Per Rådberg Nagbøl, Oliver Krancher, Oliver Müller

Challenges and Design Principles for the Evaluation of Productive AI Systems in the Public Sector

Accepted Manuscript

Published in: Charalabidis, Y., Medaglia, R., & van Noordt, C. (Eds.). (2024). Research Handbook on Public Management and Artificial Intelligence. Edward Elgar Publishing.

Challenges and Design Principles for the Evaluation of Productive AI Systems in the Public Sector

Abstract

While research on the development and adoption of AI systems is growing, organizations will harness benefits and avoid harm from AI systems only if AI systems maintain high performance after they are developed and adopted. A key activity in this regard is the evaluation of productive AI systems. In this Action Design Research study, we built, implemented, and evaluated an infrastructure for evaluating productive AI systems at the Danish Business Authority and examined the challenges that such an infrastructure needs to address. We found that key challenges revolve around tedious work, resource availability, maintaining an overview, ensuring sufficient priority, and timing of evaluations. We propose that these challenges can be addressed by a digitized evaluation infrastructure that automatically stops systems not evaluated, by aligning evaluation timing with patterns of change in the real world, by making evaluation work meaningful, and by leveraging synergies between evaluation and other activities. Our study provides unique insights into the challenges of ongoing AI system evaluation in organizational realities, into emergent solution strategies, and their theoretical foundations.

Keywords: Artificial Intelligence, Evaluation, Government, Control, Maintenance, Machine Learning

Per Rådberg Nagbøl[1], Oliver Krancher[1], And Oliver Müller[2]

1 IT University of Copenhagen, Rued Langgaards Vej 7, 2300 Copenhagen, Denmark{pena,olik}@itu.dk

2 Paderborn University, Warburger Str. 100, 33098 Paderborn, Germany oliver.mueller@upb.de

Introduction

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

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

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

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

Although existing work does not explicitly focus on the evaluation of AI systems during maintenance, it offers a critical insight relevant to the design of evaluation systems, namely that effective use of AI requires integrating domain and AI knowledge. For example, a study on the design of pre-production risk assessment emphasizes the importance of a multi-perspective expert assessment involving both AI and domain specialists. Such multi-perspective expert assessments go beyond accuracy metrics and rely on the stakeholders' diverse experience and expertise for assessing AI systems (Nagbøl et al., 2021). An ethnographic study describes the interplay of machine learning (ML) expertise and domain expertise in AI-supported hiring. It finds that developers and domain experts are in an

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

While these studies provide important background knowledge, we know little about how organizations can ensure the effective ongoing evaluation of their AI systems in production. This study aims to address this knowledge gap by exploring the following two research questions: (1) What are the challenges in planning and enforcing the evaluation of productive AI systems? (2) What are the design principles for AI evaluation infrastructure that addresses these challenges?

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

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

Methods

Research Design

The project's methodological approach is Action design research (ADR) which creates generalizable knowledge by solving practical problems through the combination of action and design research (Sein et al., 2011). Key outcomes of ADR are one or more artifacts and design principles. In our case, the artifact is a method for evaluating productive AI systems, which we call Evaluation Plan (see the section on the design artifact below for more details). ADR proceeds along the four stages of (1) problem formulation, (2) building, intervention and evaluation (BIE), (3) reflection and learning, and (4) formalization of learning.

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

[1] The DBA was nominated for the Danish digitization price sammenhængsprisen for the public sector (2022) for their AI supported work with the covid-19 compensation /https://www.digitaliseringsprisen.dk/

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

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

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

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

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

Empirical work

The first author of this article has been working with the DBA since September 2017, initially as an external consultant and from august 2018 as a collaborative Ph.D. fellow, spending about half of his time in the Machine Learning Lab at the DBA where he took part in everyday work-life activities. He kept a field diary with notes from observations in the organization and meetings and conversations with colleagues and consultants. These were supplemented with insights from reading and writing emails and documentation on platforms such as Git, Teams, Outlook, Jira, and Confluence.

