Achieving a Data-Driven Risk Assessment Methodology for Ethical AI

Felländer, Anna; Rebane, Jonathan; Larsson, Stefan; Wiggberg, Mattias; Heintz, Fredrik

Published in:
Digital Society

DOI:
10.1007/s44206-022-00016-0

2022

Document Version:
Publisher's PDF, also known as Version of record

Digital Society (2022) 1:13
https://doi.org/10.1007/s44206-022-00016-0

ORIGINAL PAPER

Achieving a Data-Driven Risk Assessment Methodology for Ethical AI

Anna Felländer[1] · Jonathan Rebane[1,2] · Stefan Larsson[1,3] · Mattias Wiggberg[1,5] · Fredrik Heintz[1,4]

Received: 2 March 2022 / Accepted: 28 July 2022

Abstract

The AI landscape demands a broad set of legal, ethical, and societal considerations to be accounted for in order to develop ethical AI (eAI) solutions which sustain human values and rights. Currently, a variety of guidelines and a handful of niche tools exist to account for and tackle individual challenges. However, it is also well established that many organizations face practical challenges in navigating these considerations from a risk management perspective within AI governance. Therefore, new methodologies are needed to provide a well-vetted and real-world applicable structure and path through the checks and balances needed for ethically assessing and guiding the development of AI. In this paper, we show that a multidisciplinary research approach, spanning cross-sectional viewpoints, is the foundation of a pragmatic definition of ethical and societal risks faced by organizations using AI. Equally important are the findings of cross-structural governance for implementing eAI successfully. Based on evidence acquired from our multidisciplinary research investigation, we propose a novel data-driven risk assessment methodology, entitled DRESS-eAI. In addition, through the evaluation of our methodological implementation, we demonstrate its state-of-the-art relevance as a tool for sustaining human values in the data-driven AI era.

Keywords Ethical AI · Sustainability · Risk assessment

✉ Jonathan Rebane
jonathan@dsv.su.se

[1] anch.AI (formerly AI Sustainability Center), Stockholm, Sweden

[2] Department of Computer and Systems Sciences, Stockholm University, Stockholm, Sweden

[3] Department of Technology and Society, Lund University, Lund, Sweden

[4] Department of Computer and Information Science, Linköping University, Linköping, Sweden

[5] Department of Industrial Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

Published online: 18 August 2022

1 Introduction

The evolving data-driven technology sector has resulted in AI solutions becoming pervasively implemented throughout much of society, such as for personalized recommendations, health applications, and optimized processes (Jameson et al., 2002; Kalis et al., 2018; Tseng et al., 2021; Javaid et al., 2022; Groshev et al., 2021). These implementations demand a myriad of legal, ethical, and societal considerations which must be accounted for in order to develop ethical AI (**eAI**) solutions which sustain human values in an emerging data-driven era (Cath, 2018). The cost of ignoring eAI issues can be very high, with several high-profile AI systems ultimately needing to be shut down after risks inadvertently materialized and massive reputational losses occurred (Wolf et al., 2017; Analytica, 2018; Lauer, 2021). These problems are not limited to isolated events, with over 1400 reports of AI causing harm being reported in the Artificial Intelligence Incident Database (McGregor, 2020).

Effective and continuous risk management, utilizing risk assessment, is a vital component of ethics and compliance programs to anticipate and mitigate eAI risks before they occur. Risk management has backings in generalized ISO standards and should not be confused with ethical, financial, or operational auditing (Purdy, 2010). As of today, most means of risk management specific to the eAI landscape consists of a variety of guidelines, recommendations, and a handful of specified tools to account for and tackle individual challenges (Hagendorff, 2020; Jobin et al., 2019; Bellamy et al., 2019; Canca, 2020; Larsson, 2020). However, such resources have been criticized for being too abstract or technology-centered, lacking a direct focus on cross-functional organizational viewpoints and needs, such as the compatibility with standardized risk assessment models to bring principles to practice (Theodorou & Dignum, 2020). In addition, existing governance methodologies for eAI focus on research-focused processes rather than risk assessment, such as d'Aquin et al. (2018), or lack validation, emphasis on human rights, and cross-functional perspectives (Brendel et al., 2021). Proposals do exist for self-assessment, such as HLEGAI (2020), and structured approaches on "ethics-based auditing" such as Brown et al. (2021); Floridi and Cowls (2019). However, requests have been made for the development and validation of methods that can be applied in reality to managing eAI organization risk, as a means to assure the legality, ethics, and robustness of AI systems Wright (2020); Theodorou and Dignum (2020); Brendel et al. (2021). Furthermore, the proposal for an AI Act, published by the European Commission in April 2021 (Commission, 2021), puts much emphasis on assessments as a way to manage and mitigate high-risk use of AI systems. This proposal has faced critique in regard to its broad definition and potential for over-regulation of AI according to Glauner (2021), while others have characterized the proposal as an "auditing" regulation (Mökander et al., 2021).

From an organizational viewpoint, risk assessment methodologies for technical systems exist to provide a linear structure for identifying and mitigating unregulated business risks with individualized risk assessment phases (Pandey, 2012). However, due to the multidisciplinary nature of AI solutions, what is needed

are novel approaches, developed from a holistic multidisciplinary approach which incorporates technical, legal, and societal perspectives, with the objective of detecting negative eAI externalities of organizations that would otherwise infringe on legal and human rights alongside organizational principles (Dignum, 2020). Research gaps needed to be filled are in relation to multi-stakeholder organizational perspectives for establishing core concepts with regard to the eAI risk landscape (Rodrigues, 2020). Such knowledge can then be leveraged as a basis for filling a gap in relation to the development of an eAI risk assessment methodology. Due to the rapidly progressing eAI landscape, such a methodology must be flexible in the sense that defined risks, concepts, and methods are flexible enough to accommodate an evidence-based evolution.

It is well established that many organizations face practical challenges in navigating eAI considerations (Lauer, 2021; Rakova et al., 2021; Desouza et al., 2019). There is a general discussion regarding how ethical AI in organizations could be handled (Clarke, 2019). Yet the presence of proof-of-concepts where real organizations and real data have been tested is low. Therefore, new methodologies are needed to provide a structure and path through the checks and balances needed for ethically assessing an AI. The question we wish to answer in this paper is: *How can a standardized approach to ethical AI risk assessment be constructed that is compatible with cross-functional organizational demands over a large variety of contexts?*

Our contribution addresses this question as follows:

- Firstly, in Sect. 2, we report the findings of a multidisciplinary research investigation. This investigation was initiated as an eAI landscape review of risks, and then followed by cross-sector expert discussions which provided categorizations of unintended root causes of risks, which we call *pitfalls*. These discussions also identified *fundamentals*, which must be enacted by organizations within the eAI domain, in order to prevent pitfalls. These results, and subsequent content analysis, helped to formulate the requirements needed for a standardized eAI risk assessment methodology that is compatible with regulatory demands across a large variety of contexts.
- Next, in Sect. 3, we propose a novel data-driven methodology as a means to ensure that human values and rights are sustained for data-driven AI applications based on these results. We then leverage the discovered results to present a data-driven, cross-functional methodology implemented in the real world as a means to help ensure human values and rights are sustained in the data-driven AI era.
- In Sect. 4, an initial evaluation of the methodology implementation is provided in the context of two case studies that demonstrate the effectiveness and feasibility of the methodology including its data-driven development through iterative improvements.
- Finally, in Sect. 5, we demonstrate how repeated applications of the implementation on use cases have guided refinements. In addition, we outline the general data strategy of our implementation to provide a better understanding of how group-level data can and should be leveraged to refine implementations and provide insights.

