered crime data (e.g. the number of arrests) are used to predict in which areas new crimes will arise, the police surveillance can be intensified in those areas (Dobbe et al., 2018). This way, it is likely that the discovered crime rate increases. If this new discovered crime rate is used for updating the model, the police will be repeatedly sent back to that area, regardless of the true crime rate (Dobbe et al., 2018). This shows that feedback used to update the model can impose emergent bias over time, which is called a runaway feedback-loop.

Lack of Reproducibility

Developing and operating ML applications consists of many steps, often performed by multiple people. If the ultimate decisions that are made based on a ML application are not reproducible, it is not possible to trace back the cause when something goes wrong (Ruf et al., 2021). To reproduce output, training data should be versioned, experiments should be versioned, models should be versioned, it should be tracked which models is used for which prediction, and which input data is used for which prediction (Kohli et al., 2021; Ruf et al., 2018).

Decentralized Data Collection

When models are developed in different parts of an organisation than data collection efforts, it can easily happen that changes in data collection practices occur, without

the people responsible for model performance being aware of those changes (Veale et al., 2018). It can be that data collectors not even know that the data is used for a ML application, particularly when data is collected in a decentralized manner, for example by auditors or police patrol officers (Veale et al., 2018). Better communication between data collectors and ML developers might help, but becomes increasingly difficult if more and more models are developed. Moreover, changes in data collection might be not explicit at all, but can emerge from cultural change or day-to-day choices (Veale et al., 2018). As one can see, it might be impossible to communicate all changes in the data collection. Another approach could be to detect changes in the data distribution itself, by means of concept drift detection (Veale et al., 2018).

Data Cascades

Data cascades are compounding events that cause negative, downstream effects from data issues, that result in technical debt over time (Sambasivan et al., 2021). Data cascades are triggered when ML practices undervalue the importance of data quality, while data largely determines performance, fairness, robustness, safety and scalability of ML applications (Sambasivan et al., 2021). If not enough effort is put in ensuring data quality in the beginning, this will cause negative impacts in the remainder of the ML lifecycle. Data cascades are typically triggered in the beginning of the ML lifecycle, but appear unexpectedly when models are deployed and used in production, resulting in harm to people, discarded models, and redoing data collection (Sambasivan et al., 2021).

Data Availability All interview data is saved in a public repository. This repository is currently not suitable for double-blind review. We assume it is not necessary, but if it is, the authors are willing to share the data in a repository suitable for double-blind review.

Declarations

Conflict of interest This work was supported by the Gravitation Programme Hybrid Intelligence, funded by Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) Financial interests: Anouk Wolters was employed as a research intern and received research support from Deeploy, including access to the use cases. Roel Dobbe declares to have no financial interest in the research project, and has no further relevant financial or non-financial interests to disclose.

Informed Consent For all interview data used in this study, informed consent was received from the interviewees.

Research Involving Human Participants and/or Animals The research plan for the interviews performed as part of this paper study was approved by the Research Ethics board of Delft University of Technology.

References

Ackerman, M. S. (2000). The intellectual challenge of CSCW: The gap between social requirements and technical feasibility. *Human–Computer Interaction, 15*(2–3), 179–203.

AFM (n.d.). Wet ter voorkoming van witwassen en financieren van terrorisme (Wwft) Onderwerpinformatie van de AFM AFM Professionals. Retrieved from https://www.afm.nl/nlnl/professionals/onderwerpen/wwft-wet

Alter, S. (2010). BRIDGING THE CHASM BETWEEN SOCIOTECHNICAL AND TECHNICAL VIEWS OF SYSTEMS IN ORGANIZATIONS. In *International conference of information systems* (Vol. 54, pp. 1–23).

Balayn, A., & Gürses, S. (2021). Beyond debiasing regulating AI and its inequalities (Tech. Rep.). EDRi.

Behymer, K. J., & Flach, J. M. (2016). From autonomous systems to sociotechnical systems: Designing effective collaborations. *She Ji: The Journal of Design, Economics, and Innovation, 2*(2), 105–114.

Brom, D. (2021). AI Governance in the City of Amsterdam: Scrutinising vulnerabilities of public sector AI systems (Tech. Rep.). TU Delft.

de Bruijn, H., & Herder, P. M. (2009). System and actor perspectives on sociotechnical systems. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans, 39*(5), 981–992. https://doi.org/10.1109/TSMCA.2009.2025452

Dekker, S. (2016). *Just culture: Balancing safety and accountability*. CRC Press. https://doi.org/10.4324/9781315251271

Dobbe, R. (2022). System safety and artificial intelligence. In *Oxford handbook on AI governance*. Retrieved from arXiv:2202.09292

Dobbe, R., Dean, S., Gilbert, T., & Kohli, N. (2018). A broader view on bias in automated decision-making: Reflecting on epistemology and dynamics. Retrieved from arXiv:1807.00553v2

Dobbe, R., Gilbert, T. K., & Mintz, Y. (2021). Hard choices in artificial intelligence. *Artificial Intelligence*. https://doi.org/10.1016/j.artint.2021.103555

Ensign, D., Friedler, S.A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2018, February). Runaway feedback loops in predictive policing. In *Conference on fairness, accountability and transparency*, New York.