Design Artifact

The design artifact, the Evaluation Plan, is part of a broader framework for responsible AI use named X-RAI (Nagbøl & Müller, 2020). One element of X-RAI is the Artificial Intelligence Risk Assessment (AIRA) tool (Nagbøl et al., 2021), which supports AI risk assessment before production, and thus creates the foundation for post-production evaluation and the retraining of AI systems. The Evaluation Plan inherits, due to its supplementary

nature, the theory ingrained into AIRA, including principles such as multi-expert assessment and structured intuition (i.e., providing some structure while leaving experts room for their judgment) (Nagbøl et al., 2021). In line with the principle of structured intuition, the Evaluation Plan is a questionnaire that provides some structure while leaving room for expert judgment. Table 1 shows the Evaluation Plan implemented at the DBA during iteration 3 (see below for a description of iterations).

Table 1: Evaluation Plan Artifact

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

BIE Iterations

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

Iteration 1: Usability and Content

We evaluated the Evaluation Plan following the ADR principles of authentic and concurrent evaluation. The Evaluation Plan was introduced to organizational work practices in a word format. In this phase, the evaluation focused on the understandability of the questions and on their suitability for estimating the resource needed. The evaluations led to minor changes to the artifact. Afterward, the artifact was transformed into YAML format, which facilitates integration into an IT infrastructure.

Iteration 2: Behavioral Impact

The second evaluation iteration focused on evaluating the extent to which the Evaluation Plan fulfilled its expected behavioral impact, i.e., the impact of securing and structuring the evaluation of AI systems in the DBA. To this end, we gathered the compiled Evaluation Plans and other relevant documentation, such as Evaluation Schemas. We then analyzed the compiled Evaluation Plan frameworks. The analysis revealed that 16 AI systems had

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

Iteration 3: Challenges

The third evaluation iteration focused on discovering overlooked evaluation practices and gaining a deeper insight into the circumstances impacting evaluation. Therefore, we conducted seven semi-structured interviews in January and February 2022 with stakeholders named in the Evaluation Plans. The stakeholders held diverse positions related to IT development, the ML Lab, and different departments using AI systems. Interview durations varied from 44 to 80 minutes. The interviews were structured around the following themes: introduction questions and background, AI systems purpose and use, quality assurance, evaluation, accountability, risk, challenges, and trust. The interviews were transcribed and coded in Nvivo. In coding, we followed an inductive process where we aggregated lower-level challenges and design principles into a few higher-order categories, similar to data analysis approaches in case study research and grounded theory research (Charmaz, 2006; Yin, 2009). The design principles are planned to be implemented in a subsequent, digitized version of the Evaluation Plan.