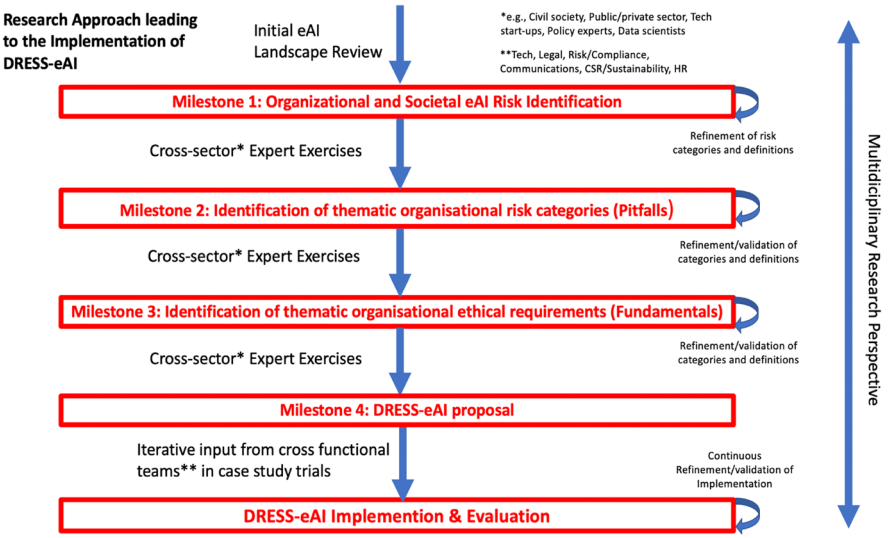

Fig. 1 Multidisciplinary research approach leading to the development, implementation, and evaluation of the proposed DRESS-eAI methodology

2 Research Approach

To answer the question of how to provide an approach to eAI risk assessment that is compatible with different organization demands, we systematically approached the problem from a multi-disciplinary research perspective, which is further clarified below, to establish requirements needed for a well-vetted and real-world applicable risk assessment methodology in eAI. We specify the importance of a multi-disciplinary approach as vital in order to capture computer scientific notions of AI as well as humanistic and social-scientific notions of ethics-based governance. This process was based on steps which have been visualized in Fig. 1, including organizations specified in Table 1, and experts specified in Table 2, whom participated in cross-sector expert exercises. Details are further specified in subsequent sections.

2.1 Organizational and Societal eAI Risk Identification

To establish the needs of a risk assessment methodology for organizational eAI risks which pose a threat to human values and rights, a systemic research process was conducted. As a first step, a literature review was performed by Larsson et al. (2019). This review included an assessment both quantitative and qualitative of the literature on fairness, transparency, and accountability in AI, in order to exhaustively examine the eAI landscape while identifying and discussing societal eAI risks which do not cause intentional harm. The review helped form a springboard for future expert exercises on the topic. One outcome of this study was

Table 1 Participating organizations in the cross-sector expert exercises

Organization	Sector
Boston Consulting Group	Cross-sector
Cirio (Law firm)	Legal
City of Malmö	Public
City of Stockholm	Public
Civil Rights Defenders	Human rights
Ericsson	Private tech
Google Sweden	Private tech
Human Rights Watch	Human rights
Karolinska Institute	Public research
KTH Royal Institute of Technology	Public research
Microsoft Sweden	Private tech
Sana Labs (AI for individualized learning)	Tech start-up
Stockholm School of Economics	Public research
Swedish institute for Standards	Cross-sector
Swedish Tax Agency	Public
Södertörn University	Public research
Telia	Private tech
The Institute for Futures Studies	Public research

a realization that much of the eAI landscape places a strong focus on technical rather than organizational risks. This realization motivated the need for highlighting organizational risks with society as the primary stakeholder.

Through an initial evaluation and iterative expert reviews of the topic, a list of the most common and distinctive risks related to eAI was composed. This consisted of eight organizational risks, whose definitions are backed by global human rights legislation and other external analysis by (General Assembly, 1948; Meek et al., 2016), including the commonality of issues addressed across ethics guidelines (Jobin et al., 2019). The eight identified organizational risks are as follows:

- **Privacy intrusion**—AI and data-driven solutions interfering with personal or sensitive data without regarding consent of the individual or groups whose data is collected, how data is shared or stored, agreement of the law, or other legitimate needs to protect the best interests of an individual or groups (right to privacy)
- **Amplified discrimination**—AI and data-driven solutions which cause, facilitate, maintain, or increase prejudicial decisions or treatment and/or biases towards race, sex, or any other protected groups obliged to equal treatment (right to fair treatment)
- **Violation of autonomy and independent decision making**—AI and data-driven solutions which intentionally or unintentionally, and without consent, facilitate behavioral changes that manipulate independent decision making and social well-being (right to autonomy)

Table 2 Description of experts from participating organizations in cross-sector exercises

Expert role	No. of experts
Policy area specialist	3
Regulator	2
Ethics philosopher	4
Standardization practitioner	4
Legal practitioner	10
Legal researcher	4
Policy think tank member	5
Human rights advocate	3
Human rights lawyer	2
Gender studies researcher	2
Government legal expert	3
Digital law expert	4
Tech company public affairs spokesperson	1
Data scientist	5
Economist	2
Global studies expert	1
Children's rights advocate	1
AI startup founder	4
Municipal chief digital officer	3
EU law expert	2
Consultancy firm partner	4
Health technology researcher	2
Clinician	1
Machine learning researcher	10
Machine learning expert	5
Member of Swedish parliament	2
Journalist	2
Board representative from cross-industry multinational company	11
Data collection expert	1

- **Social exclusion and segregation**—AI and data-driven solutions contributing to or maintaining an unfair denial of resources, rights, goods, and ability to participate in normal relationships and activities, whether in economic, social, cultural, organizational, or political arenas (right to inclusion)
- **Harm to safety**—AI and data-driven solutions facilitating unwanted physical harms to an individual or organization stemming from underdeveloped AI, and attributed to negligence from an organization (right to physical safety)
- **Harm to security of information**—AI and data-driven solutions facilitating potential damage from unauthorized access of private data, due to faulty data protection and processing, or criminal activity (right to security of information)

- **Misinformation and disinformation**—AI and data-driven solutions which intentionally or unintentionally distribute information that is regarded as false and harmful to society (right to be informed)
- **Prevention of access to public service**—AI and data-driven solutions contributing to or maintaining a denial of public social assistance and service (right to public service access)

2.2 Identification of Thematic Organizational Risk Source Categories (Pitfalls)

The output of the initial eAI landscape review formed the basis for a series of cross-sector expert-based exercises in which thematic categories of root causes of eAI risks and in an organizational context were identified. The aims of the exercises were to answer the following questions which arose from the landscape review:

- How should we define thematic categories as root causes of eAI risks in an organizational and societal context?
- How can these risks be mitigated from broad society- and organizational-based perspectives?