European Commission. (2021). EUR-Lex-52021PC0206-EN-EUR-Lex. European Union.

Financial Intelligence Unit. (n.d.). Banken, FIU-Nederland. Retrieved from https://www.fiu-nederland.nl/nl/meldergroep/8

Friese, S., Soratto, J., & Pires, D. (2018). Carrying out a computer-aided thematic content analysis with ATLAS.ti. Göttingen. Retrieved from http://www.mmg.mpg.de/workingpapers

Goodwin, P., & Fildes, R. (1999). Judgmental forecasts of time series affected by special events: Does providing a statistical forecast improve accuracy? *Journal of Behavioral Decision Making, 12*, 37–53. https://doi.org/10.1002/(SICI)1099-0771(199903)12:1

Green, B. (2021). The contestation of tech ethics: A sociotechnical approach to technology ethics in practice. *Journal of Social Computing, 2*(3), 209–225.

Green, B., & Chen, Y. (2019a). Disparate interactions: An algorithm-in-theloop analysis of fairness in risk assessments. In *FAT* 2019—proceedings of the 2019 conference on fairness, accountability, and transparency* (pp. 90–99). https://doi.org/10.1145/3287560.3287563

Green, B., & Chen, Y. (2019b). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*. https://doi.org/10.1145/3359152

Green, B., & Chen, Y. (2020). Algorithm-in-the-loop decision making. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 13663–13664). https://doi.org/10.1609/aaai.v34i09.7115

Hevner, A., Chatterjee, S., Hevner, A., & Chatterjee, S. (2010). *Design research in information systems: theory and practice* (pp. 9–22). Springer.

Jabbari, R., bin Ali, N., Petersen, K., & Tanveer, B. (2016). What is DevOps? A systematic mapping study on definitions and practices. In *Proceedings of the scientific workshop proceedings of XP2016* (pp. 1–11).

Janssen, M., Hartog, M., Matheus, R., Yi Ding, A., & Kuk, G. (2020). Will algorithms blind people? The effect of explainable AI and decision-makers' experience on ai-supported decision-making in government. *Social Science Computer Review*. https://doi.org/10.1177/0894439320980118

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399.

Kahneman, D., Rosenfield, A.M., Gandhi, L., & Blaser, T. (2016). Noise: How to overcome the high, hidden cost of inconsistent decision making. Retrieved from https://hbr.org/2016/10/noise

Kohli, N., Barreto, R., & Kroll, J. A. (2018). Translation tutorial: A shared lexicon for research and practice in human-centered software systems. In *1st conference on fairness, accountability, and transparance*.

Koops, B.-J. (2021). The concept of function creep. *Law, Innovation and Technology, 13*(1), 29–56. https://doi.org/10.1080/17579961.2021.1898299

Kuwajima, H., Yasuoka, H., & Nakae, T. (2020). Engineering problems in machine learning systems. *Machine Learning, 109*(5), 1103–1126. https://doi.org/10.1007/s10994-020-05872-w

Leveson, N. (2012). *Engineering a safer world*. MIT.

Loi, M., & Spielkamp, M. (2021). Towards accountability in the use of artificial intelligence for public administrations. In *Proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society* (pp. 757–766). Association for Computing Machinery. Retrieved March 7, 2023, from https://doi.org/10.1145/3461702.3462631

Makarius, E. E., Mukherjee, D., Fox, J. D., & Fox, A. K. (2020). Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research, 120*, 262–273. https://doi.org/10.1016/j.jbusres.2020.07.045

Martin, Jr., D., Prabhakaran, V., Kuhlberg, J., Smart, A., & Isaac, W. S. (2020). Extending the machine learning abstraction boundary: A Complex systems approach to incorporate societal context. arXiv preprint. arXiv:2006.09663.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/J.ARTINT.2018.07.007

Milli, S., Miller, J., Dragan, A. D., & Hardt, M. (2019, January). The social cost of strategic classification. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 230–239). ACM. Retrieved October 6, 2023, from https://doi.org/10.1145/3287560.3287576

Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology, 33*, 659–684.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220.