Findings: Challenges and Solutions

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

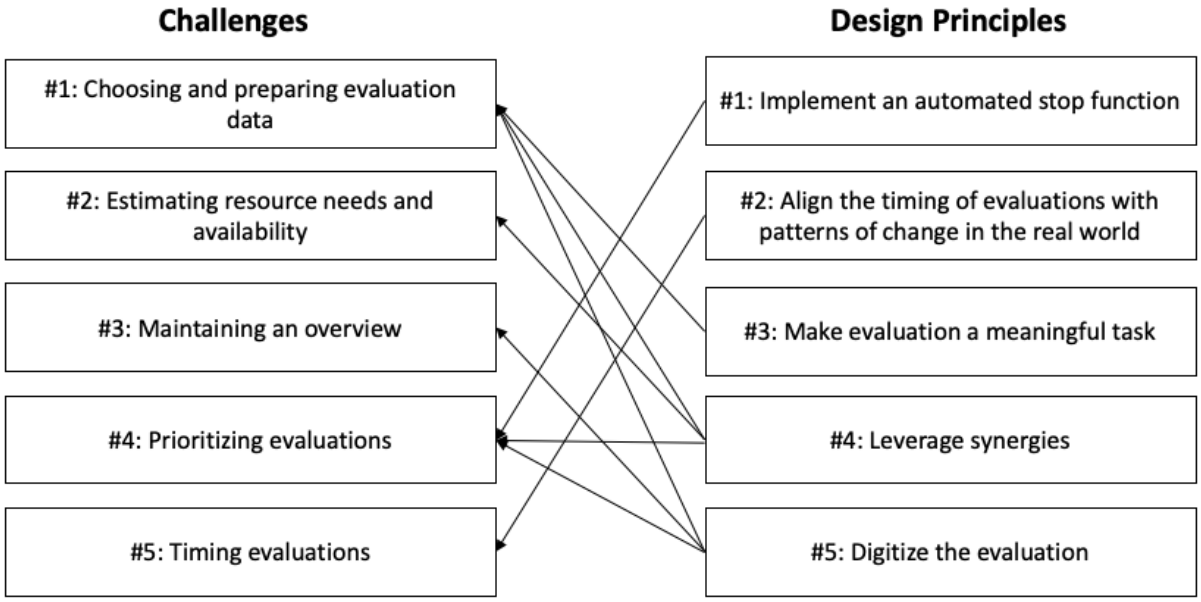

Figure 1: Challenges of and Design Principles for AI Systems Evaluation

Challenges

#1: Choosing and preparing evaluation data

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

Our participants perceived annotating data for supervised ML as a rather tedious, resource-intensive task. Much like during the initial training of an AI system, post-production

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

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

#2: Estimating resource needs and availability

Our informants reported that it was often difficult to anticipate the resources available and needed for evaluation because the use of AI fundamentally changed business processes and priorities. For instance, Torben mentioned that the work in his team increased rather than decreased after AI was introduced. AI provided the opportunity of more effectively detecting fraudulent applications early in the process, helping employees "to take out the right ones" early in the process rather than to "waste a lot of businesses' time" or to perform manual checks later. Because of this more effective, AI-enabled checking process, the DBA found it economically advantageous to increase the team size from 4 to 18 employees, who were now responsible for following up on the alarms triggered by the system. These and other

fundamental changes in the business process and the role of human labor in the business process would have made it very difficult, if not impossible, for the DBA to anticipate what amount of resources would be available for evaluation and not bound by the new work activities that are enabled by the AI system.

#3: Maintaining an overview

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

#4: Prioritizing Evaluations

The Evaluation Plans were initially followed until the start of the Covid-19 pandemic. The pandemic caused an exceptional situation at the DBA where enormous resources were needed

to rapidly develop systems, such as systems for administering the compensation packages the Danish government granted to businesses suffering from lockdowns. This exceptional situation made it difficult for the DBA to allocate resources to evaluate existing AI systems.

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

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

#5: Timing evaluations

There was substantial uncertainty about when to evaluate. Initially, a rule of thumb was that the first evaluation should occur 14 days after go-live and that subsequent evaluations should be performed every third month. Our informants agreed that deciding on the time and scope for the first evaluation was difficult. For example, are child diseases and early implementation issues something to include, or should the first evaluation only touch on matters aligned with the subsequent evaluations? Some informants argued that 14 days was

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

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

Design principles

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

Table 2: Design Principles

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

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

#1: Implement an automated stop function

A key challenge, especially during the Covid-19 pandemic, was ensuring evaluations receive sufficient priority. To address this challenge, the DBA chose to implement an automated emergency stop function in its Intelligent Control Platform, i.e., the infrastructure developed to digitize the Evaluation Plan. The automated stop function is a feature that ensures a productive AI system stops running if it is not evaluated as per the evaluation plan. Such a function is similar to an automatic train stop system, which stops a train automatically if the train conductor fails to push a button regularly. As one informant told us: "[The head of the department] is going towards setting something up in the Intelligent Control Platform so that we switch off a model if it is not confirmed that the model has been evaluated" (Ida).