Cross-sector exercises were conducted as part of an initiative together with the Swedish Innovation Agency (Vinnova), a government agency that administers state funding for research and development. Starting in 2018, participants were gathered to perform a series of exercises in Stockholm, Sweden which involved cross-sector experts in the domains such as civil society, public/private sector, tech start-ups, and policy. This included organizations spanning legal, technical, business, communication, and sustainability/CSR from public and private domains, along with organizations associated with the Stockholm, Sweden-based AI Sustainability Center startup. See Tables 1 and 2 for a full list of these organizations and the expert roles. Exercises were performed with the listed organizations, over the course of several years, consisting of round-table discussions, panel discussions, seminars, and joint analyses of ethical AI topics. Following these exercises, a content analysis of notes taken was performed by principle researchers to further refine and validate thematic content which emerged. Such notes were subsequently shared with a total of 20 independent experts groups from Table 1 to provide independent perspectives and analysis on thematic content. This feedback was also integrated into the content analysis to validate and refine definitions surrounding the emergent themes. Evidence continues to be collected through use cases to ensure identified categories are exhaustive of the eAI risk landscape from societal and organization perspectives.

This selection was based on a need to form a multidisciplinary organizational perspective on societal, ethical, and legal considerations towards the eAI risk landscape. The goal of exercises performed was to reflect on how the identified eAI risks would appear in each of these domains and identify which thematic categories form the root cause of each of these risks across all domains. Through this process, the experts reached a consensus on four common themes, which we refer to as *pitfalls*.

Notes taken during these exercises were used in the previously specified content analysis approach to concretely narrow down and define the pitfall themes which emerged. These four pitfalls are:

- **Misuse/overuse of data**—The AI application/solution could be overly intrusive, using private data, or it could be used for unintended purposes by others. This can include misinterpretations of primary users regarding implementation or deployment the AI application/solution.[1]
- **Bias of the creator**—Values and bias are intentionally or unintentionally programmed by the creator who may also lack knowledge/skills of how the solution could scale in a broader context.[2]
- **Immature data and AI**—Insufficient training of algorithms on datasets as well as lack of representative data could lead to incorrect and unethical recommendations.[3]
- **Data bias**—The data available is not an accurate reflection of reality or the preferred reality and may lead to incorrect and unethical recommendations.[4]

2.3 Identification of Thematic Organizational Ethical Requirements (Fundamentals)

The next step of in the series of exercises was to discuss thematic categories for how to prevent and overcome such pitfalls in an organizational context. The results of this were the establishment of organization structural eAI foundations as thematic categories. Through this process, the experts identified four common themes, which we refer to as *fundamentals*, and are also echoed in much recent principled work on AI (Jobin et al., 2019). These 4 fundamentals consisted of:

- **Accountability**—The need to stand accountable and justify one's decisions and actions to its partners, users, and others with whom the system interacts.
- **Governance**—Establishment of policies, principles, and/or protocols, and continuous monitoring of their proper implementations.
- **Explainability**—Ensure that algorithmic decisions, as well as any data driving those decisions, can be explained to and understood by *end users and other stakeholders using nontechnical terms.* Explainability demands must meet a contextually appropriate level to establish trust across stakeholders.
- **Transparency**—It must be possible to discover, trace, and detect how and why a system made a particular decision or acted in a certain way, and, if a system causes harm, to discover the root cause. Transparency demands must meet a contextually appropriate level across entire systems to establish stakeholder trust.

[1] Further discussion on misuse/overuse of data: Brundage et al. (2018); Larsson (2021)
[2] Further discussion on bias of the creator: Whittaker et al. (2019); Noble (2018)
[3] Immature data and AI examples: Buolamwini and Gebru (2018); Shankar et al. (2017); Larsson (2019)
[4] Data bias examples: Buolamwini and Gebru (2018); Shankar et al. (2017)

It was discovered through expert-based exercises and subsequent content analysis of notes and independent feedback that these categories have to be addressed as minimum requirements for any organization wishing to achieve eAI. In addition, it was clear from within these discussions that meeting such requirements means that cross-functional considerations between roles must be taken into account from organizational levels to technical systems levels. Again, evidence to better define thematic content continues to be collected through use cases and further exercises to ensure these categories are exhaustive of the eAI risk landscape from societal and organization perspectives.

3 Proposal of an eAI Risk Assessment Methodology

As a realization of our results and to answer the question posed by this paper, we propose the following methodology, entitled the *Data-driven Risk Assessment Methodology for Ethical AI* (DRESS-eAI). DRESS-eAI is designed to focus on the detection of pitfalls and enact the fundamentals relevant to most eAI use cases while being structured as a process that is familiar to organizations as it is comparable to the International Organization for Standardization (ISO) standard 31000:2009 for risk management (Purdy, 2010). This is an accepted standard for risk management developed by hundreds of risk management professionals over the course of four years, and has previously seen utilization within sustainability-focused methodology (Tiganoaia et al., 2019). The six process phases of the methodology are inspired directly by the ISO 31000:2009 risk management process with each phase being identified as a necessary step for systematically ensuring rigorous eAI practices of an organization. In addition, our aim has been to make the DRESS-eAI methodology compatible with any phase of an AI systems life cycle, while being fully compatible with a recent Declaration of eAI[5]. This declaration was issued as a response to aid organizations in preparing for the upcoming AI regulation recently proposed by the European Commission[6]. The declaration can be fulfilled directly through applying DRESS-eAI to achieve fundamentals as minimum requirements and overcome pitfalls which are root causes of eAI risks. We envision DRESS-eAI as a formative step towards establishing the requested common normative standards for high-risk AI solutions which may pose a risk to health, safety, and fundamental rights.

Lack of cross-functional teams tackling eAI is a thematic issue that emerged within panel discussion exercises. To accommodate for this, we advocate that all these roles/functions spanning technical, legal, risk, compliance, communications, CSR/sustainability, and HR are part of the process. Secondly, due to the specific individual risks of eAI projects, we acknowledge their relevancy in detection and mitigation as a core part of this framework. Due to eAI risk landscape facing ongoing changes, we also acknowledge the need for DRESS-eAI and

[5] https://aisustainability.org/the-code/

[6] https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

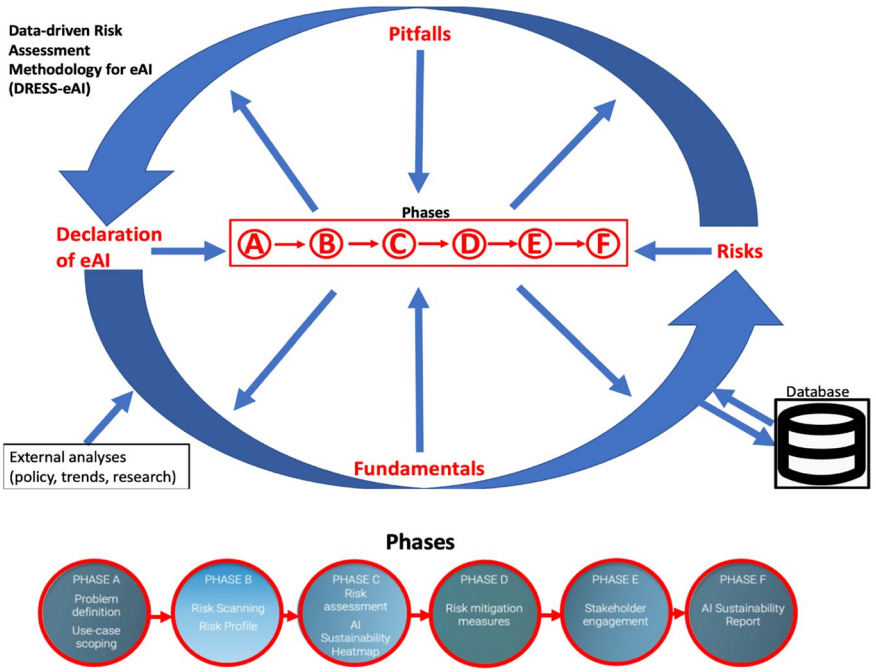

Fig. 2 An overview of DRESS-eAI. The main use-case process phases are shown in the center. Linkages show the conceptual flow between core concepts and how collected data can be used to construct and refine implementations of the phases. Also shown is how data outputs of use cases and external analyses are recorded in a database and used for generating insights and the iterative refining of existing categories and internal processes

