# Designing a Risk Assessment Tool for Artificial Intelligence Systems

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

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

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

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

# 1 Introduction

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

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

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

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

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

# 2 Literature Background

## 2.1 The AI Literature

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

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

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

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

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

### 2.2 Risk Management

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

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

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

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

# 3 The Action Design Research Project

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

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

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

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

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

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

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

# 4 The Artificial Intelligence Risk Assessment Tool

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

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

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

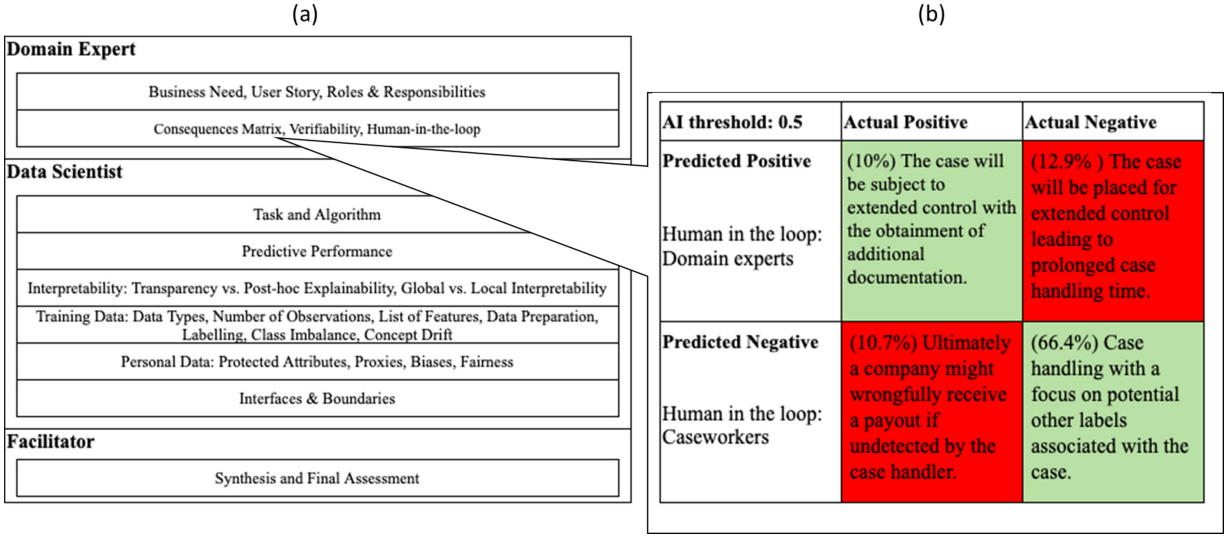


**Fig. 1.** (a) Schematic overview of the artificial intelligence risk assessment tool with (b) an Example of a consequence matrix

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

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

# 5 Reflection, Learning, and Formalization of Design Principles

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

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

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

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

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

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

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

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

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

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

# 6 Discussion

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

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

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

# References

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

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

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