#2: Align the timing of evaluations with patterns of change in the real world

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

While the timing of the first evaluation often depended on the question of when enough data is available, the timing of subsequent evaluations followed one of the following logics: frequency-based, event-based, seasonal, and autonomous-driven. *Frequency-driven evaluation* implies that evaluations are conducted as a reoccurring event after a fixed timespan, for example, every third month. Our informants pointed to several considerations to be made when deciding on the frequency. One question is how long it is tolerable to run on a false premise, given that the pattern in behavior could, in principle, change the day after the evaluation. A question to ask when planning the evaluation is "for how long time can we tolerate that it answers incorrectly [and for how long are we willing]…to live with not discovering that there suddenly is something that we do not catch although we think we catch right..." (Ida). Our informants mentioned the impact of the AI system, the thoroughness of prior evaluations, and the amount of dynamism as further factors that affect the needed frequency: "How big an impact it has but also how the earlier assessments have looked like. If we had the first evaluation rather quickly and then held another one after three months, and everything looks fine, and this is not an area where something is going to happen, and it is probably business as usual, then there is no reason we should meet again in three months, and then we can set it up to be biannual" (Torben).

Event-driven evaluation is based on events that impact the AI system's performance or change the context of the AI systems so that the predictions are no longer suitable. Examples of such events include changes in technical standards and industrial classification codes: "We also have XBRL with taxonomies that are changing all the time" (Daniel), "… revision of industrial classification code, yes, we know that would happen in 2024, I think, maybe 2025… (Oscar).

Other AI systems have a *seasonal* flow, with activity fluctuating depending on the time of the year or other recurrences. It is essential to consider when planning the evaluation:" I will say

that there should be more meetings the closer we get to the big filling period occurring from around the end of April until the end of June…" (Daniel)

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

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

#3: Make evaluation a meaningful task

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

an opportunity for competence development by emphasizing that the evaluator will obtain a first-hand feeling of how the AI system performs on negative and positive classifications. Evaluation can also be framed as an opportunity for autonomy by communicating that the evaluator plays the role of an educator of the AI system by ingraining their expert knowledge or by including the management of running AI systems into individuals' job descriptions. Moreover, evaluation can be seen as an opportunity for relatedness by involving multiple evaluators, which may provide opportunities for knowledge sharing and learning from each other. While these strategies may help make evaluation work more enjoyable, other strategies focus on communicating the benefits and rewards of evaluation. For example, it may be helpful to communicate why the evaluation is essential, how it benefits the quality of everyday work, and what consequences can occur without evaluation.

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

While it was difficult to ensure sufficient, motivated resources were available (as captured through challenges #1, #2, and #4), our informants shared that several strategies help leverage synergies between evaluation and other activities, which may help relax resource issues and reduce tedious elements of evaluation work. Specifically, our informants recommended leveraging synergies between regular work, AI system evaluation, human training, and AI system training. Human oversight was often a mandatory quality insurance and validation mechanism for at least one of the classification categories (typically for negative classifications, e.g., for rejected applications) and, in some AI systems, both for negative and positive classifications. The informants considered quality insurance critical. Indeed, there were different kinds of ongoing quality insurance and evaluation mechanisms of the AI systems despite the lack of use of X-RAI's evaluation framework. "In the audit unit, we have weekly meetings about the model and our experiences with the model both on the caseworker level and with our boss ... We conduct these meetings, among other things, to collect and

deliver feedback back to ML (ML lab).” (Kim) Hence, meetings that aimed at reflecting on and improving AI-enabled work practices provided important potential input for evaluations.

Our informants also pointed us to potential synergies between evaluation and human training. Indeed, the formalized and standardized evaluation flow allowed acquiring, storing, and sharing of knowledge and experience from the evaluation of the AI system, contributing to continuous individual and organizational learning and the development of best practices, including utilizing the experiences already ingrained in the evaluation schema. For example, evaluation activities can be included into individual competence development processes: “We have just hired a new employee in the team who needs training. We always focus them on [the AI system] because there are some good cases to get out about of training” (Torben). The data annotation is an opportunity to work dedicated to one specific interpretation of, for example, a law repeatedly, thus stipulating learning. It is then relevant when working through the cases to annotate the data.