its implementations to be evolving through *data-driven projects* where generated structured data from use cases and external analyses are recorded and used for iterative refining of existing categories and internal processes. This will permit implementations of the methodology to remain up-to-date alongside performing as an analytical tool. Thus, when highlighting the methodology as being *data-driven*, this refers to the inclusion of a central database to store structured data generated from the different phases which can later be used to provide data-driven insights and refine implementations. Figure 2 provides a complete overview of the proposed methodology.

We break the general process of the methodology into the following process phases for each eAI risk assessment use case. Such stages were based primarily on the ISO 31000:2009 risk management process:

- **Phase A:** Problem definition/use case scoping—Establishing a use-case definition including summary of challenges and identifying the project team. This should achieve a detailed description of the use case, including understanding of guiding policies/codes/values, key stakeholders, and technical specification.

- **Phase B:** Risk scanning/profiling—Capturing structured data related to the current state of achievement towards fundamentals and vulnerability towards pitfalls with data collection form multiple organization roles. In addition, if a screening is performed from multiple perspectives, this phase can provide a gap analysis which is an indication how well a specific use case is conforming to the organizational standards.
- **Phase C:** Risk assessment—Identification, evaluation, and prioritization of risk scenarios. For example, workshops should be used with a cross-functional team to identify ethical risk scenarios and what/which stakeholders could be impacted based on the risk exposure to pitfalls and fundamentals.
- **Phase D:** Risk mitigation measures—Identification of technical and non-technical mitigation measures and assigning ownership for actions. Identify risk mitigation measures that target the root cause of a risk scenario, or its effect. Plan for monitoring implementation of mitigating measures.
- **Phase E:** Stakeholder engagement—Capturing stakeholder feedback. A necessary validation step for identified risk mitigation activities; focusing on those affected by the organization's identified risk mitigation activities and what should be done to manage actual and potential impacts.
- **Phase F:** Review and maintain—Conclusions from the completion of each phase and recommendations going forward.

3.1 Implementation of DRESS-eAI

The above structure outlines and defines a generalized methodology for risk assessment within eAI. To explain how such a process can be enacted in the real world, we explain our implementation which has been applied and refined in relation to the case studies under examination.

The chosen implementation relies on collecting structured data through cross-functional self-assessment surveys. It is important to note that the chosen implementation may be prone to closed feedback loops, which can erroneously verify its own effectiveness and introduce data bias due to survey responses not reflecting true reality. As such, the implementation also collects and records qualitative feedback on the implementation directly through organizational stakeholders, permitting a deeper understanding of the implementation validity, rather than only relying on quantitative evidence acquired through repeated surveys which may possess respondent errors.

- In *phase A:* Problem definition/use case scoping—We perform workshops for the identification and detailing of an appropriate eAI use case. The outcome is a use-case definition including summary of challenges and detailed use-case description based on a pre-defined template. To leverage the data-driven nature of the methodology, we administer three structured surveys to capture data which can later be leveraged for data-driven group-level insights between phases. Firstly, an organizational survey to capture general questions such as the organization's size and domain. We also administer an organizational maturity survey to screen

for the organizations preparedness for ethically high-risk AIs. Finally, a use-case scoping survey is administered to capture a description of the AI solution that is to be assessed.

- In *phase B:* Risk scanning/profiling—We cover the eAI risk landscape with an exhaustive risk scan survey of over 150 questions tagged to and equally balanced according to relevant fundamental, pitfalls, and organizational role. These questions emerged as part of the same expert-based iterative process to exhaustively, and in a balanced manner, establish where an organization lies in the eAI risk landscape. This entails that each pitfall and fundamental is treated with equal priority in order to appropriately cover exposure to eAI pitfalls in the technically, legally, and societal defined risk landscape. Conceptually speaking, pitfalls may overlap with each other. However, for simplicity with our implementation, each phase B question is tagged to a single pitfall. The decision of how to tag each question to a pitfall corresponded to which point in a AI's life cycle the question was most associated to. Figure 3 provides an overview of how each pitfall was connected to the AI life cycle for the purposes of tagging questions. Tagging of fundamentals and roles were not associated to the AI life cycle; however, as stated, efforts were made to ensure that appropriate combinations of taggings were included to comprehensively cover the eAI risk landscape. Questions can be answered by each role with four options: "yes," "in-progress," "not sure," and "no." We emphasize the role-based structuring of the questions to ensure the validity and comprehensiveness of answers, in addition to activating cross-functional cooperation across the organization. These roles include technical, legal, risk, compliance, communications, CSR/sustainability, business owner, and HR. See Table 3 for examples of these questions and their tagging structure, and Fig. 4 for an example summary report. Importantly, all structured data from this stage is captured in our database and used to produce group-level insights which can verify the ability of this phase to exhaustively cover the risk landscape, along with using outputs from this phase to provide insights within other phases. The output of this phase can be used to provide a gap analysis which is specified further in Sect. 3.2.
- In *phase C:* Risk assessment—We identify and characterize risk scenarios guided by information acquired in phase B, constructing a traditional heat map of risk scenarios to aid in prioritizing risk mitigation procedures on organizational and use-case levels. Each risk scenario is tagged to a fundamental and a pitfall as well as one or more of the eight identified risks from Sect. 2. After a sufficient data collection period, we exploit our acquired database of risk scan surveys and risk scenarios to aid in the data-driven insight generation within and across phases. Risk scenarios are prioritized based on a qualitative analysis of likelihood and severity. Prioritized risk scenarios are characterized further, with input from additional interviews and focus meetings with the client if needed.
- In *phase D:* Risk mitigation measures—Risk mitigation tools and recommendations are determined which can be technical or non-technical. We identify risk owners for the prioritized risk scenarios, either taken in the project or identified improvements needed. We provide risk mitigation from both organization and use-case levels based on evidence acquired during evaluation. Each mitigation

Table 3 Selected examples of eAI risk scanning questions demonstrating the tagging structure of fundamentals, pitfall, and cross-functional organizations roles