Nouws, S., Janssen, M., & Dobbe, R. (2022). Dismantling digital cages: Examining design practices for public algorithmic systems. In M. Janssen (Ed.), *Electronic government* (pp. 307–322). Springer. https://doi.org/10.1007/978-3-031-15086-9_20

Offermann, P., Blom, S., Schönherr, M., & Bub, U. (2010). Artifact types in information systems design science —a literature review. In *International conference on design science research in information systems 2010*. LNCS (Vol. 6105, pp. 77–92). Springer. https://doi.org/10.1007/978-3-642-13335-0_6

Oosthuizen, R., & Van't Wout, M. C. (2019). Sociotechnical system perspective on artificial intelligence implementation for a modern intelligence system. In *International command and control research & technology symposium*.

Orr, W., & Davis, J. L. (2020). Attributions of ethical responsibility by artificial intelligence practitioners. *Information, Communication & Society, 23*(5), 719–735.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*(3), 381–410.

Peeters, R., & Widlak, A. (2018). The digital cage: Administrative exclusion through information architecture—the case of the Dutch civil registry's master data management system. *Government Information Quarterly, 35*(2), 175–183. https://doi.org/10.1016/j.giq.2018.02.003

Perdomo, J., Zrnic, T., Mendler-Dünner, C., & Hardt, M. (2020). Performative prediction. In *Proceedings of the 37th international conference on machine learning* (pp. 7599–7609). Retrieved October 6, 2023, from https://proceedings.mlr.press/v119/perdomo20a.html

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 33–44).

Rasmussen, J. (2000). Designing to support adaptation. In *Proceedings of the human factors and ergonomics society annual meeting* (pp. 554–557).

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/S42256-019-0048-X

Ruf, P., Madan, M., Reich, C., & Ould-Abdeslam, D. (2021). Demystifying MLOPS and presenting a recipe for the selection of open-source tools. *Applied Sciences (Switzerland)*. https://doi.org/10.3390/app11198861

Salwei, M. E., & Carayon, P. (2022). A sociotechnical systems framework for the application of artificial intelligence in health care delivery. *Journal of Cognitive Engineering and Decision Making, 16*(4), 194–206.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI (Tech. Rep.). Retrieved from https://doi.org/10.1145/3411764.3445518

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology, 34*, 1057–1084.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 59–68).

Selbst, A. D., Friedler, S. A., Venkatasubramanian, S., Vertesi, J., Boyd, D., & Venkatasubrama, S. (2019). Fairness and abstraction in sociotechnical systems. In *Fat* '19: Proceedings of the conference on fairness, accountability, and transparency* (pp. 59–68). https://doi.org/10.1145/3287560.3287598

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle; A framework for understanding sources of harm throughout the machine learning life cycle. In *ACM conference on equity and access in algorithms, mechanisms, and optimization, EAAMO*. https://doi.org/10.1145/3465416.3483305

Torraco, R. J. (2002). Research methods for theory building in applied disciplines: A comparative analysis. In *Advances in developing human resources* (Vol. 4, pp. 355–376). SAGE.

Treveil, M. (2020). *Introducing MLOps: How to scale machine learning in the enterprise*. O'Reilly Media. Retrieved from http://www.dataiku.com

Tsymbal, A. (2004). *The problem of concept drift: Definitions and related work*. Trinity College.

Van de Poel, I. (2015). Conflicting values in design for values. In *Handbook of ethics, values, and technological design: Sources, theory, values and application domains* (pp. 89–116). Springer.

Van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. *Minds and Machines, 30*(3), 385–409.

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*. https://doi.org/10.1177/2053951717743530

Veale, M., Kleek, M. V., & Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *CHI '18: Proceedings of the 2018 CHI conference on human factors in computing systems*. https://doi.org/10.1145/3173574.3174014

Wagner, B. (2018). Ethics as an escape from regulation. From "ethics-washing" to ethics-shopping? In *Being profiling. Cogitas ergo sum. Legal and political theory in data-driven environments*. Amsterdam University Press.

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kazunias, E., Mathur, V., & Schwartz, O. (2018, December). AI Now Report 2018 (Annual report). New York City: AI Now Institute, New York University.

Winby, S., & Mohrman, S. A. (2018). Digital sociotechnical system design. *The Journal of Applied Behavioral Science, 54*(4), 399–423.

Wolters, A. (2022). Guiding the specification of sociotechnical Machine Learning systems: Addressing vulnerabilities and challenges in Machine Learning practice (Unpublished master's thesis). Massachusetts Institute of Technology.

Zejnilović, L., Lavado, S., Soares, C., Martínez De Rituero de Troya, I., Bell, A., & Ghani, R. (2021). Machine learning informed decision-making with interpreted model's outputs: A field intervention. In *81st Annual meeting of the academy of management 2021: Bringing the manager back in management, AoM 2021*

Zuiderwijk, A., Chen, Y.-C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly, 38*(3), 101577. https://doi.org/10.1016/j.giq.2021.101577