Another synergy is to store and declare annotated evaluation data so that it can be recycled as training data when retraining the AI system. Hence, a retraining procedure starts with deciding which evaluation data to recycle. Integration of evaluation into the regular workflow is an option “… the best in the world would be that our case management is constructed in a way if I, for example, had processed a case in the signature [AI system], then I could, while closing my handling of the case, do some evaluations of the positives” (Torben).

#5: Digitize the evaluation

Among the most important challenges related to evaluation were the difficulties of maintaining an overview of and prioritizing evaluation activities. The DBA reacted to these difficulties by introducing the Intelligent Control Platform, a digitized infrastructure for managing AI systems evaluation. As control theory suggests (Eisenhardt, 1985), controlees (e.g., evaluators) are more likely to show the expected behaviors (e.g., evaluating AI systems

when needed) if the information about the controlees' behaviors is transparent. Hence, a digitized evaluation infrastructure can be an important element for not only maintaining an overview of evaluation activities but also for enforcing that evaluations are conducted as required. Interestingly, the platform also helped make evaluation and retraining more straightforward (see the link between challenge #1 and #5: Digitize the evaluation in Figure 1) because it allowed the evaluator to annotate relevant data in a system that automatically stored it and made it accessible for retraining purposes, which further helped cope with resource bottlenecks. As Daniel put it: "We want that four eyes have a look at a case… We should have created … a GUI [Graphical User Interface] …, where we could annotate: I think this [company] here is fictitious [based on the AI system prediction].Then your colleague could go into the tool and say: This is correct … and starts a case." (Daniel). The DBA started building an infrastructure for evaluating and retraining AI systems. That infrastructure should allow for easier evaluation and faster adaptation of AI systems to changes in their environment. The expectation was that the new platform will allow for better monitoring, evaluation, and retraining: "It makes it easier for us to evaluate the possibility because we are sitting with it closely now. It becomes easier to update the model when we discover that the model starts to perform worse, making it easier to put a new model in production. It is easier to retrain the models because everything is in one place in our repository" (Oscar).

Discussion

This chapter was motivated by the observation that little work has examined the evaluation of productive AI systems in organizational realities, even though evaluating productive AI systems is critical for avoiding harm and ensuring benefits from AI systems. Against this background, we explored the challenges organizations face in planning and enforcing the evaluation of productive AI systems and design principles for an AI evaluation infrastructure that helps address these challenges. We relied on Action Design Research to answer these

research questions. We built, implemented, and evaluated our design artifact, the Evaluation Plan in the Danish Business Authority (DBA), which included conducting seven semi-structured interviews. The DBA, with its high amount of productive AI systems, was a unique environment for studying changes in the ongoing evaluation of productive AI systemns and design principles that help address these challenges.

Implications

Challenge #1: Choosing and preparing evaluation data relates to prior research emphasizing the importance and difficulties of preparing labeled data for training or evaluation (Lebovitz et al., 2021). While existing work highlights the potentially problematic nature of ground truth and the difficulties that uncertain ground truth creates for labeling (Lebovitz et al., 2021), our findings point to another challenging facet of labeling work: its tedious, effortful nature. We not only reveal this as a factor that challenges the evaluation of AI systems; we also propose design principles for evaluation infrastructure that can help address this challenge. Specifically, our data suggest that the tedious nature of evaluation work can be mitigated by a digitized evaluation infrastructure that enables evaluation planners to make evaluation more meaningful and leverage synergies between evaluation and human and AI training. These insights align with but also go beyond prior research that emphasizes the tightly connected nature of human and AI learning in AI implementations (Lebovitz et al., 2021; Lou & Wu, 2021; Nagbøl et al., 2021; van den Broek et al., 2021).