Risk scanning question examples			
Question	**Fundamental**	**Pitfall**	**Organization role**
Have you tested model results for fairness with respect to different affected groups (e.g., tested for disparate error rates)?	Governance	Data bias	Technical
Have you defined what human bias means in the context of the solution and with regards to your organizations values or policies?	Explainability	Bias of the creator	CR/CSR
Are you confident in your organization's ability to detect, then shut down a malfunctioning solution(s) in a timely manner, i.e., before any harm to people or society is caused?	Governance	Misuse/overuse	Business owner
Are the explanations that you provide about your solution easily accessible and in clear terms to external parties?	Transparency	Misuse/overuse	Communications
Do you have a person/function who is responsible for deciding when the algorithm(s) in the solution are mature enough/market ready?	Accountability	Immature data/AI	Technical

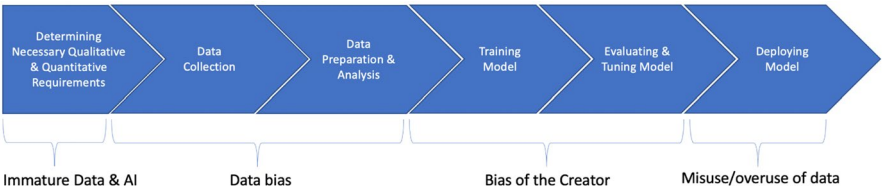

Fig. 3 Demonstrating how the pitfalls were mapped along the AI life cycle to provide mutually exclusive categorization of phase B risk scan questions for the DRESS-eAI implementation. The stages shown represent universally standard steps taken by organizational teams in developing and deploying AI solutions

measure is also tagged to a fundamental and pitfall as well as one or more of the eight identified risks from Sect. 2. After a sufficient data collection period, we leverage our database to provide data-driven recommendations and insights generation surrounding risk mitigation activities. Technical or non-technical risk-mitigating measures are identified and implemented in broader risk management/existing processes and assigned risk owners. Examples of risk mitigation measures: updated legal documents and processes, synthetic data for avoiding bias or to preserve privacy, tailored explainability models, training, establishing AI Ethical Principles or establishing an AI Ethical Board. Risk owners are identified within the organization and a plan is created for implementation and follow-up of actions.

- In *phase E:* Stakeholder engagement—we provide a summary of issues and recommendations on the topic of risk mitigation, and how these can be addressed and enacted through stakeholder engagements. Steps include identifying and prioritizing stakeholders to engage with; deciding what type of input is needed

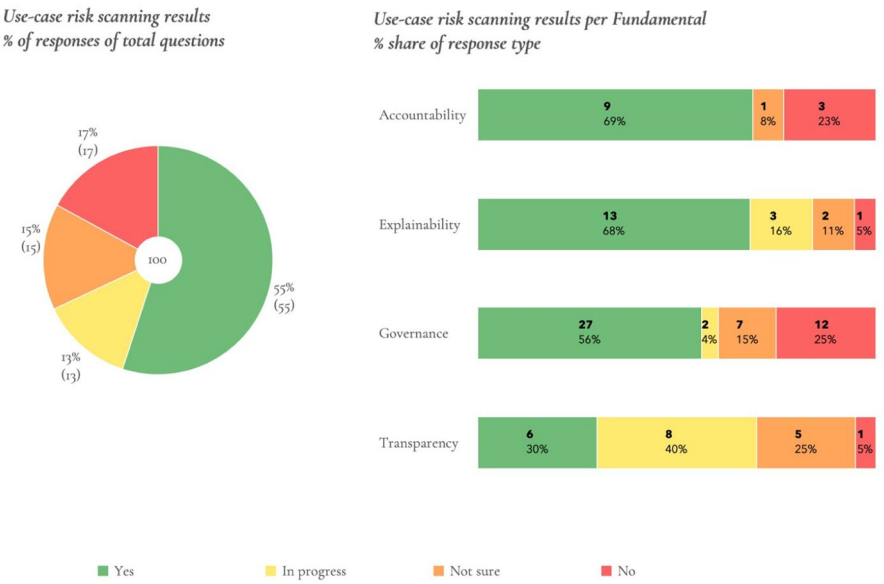

Fig. 4 Example overview of use-case output for phase B categorized by fundamentals

and whether to use existing stakeholder engagement forums /channels or targeted activities; collecting and analyzing information on key topics and problems addressed by stakeholders; stakeholder feedback captured in workshops and analyzed in a report summarizing the activities and concerns raised; and potentially creating a modified risk mitigation plan to address feedback.

- In *phase F:* Review and maintain—we report a summary of findings from applying the DRESS-eAI implementation. An updated risk scan of the use case is conducted in order to track the effectiveness of risk mitigation activities taken over time. We also provide recommendations on how internal frameworks can be strengthened. Qualitative feedback on the implementation's true impact is acquired.

3.2 Gap Analysis

eAI principles and commitments made by organizations are often high level, and analyses are needed to ensure a minimization of gaps between higher aspirations and what is actually happening on product and developer levels (Mittelstadt, 2019). Such principles ultimately have little effect on practices if they are not directly tied to structures of accountability, incentives, and the ways of working in an organization. AI principles, codes, and guidelines also need to be combined with monitoring of their implementation, as well as consequences if they are not met.

The phase B risk scanning survey output can be further used as a tool to identify possible gaps between stated ethical principles and higher aspiration and what might be happening on product or organizational level. For our DRESS-eAI implementation, we also map organizational AI principles directly to risk scanning question results to facilitate the gap analysis described in Sect. 3.2. A general example of this output can be seen in Fig. 5.

4 Evaluation and Iterative Evolution of DRESS-eAI Implementation

As described, the DRESS-eAI risk assessment methodology has been structured to follow a data-driven iterative approach for refining implemented processes and concepts. We have implemented and tested this to our knowledge unique methodology for assessing AI which is compatible to typical organizational structure and usable at any point in the life-cycle of an AI-system. We propose that any implementation of our methodology should not remain a static snapshot, but a data-driven, iteratively evolving system, capturing information from each use-case for insights into the developing eAI landscape and for refining DRESS-eAI methodological implementations.

In this section we outline the application of the DRESS-eAI methodology to two real organizational case studies, reporting the effectiveness of the current implementation's ability to detect and mitigate risk, along with reporting the data-driven

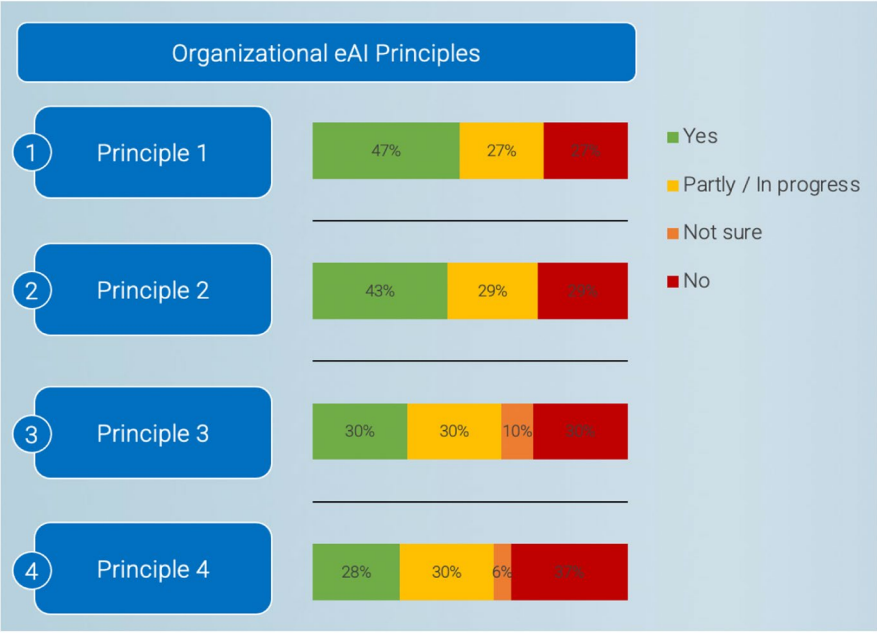

Fig. 5 Example output of a gap analysis, showing how an organization's AI principles are being achieved according to mappings to relevant DRESS-eAI's use-case risk scan questions

evolution of our implementation dictated by quantitative and qualitative evidence acquired from each case study.