Challenge #2: Estimating resource needs and availability highlights the difficulties of planning and ensuring the availability of resources for evaluation. This finding echoes socio-technical research showing that system implementations often yield many planned and unplanned changes (Lehrig & Krancher, 2018; Orlikowski, 1996; Robey et al., 2002). It may, hence, be of little surprise that work systems after implementing an AI system may look different from what the designers of the systems had anticipated, which also implies that it is

difficult to plan the resources required for evaluation. Challenge #2 also resonates with project management research that generally emphasizes the difficulties of ensuring domain expert availability during IS projects (e.g., Wallace et al., 2004). Going beyond these findings, however, our study highlights that, for AI systems, the availability of domain experts is a key challenge not only during development but also during the maintenance of AI systems, given the continuous need for domain experts to ascertain that the systems work as intended.

Challenge #3: Maintaining an overview and *Challenge #4: Prioritizing evaluations* point to two further difficulties in organizational realities after the go-live of AI systems. Given the relatively scant focus on maintenance (Luccioni et al., 2022) and on the social context (Selbst et al., 2019) in most existing AI research, we are not aware of prior work that has pointed to these two challenges, which arise as the number of live AI systems in organizations is increasing. Nonetheless, our paper suggests that existing organizational theories may be well suited for explaining the organizational structures required to ensure coordination of and focus on evaluations. Control theory (Eisenhardt, 1985; Kirsch, 1996; Krancher et al., 2022) can help explain the information systems required to ensure enough focus on evaluation. Organizational research on the role of slack in improvement activities (e.g., Repenning & Sterman, 2002) can help explain why slack resources (i.e., domain experts having capacity beyond their daily tasks) can be an essential foundation for evaluations.

Last but not least, *Challenge #5: Timing evaluations* and the design principles of aligning the timing of evaluations with patterns in the real world raises the timing of evaluation as an important issue that yet awaits to be addressed by AI research. We propose that representation theory (Recker et al., 2019) can be a fruitful theoretical basis for future research on the issue.

Practical contribution

Our paper and the collaborative development of X-RAI with the DBA offer several practical implications for organizations setting up or needing to improve structures to evaluate productive AI systems. First, they should consider implementing automated stop function that switch off critical AI systems if they are not evaluated. Second, they should consider the pattern of change in the application domain of a system and align the timing of evaluations with that pattern. Third, they design work arrangements that make evaluations a meaningful task despite its potentially tedious nature. Fourth, they should leverage synergies between regular work, AI system evaluation, human training, and AI system training. Five, they should consider building digital infrastructures for AI system evaluations, which help provide an overview of evaluations and make use of the data generated during evaluation and during human-in-the-loop activities.

The use of the X-RAI artifact (Nagbøl & Müller, 2020; Nagbøl et al., 2021) with its four elements Artificial Intelligence Risk Assessment, Evaluation Plan, Evaluation, and Retraining frameworks has contributed to the recognition of the DBA as an industry leader in AI Ethics. The Joint Research Center under the European Commission has found the DBA´s Ethical Data Governance to be the most advanced among public sector organizations within the European Union (Tangi et al., 2022). In addition, the consultancy firm Gartner has suggested that AI Leaders adopt the DBA approach to ensure an ethically defensible development and use of AI Systems (Gartner, 2021). The importance of this topic for the industry is evident by Gartner, including it in their Top 10 Strategic Technology Trends for 2023 as AI Trust, Risk and Security Management (AI TRiSM). Furthermore, Gartner highlights the DBA´s methodological framework as an example of how to tie ethical principles to concrete actions (Gartner, 2022).

Limitations and Boundary Conditions

In this ADR study, we focused on challenges and design principles that help address them, as our work with the artifact and interview data indicated. Implementing an artifact that fully embodies these design principles remains future work. As such, it is well possible that implementing an artifact based on the design principles proposed here may produce unanticipated effects and call for further changes to the artifact. Similarly, it is worth highlighting that our artifact, the Evaluation Plan, was subject to authentic, concurrent evaluation by examining its use in action at the DBA and deriving design principles from these observations. A rigorous summative evaluation of the effect of the Evaluation Plan on outcomes such as decision accuracy or harm is beyond the possibilities of this study. It is also important to point out that this study solely focused on evaluation based on the X-RAI framework as an artifact accordingly to the design principles of ADR. The quality insurance mechanism and evaluation of governmental conduct and work in the DBA were beyond the scope of this study. Pointing toward a lack of formal evaluation of a given AI system must not be interpreted as a lack of quality insurance.

Conclusion and Future Research

While most existing research on AI focuses on technical issues or on socio-technical aspects of the development and adoption of AI, our study reveals challenges that arise post-implementation when organizations rely on a multitude of AI systems and make efforts to ensure the proper functioning of these systems through evaluation. The challenges and design principles for addressing the challenges open up many avenues for future research. For example, researchers interested in the future of work and humanistic aspects of work could explore how individuals cope with the tedious nature of evaluation work and how organizations can further address these challenges beyond the strategies unveiled in this paper. Researchers interested in organizational structures and control could explore portfolios

of control mechanisms that help ensure sufficient evaluation. Researchers interested in dynamics and formal systems could expand on the evaluation timing strategies uncovered in our study and examine the effectiveness and efficiency of these strategies. Design researchers could focus on infrastructure that integrates an organization's AI systems and organizes the data required for evaluation in an efficient, user-friendly way. Further research could focus on conducting and optimizing the post-production performance evaluations to secure continued fulfillment of the business objectives, here among how to performance measure AI systems where the business value differentiates within the positive or negative categories and how to continuously monitor business objectives going beyond the accuracy of the AI system such as systematically checking and mitigating for bias and other harmful outcomes.

References

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

Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T., & Salovaara, A. (2021). Socio-technical Envelopment of Artificial Intelligence: An Approach to Organizational Deployment of Inscrutable Artificial Intelligence Systems. *Journal of the Association for Information Systems*, *22*(2), 325–352. https://doi.org/10.17705/1jais.00664

Baird, A., & Maruping, L. M. (2021). The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts. *MIS Quarterly*, *45*(1).

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Q*, *45*(3), 1433–1450.

Charmaz, K. (2006). *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*. Sage.

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *ArXiv:1702.08608 [Cs, Stat]*. http://arxiv.org/abs/1702.08608

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

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

Gartner. (2021). *Case Study: How to Apply Ethical Principles to AI Models (Danish Business Authority)* (Case Study No. G00749866). Gartner. https://www.gartner.com/en/documents/4004387

Gartner. (2022). *Top Strategic Technology Trends 2023*. https://www.gartner.com/en/information-technology/insights/top-technology-trends

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

Hernández-Orallo, J. (2017). Evaluation in artificial intelligence: From task-oriented to ability-oriented measurement. *Artificial Intelligence Review*, *48*(3), 397–447. https://doi.org/10.1007/s10462-016-9505-7

Hill, K. (2020). Wrongfully accused by an algorithm. In *Ethics of Data and Analytics* (pp. 138–142). Auerbach Publications.

Jussupow, E., Spohrer, K., Heinzl, A., & Gawlitza, J. (2021). Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. *Information Systems Research*, *32*(3), 713-735.

Kirsch, L. J. (1996). The management of complex tasks in organizations: Controlling the systems development process. *Organization Science, 7*(1), 1–21.

Kirsch, L. J. (2004). Deploying common systems globally: The dynamics of control. *Information Systems Research, 15*(4), 374–395.

Krancher, O., Oshri, I., Kotlarsky, J., & Dibbern, J. (2022). Bilateral, collective, or both? Formal governance and performance in multisourcing. *Journal of the Association for Information Systems, 23*(5), 1211–1234.

Lebovitz, S., Levina, N., & Lifshitz-Assaf, H. (2021). Is AI ground truth really "true"? The dangers of training and evaluating AI tools based on experts' know-what. *Management Information Systems Quarterly*.

Lehrig, T., & Krancher, O. (2018). *Change of Organizational Routines under Malleable Information Technology: Explaining Variations in Momentum*. Proceedings of the Thirty Ninth International Conference of Information System.