4.1 Case Study 1: AI for Assisting Job Hiring Practices

4.1.1 Description

This use case revolves around an AI system being used to classify job seekers based on personal data pertaining to job hiring, education, and language proficiency, as well as data about the condition and functioning of the labor market. This has, in other studies, been shown to be an area with inherent risks (Lopez, 2021). The data used for training input was generated from various sources. For new job seekers, a self-assessment survey was answered, and personal data was generated. For job seekers already known, data was gathered from a data lake where existing data about the job seeker was stored. Job seekers are then profiled using a deep learning model on 64 features.

The output of the model was a prediction of how far from the labor market a job seeker is. Based on the outcome, job seekers are placed into three categories based on a rule-based selection. A human case worker would also be able to change the category

a job seeker is placed in. Primary early concerns were raised regarding the risk of discriminating against sensitive groups, such as foreign-born women.

4.1.2 Outcomes

After applying DRESS-eAI phases A, B, and C, it was identified that this use case was exposed to several ethical and societal risks. These mainly pertained to the pitfalls Bias of the creator and Data bias. Weaknesses in accountability and AI governance were identified. Risk scenarios were identified and nine of them were likely to occur and could result in severe impacts on people and society. These risk scenarios were prioritized for mitigation. Mitigating measures and risk owners were identified for each risk scenario identified. The mitigating actions taken in the project altered the solution to prevent misuse, such as updating UX interface to case workers to prevent misuse of model outputs; creating and communicating purpose statements to various stakeholders; and implementing methods for explaining model outputs and defining what needs to be explained and for whom (based on current and proposed future regulation). The effect is that the solution could then be scaled.

Furthermore, applying DRESS-eAI highlighted the need for better AI governance broadly across the organization. One key finding is a lack of ownership of an ethical AI framework internally. Following this, a cross-functional group of internal stakeholders has now been initiated as a permanent ethical AI group, with the responsibility of supporting developers of AI solutions and advancing the organizations ethical AI maturity. In addition, it is highlighted that an approach for AI fairness and explainability was needed in order to serve future AI solutions.

4.1.3 Input into Evolution of Implementation

- A need for separate use case and organizational risk mitigation was identified.
- Distinction added between risk mitigation measures that can be taken in the project as opposed to what needs to change in the line organization.
- When implementing DRESS-eAI, the need to involve representatives from business operations was identified. A role was then added to the survey, entitled *business owner*. It was found that when an AI system is part of a larger organizational process such as in this project, many risks are associated with lack of effective collaboration and/or instructions on how to use the AI system by the business unit. Specifying a business owner role enabled further cross-functional collaboration in identifying potential risk exposure and taking effective mitigation actions.
- Validation that DRESS-eAI can be applied to identify and mitigate eAI risk for a use case in the development life-cycle phase.

4.2 Case Study 2: AI for Detecting Tax Fraud

4.2.1 Description

This use case began in an idea life cycle stage, where an AI system was used to monitor and select transactions on third-party marketplace platforms that should be reviewed for potential tax fraud. The AI system would be implemented in the own environment of third-party platforms. The AI system used information about the individual and the transaction as features for classification.

4.2.2 Outcomes

Applying DRESS-eAI identified which types of eAI risks can occur when using AI to detect tax fraud, leading to an increased understanding and awareness of how prepared the organization was to handle such risks. Several eAI risk aspects were highlighted including a lack of clear organisational strategy for eAI; a lack of a systematic approach to detect and handle ethical AI risks; lack of accountability for ethical AI risks; an inability to monitor AI systems; a large exposure to the pitfalls "Data bias" and "Bias of the creator"; a need to instate a central steering committee for overseeing eAI operations; and a need for competence development. Mitigating actions were then performed, resulting in the organization, a year later, having an established eAI policy and plan to establish an eAI steering committee.

4.2.3 Input into Evolution of Implementation

- Two risk scannings were required, with one on the organization level and one on the standard use-case level. This leads to the inclusion of the phase A organizational survey to aid in streamlining the implementation.
- To foster better understanding around the terminology, a clearer distinction was made between transparency and explainability.
- Validation that DRESS-eAI can be applied to a planning stage use case.

5 Leveraging Group-Level Results—Refinement of Data Collection Tool, Data Strategy, and Insights

In this section, we examine the the data-driven aspect of the DRESS-eAI from two perspectives. Firstly, we demonstrate how repeated applications of use cases have guided the refinement of our implementation. Secondly, we outline the general data strategy of our implementation to provide a better understanding of how group-level data can and should be leveraged to refine implementations and provide eAI insights both within and across DRESS-eAI phases.

Table 4 Selected examples of eAI risk scanning questions showing the iterative evolution using case study evidence

Risk Scanning Question Example 5 Version 1			
Question	**Fundamental**	**Pitfall**	**Organization Role**
Do you have a diversity policy and procedures to ensure diversity in your organization?	Governance	Bias of the creator	HR
Do you have processes/approaches in place to ensure that there is diversity within your pool of designers and managers involved in the creation of the solution in terms of gender, culture, age, etc.?	Governance	Bias of the creator	HR
Updated Risk Scanning Question Example 5			
Question	**Fundamental**	**Pitfall**	**Organization Role**
- (Question removed due to overlap with question below)	Governance	Bias of the creator	HR
Do you have processes/approaches in place to ensure that there is diversity within your pool of designers and managers involved in the creation of the solution in terms of gender, culture, age, etc.?	Governance	Bias of the creator	HR

5.1 Refinement of Survey Tool Through Use-Case Insights

Any implementation of DRESS-eAI will demand the refinement of process tools to better accommodate for organizational needs and the developing eAI landscape. For our implementation, we recorded common feedback acquired during use cases in our database and made refinements based on group-level evidence. To help demonstrate this process, we present representative examples of how repeated use cases of our implementation resulted in the refinements of our data collection process. A complete overview of these examples can be viewed in Table 4.

In the first example, we examine a question pertaining to the data bias pitfall and governance fundamental. The question pertains to the needs of building a solution on the same data distributions in which it will be deployed. This is to help ensure it is not simply well fit to a training dataset and then underperforms on unfamiliar data examples in the real world. The original question was reported as incorrectly capturing the intention behind the question due to ambiguity. Since a technical role was intended to answer such a question, more detailed terminology about datasets and statistical distributions was included. In general, more exact terms for technical roles questions were added across questions.

In the second example, an additional question was added to better capture the product owner's input on whether or not they oversee the compatibility of the solution to their organization's values. This was part of the general refinement of the the risk scan to have the organizational product owner more involved in the risk scanning. This inclusion was noted as being crucial as product owners tended to understand the intended use and value of the solution more than other roles. To better identify potential vulnerabilities, more questions were added to have the product owner role as a larger part within the risk scan process, specifically asking them more organizationally related questions surrounding the governance and accountability pitfalls.