Li, J., Li, M., Wang, X., & Thatcher, J. B. (2021). STRATEGIC DIRECTIONS FOR AI: THE ROLE OF CIOS AND BOARDS OF DIRECTORS. *MIS Quarterly, 45*(3).

Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue, 16*(3), 31–57.

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

Luccioni, A. S., Corry, F., Sridharan, H., Ananny, M., Schultz, J., & Crawford, K. (2022). A Framework for Deprecating Datasets: Standardizing Documentation, Identification, and Communication. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 199–212. https://doi.org/10.1145/3531146.3533086

Mayer, A.-S., Strich, F., & Fiedler, M. (2020). Unintended Consequences of Introducing AI Systems for Decision Making. *MIS Quarterly Executive*, *19*(4).

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

Nagbøl, P. R., Müller, O., & Krancher, O. (2021). Designing a Risk Assessment Tool for Artificial Intelligence Systems. In L. Chandra Kruse, S. Seidel, & G. I. Hausvik (Eds.), *The Next Wave of Socio-technical Design* (pp. 328–339). Springer International Publishing.

Nations, U. (2018). *United Nations E-Government Survey 2018*. United Nations. https://www.un-ilibrary.org/content/books/9789210472272

Orlikowski, W. J. (1996). Improvising organizational transformation over time: A situated change perspective. *Information Systems Research*, *7*(1), 63–92.

Recker, J., Indulska, M., Green, P., Burton-Jones, A., & Weber, R. (2019). Information Systems as Representations: A Review of the Theory and Evidence. *Journal of the Association for Information Systems*, *20*(6), 5.

Repenning, N. P., & Sterman, J. D. (2002). Capability traps and self-confirming attribution errors in the dynamics of process improvement. *Administrative Science Quarterly*, *47*(2), 265-295.

Robey, D., Ross, J., & Boudreau, M.-C. (2002). Learning to Implement Enterprise Systems: An Exploratory Study of the Dialectics of Change. *Journal of Management Information Systems*, *19*(1), 17–46.

Russell, S., & Norvig, P. (2002). *Artificial intelligence: A modern approach*.

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

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

Sein, M., Henfridsson, O., Purao, S., Rossi, M., & Lindgren, R. (2011). Action Design Research. *Management Information Systems Quarterly*, *35*(1), 37–56.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. https://doi.org/10.1145/3287560.3287598

Strich, F., Mayer, A.-S., & Fiedler, M. (2021). What do I do in a world of artificial intelligence? Investigating the impact of substitutive decision-making AI systems on employees' professional role identity. *Journal of the Association for Information Systems*, *22*(2), 9.

Sun, T. Q., & Medaglia, R. (2019). Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly*, *36*(2), 368–383.

Tangi, L., Van Noordt, C., Combetto, M., Gattwinkel, D., & Pignatelli, F. (2022). *AI Watch. European landscape on the use of Artificial Intelligence by the Public Sector* (ISBN 978-92-76-53058-9). Publications Office of the European Union. doi:10.2760/39336

Teodorescu, M. H., Morse, L., Awwad, Y., & Kane, G. C. (2021). FAILURES OF FAIRNESS IN AUTOMATION REQUIRE A DEEPER UNDERSTANDING OF HUMAN-ML AUGMENTATION. *MIS Quarterly*, *45*(3).

United Nations. & Department of Economic and Social Affairs. (2020). *United Nations e-government survey 2020: Digital government in the decade of action for sustainable development*. United Nations, Department of Economic and Social Affairs.

Van den Broek, E., Sergeeva, A., & Huysman, M. (2021). When the Machine Meets the Expert: An Ethnography of Developing AI for Hiring. *MIS Quarterly*, 45(3).

Wallace, L., Keil, M., & Rai, A. (2004). How software project risk affects project performance: An investigation of the dimensions of risk and an exploratory model. *Decision Sciences*, *35*(2), 289–321.

Yin, R. K. (2009). *Case study research: Design and methods*. Sage.