In the third example, it was reported that the original question could lead to incorrect responses due to ambiguities. The intention of the original question was to establish from the product owner whether or not they possess a general open data collection strategy; a lack of which could lead to insufficient data in terms of quality. However, the original wording of the question lead to misunderstandings that the question related to communicating human biases for selecting data. This update represents an example in which ambiguities in the questions were removed.

In the fourth example, it was noted that the question was both ambiguous and interpreted incorrectly by respondents. The intention of the question was to highlight a data bias vulnerability due to having automated data bias processes. This could lead to data bias due to a lack of human oversight. The original wording of the question did not make this intention clear as thus lead to incorrect responses. Such a question is representative of similar questions that needed to be rephrased.

In the fifth example, we highlight a general case in which redundant questions needed to be removed. For the questions shown we noted from feedback that the general organisation question was sufficiently covered by a similar question and thus could be removed.

Risk Scanning Question Example 1 Version 1

Question	Fundamental	Pitfall	Organization Role
Is the distribution of demographic groups in your dataset representative of the reality you are trying to reflect?	Governance	Data bias	TECH

Updated Risk Scanning Question Example 1

Question	Fundamental	Pitfall	Organization Role
Is the distribution of demographic groups in your dataset representative of the distribution present in the population(s) where your solution is/are deployed	Governance	Data bias	TECH

Risk Scanning Question Example 2 Version 1

Question	Fundamental	Pitfall	Organization Role
- (Question needed to be added)	-	-	-

Updated Risk Scanning Question Example 2

Question	Fundamental	Pitfall	Organization Role
Are you as product owner involved in the design, development, auditing etc. of the solution to ensure that the solution conforms to your organizationial values	Accountability	Misuse/Overuse	Product owner

Risk Scanning Question Example 3 Version 1

Question	Fundamental	Pitfall	Organization Role
Do you communicate to relevant stakeholders about on what biases and values your data was selected and processed?	Transparency	Data bias	Product owner

Updated Risk Scanning Question Example 3

Question	Fundamental	Pitfall	Organization Role
Do you communicate to relevant stakeholders about on what grounds the data was selected and processed?	Transparency	Data bias	Product owner

Risk Scanning Question Example 4 Version 1

Question	Fundamental	Pitfall	Organization Role
Do you have an automated process for data validation?	Governance	Data bias	TECH

Updated Risk Scanning Question Example 4

Question	Fundamental	Pitfall	Organization Role
Do you have an automated process/approach with human oversight for data validation?	Governance	Data bias	TECH

5.2 Data Strategy and Insights

Importantly, we wish to highlight the relevancy for this methodology to be data-driven by having an underlying database capable of storing structured information from each use case to acquire group-level insights. Utilizing this data-driven backbone of the methodology permits refinements in terms of efficiency, effectiveness, and deployability across a variety of contexts. We also model DRESS-eAI as a process that is continually refined though complementary external input such as new regulation, trends, and research on assessment methodology. Most importantly, the EU proposal for an AI Act is very likely to greatly impact European markets, stressing the need for these types of assessment. The data-driven backbone permits continuous adaption of implementations to the changing eAI landscape.

In practical terms, insights acquired from group-level data are exploited to improve each implementation in the following manner:

- Providing summary reports on the general effectiveness of an implementation, and the state of the eAI landscape.
- Permitting the benchmarking of eAI organizational status on per-sector and cross-sector levels.
- Acquiring greater contextual information with less time burden on clients through personalized questions
- Identifying deficiencies with existing surveys or tools.
- Developing internal and client dashboards and PR reports

With the DRESS-eAI database, data insights can be acquired though the independent analysis of each implementation phase. Of equal importance is the potential to understand how data from each phase is connected. Within the implementation applied for this study, we utilize a general approach for mapping phases together which can provide a structure for acquiring informative results though the means of statistical analysis and AI modeling. For our implementation, we build data relations across various phases through common attributes for each output data table. More specifically, we achieve such relations by tagging all questions in phase B, risk scenarios in phase C, and mitigation measures in phase D with attributes of their respective pitfalls and fundamentals. Phase C risk scenarios and phase D risk mitigation activities are also tagged to the eight risk categories identified and defined as part of this study. See Sect. 2 for an overview of the eight risks.

5.3 Comparing DRESS-eAI to Other Frameworks

IEEE's newly released IEEE Standard Model Process for Addressing Ethical Concerns during System Design (IEEE 7000-2021) (IEEE, 2021) addresses a set of processes by which *organizations can include consideration of ethical values throughout the stages of concept exploration and development is established by this standard* IEEE 7000-2021 supports organizations managers and engineers

in transparent communication with selected stakeholders to look into ethical values elicitation and prioritization. This involves *traceability of ethical values through an operational concept, value propositions, and value dispositions in the system design*. The standard is relevant to all sizes and types of organizations. While the IEEE 7000-2021 standard provides engineers and technologists with an *implementable process aligning innovation management processes, system design approaches, and software engineering methods to help address ethical concerns or risks during system design*, the DRESS-eAI structure is more focused on using the organization's own data as input for mitigating risk in their AI implementations. Hence, DRESS-eAI provides a value to the organization even though it is deployed after processes involving potential risks have been designed. In addition, through included case studies, DRESS-eAI has now proven to work in terms of real-world implementation the last three years.

Ernst and Young recently surveyed and assessed the ecosystem of artificial intelligence risk assessment (AIRA) methodologies (Ezeani et al., 2021). Claiming to present a snapshot of the landscape at a certain point in time, the report aims to inform policymakers about the AI risk assessment landscape and provide *emerging policy trends and leading practices*. Based on the surveyed reports four leading practices have been identified: categorization of risk, risk management, requirements for trustworthiness of AI, and relevant stakeholders for identifying and mitigating AI risk. Although formulated somewhat differently, all have their parallels with the DRESS-eAI methodology.

6 Limitations and Scope of DRESS-eAI

Limitations and scope of the DRESS-eAI framework are discussed in this section to help clarify the framework's intended use and applicability.

Firstly, we highlight that DRESS-eAI exists as an ethical AI risk assessment framework for organizations actively applying, or seeking to apply AI, and thus is not explicitly designed to support all functions of an ethics-based AI auditing framework. In this regard, DRESS-eAI does not explicitly support quantitative assessments of algorithmic bias and fairness which is achieved through direct analysis of training data and AI performance. DRESS-eAI also does not inherently provide full-compliance assessments tailored to regulations, such as those proposed in the EU AI Act (Commission, 2021). Furthermore, DRESS-eAI does not provide specific checklists nor guidelines to follow. It instead demands active participation of cross-functional engagement through an outlined playbook of steps to assess, report, and monitor comprehensive eAI risk exposure and mitigation plans on both organization and use-case levels.

We would note that the demands of cross-functional work needed for DRESS-eAI are not always realistically achievable for logistical reasons. The economic, technical, and expertise resources needed to complete an organizational-specific implementation of the DRESS-eAI framework may not be available. Current demands for attaining AI expertise from IT, business, and legal perspectives suggest that the DRESS-eAI framework can most realistically be applied to organizations with

at least a basic level of AI maturity. Furthermore, although DRESS-eAI explicitly integrates stakeholder engagement, the challenges of attaining meaningful stakeholder involvement within this field are reported by Costanza-Chock et al. (2022). To remedy several of these stated issues, we would defer organizations with limited resources to utilize, and gain inspiration, from the freely available digital implementation of DRESS-eAI[7]. It should be noted this implementation does not explicitly support the entire stakeholder engagement stage.

Finally, we note that the specified implementation demands the integration of questionnaires for the risk scanning phase. This questionnaire-based approach to implementing the framework can inherently lead to self-reporting errors and bias, potentially hurting the validity of framework findings. For this reason, we would advocate having checks for data quality, such as an independent review of question responses, or multiple respondents answering and comparing their responses, to help ensure the reliability and validity of results.

7 Conclusion and Future Directions

In this paper, we have outlined and motivated the problem of developing a vetted and real-world applicable approach to ethical AI risk assessment. We report the findings of our systematic multidisciplinary research approach to building definitions and establishing requirements needed for such a methodology. Importantly, our approach to involve cross-sector experts has highlighted a need for a methodology that incorporates cross-functional considerations that build on familiar organizational processes. Leveraging this evidence, we then propose a novel methodology named DRESS-eAI. Furthermore, we fully describe our implementation and report the effectiveness and evolution of our implementation by describing several case studies and group-level insights.

As ongoing work, we are actively employing the implementation of DRESS-eAI with organizations, continuously acquiring evidence to understand how our implementation of the methodology can be further refined. Such evidence will permit additional group-level analyses, afforded by the data-driven backbone of DRESS-eAI, providing data-driven insights, while refining risk assessment tools and gap analyses.

Funding Open access funding provided by Stockholm University. Partial financial support was received from Verket for innovationssystem (Vinnova).

Data Availability The datasets generated during and/or analyzed during the current study are not publicly available due to privacy rights granted to participating organizations.

Declarations

Conflict of Interest The authors declare no competing interests.

[7] https://platform.anch.ai/auth/login

References

Analytica, O. (2018). Us fatality could slow down self-driving car testing. *Emerald Expert Briefings*.

Bellamy, R., Dey, K., Hind, M., Hoffman, S., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K., & Zhang, Y. (2019). Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development, 1*. https://doi.org/10.1147/JRD.2019.2942287

Brendel, A., Mirbabaie, M., Lembcke, T.-B., & Hofeditz, L. (2021). Ethical management of artificial intelligence. *Sustainability, 13*,. https://doi.org/10.3390/su13041974

Brown, S., Davidovic, J., & Hasan, A. (2021). The algorithm audit: Scoring the algorithms that score us. *Big Data & Society*. https://doi.org/10.1177/2053951720983865

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B. et al. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation arXiv preprint. https://doi.org/10.48550/arXiv.1802.07228

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77–91). PMLR.

Canca, C. (2020). Operationalizing ai ethics principles. *Communications of the ACM, 63*, 18–21. https://doi.org/10.1007/s00146-021-01308-8

Cath, C. (2018). Governing artificial intelligence: ethical, legal and technical opportunities and challenges. https://doi.org/10.1098/rsta.2018.0080

Clarke, R. (2019). Why the world wants controls over artificial intelligence. *Computer Law and Security Review, 35*, 423–433. https://doi.org/10.1016/j.clsr.2019.04.006

Commission, E. (2021). Proposal for a regulation of the european parliament and the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. *EUR-Lex-52021PC0206*.

Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022). Who audits the auditors? recommendations from a field scan of the algorithmic auditing ecosystem. In *2022 ACM Conference on Fairness, Accountability, and Transparency* FAccT '22 (p. 1571–1583). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3531146.3533213

d'Aquin, M., Troullinou, P., O'Connor, N. E., Cullen, A., Faller, G., & Holden, L. (2018). Towards an "ethics by design" methodology for ai research projects. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 54–59). https://doi.org/10.1145/3278721.3278765

Desouza, K., Dawson, G., & Chenok, D. (2019). Designing, developing, and deploying artificial intelligence systems: Lessons from and for the public sector. *Business Horizons, 63*. https://doi.org/10.1016/j.bushor.2019.11.004

Dignum, V. (2020). *Ai is multidisciplinary. AI Matters, 5*, 18–21. https://doi.org/10.1145/3375637.3375644

Ezeani, G., Koene, A., Kumar, R., Santiago, N., & Wright, D. (2021). *A survey of artificial intelligence risk assessment methodologies - The global state of play and leading practices identified*. Technical Report Trilateral Research.

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.8cd550d1

General Assembly, U. G. (1948). Universal declaration of human rights. *UN General Assembly, 302*, 14–25.

Glauner, P. (2021). An assessment of the ai regulation proposed by the european commission. *arXiv preprint*. https://doi.org/10.48550/ARXIV.2105.15133

Groshev, M., Guimarães, C., Martín-Pérez, J., & de la Oliva, A. (2021). Toward intelligent cyber-physical systems: Digital twin meets artificial intelligence. *IEEE Communications Magazine, 59*, 14–20. https://doi.org/10.1109/MCOM.001.2001237

Hagendorff, T. (2020). The ethics of ai ethics: An evaluation of guidelines. *Minds and Machines, 30*, 99–120. https://doi.org/10.1007/s11023-020-09517-8

HLEGAI (2020). Assessment list for trustworthy artificial intelligence. *European Commission*.

IEEE (2021). Ieee standard model process for addressing ethical concerns during system design. *IEEE Std 7000-2021*, (pp. 1–82). https://doi.org/10.1109/IEEESTD.2021.9536679

Jameson, A., Konstan, J., & Riedl, J. (2002). Ai techniques for personalized recommendation. In *Tutorial at 18th National Conference on Artificial Intelligence (AAAI)*.

Javaid, M., Haleem, A., Singh, R. P., & Suman, R. (2022). Artificial intelligence applications for industry 4.0: A literature-based study. *Journal of Industrial Integration and Management, 7*, 83–111. https://doi.org/10.1142/S2424862221300040

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature Machine Intelligence, 1*, 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kalis, B., Collier, M., & Fu, R. (2018). 10 promising ai applications in health care. *Harvard business review*.

Larsson, S. (2019). The socio-legal relevance of artificial intelligence. *Droit et societe, 573–593*. https://doi.org/10.3917/drs1.103.0573

Larsson, S. (2020). On the governance of artificial intelligence through ethics guidelines. *Asian Journal of Law and Society, 7*, 1–23. https://doi.org/10.1017/als.2020.19

Larsson, S. (2021). Ai in the eu: Ethical guidelines as a governance tool. (pp. 85–111). https://doi.org/10.1007/978-3-030-63672-2_4

Larsson, S., Anneroth, M., Felländer, A., Felländer-Tsai, L., Heintz, F., & Ångström, R. C. (2019). Sustainable ai: An inventory of the state of knowledge of ethical, social, and legal challenges related to artificial intelligence.

Lauer, D. (2021). You cannot have ai ethics without ethics. *AI and Ethics, 1*, 21–25. https://doi.org/10.1007/s43681-020-00013-4

Lopez, P. (2021). Bias does not equal bias: A socio-technical typology of bias in data-based algorithmic systems. *Internet Policy Review*, *10*, 1–29. https://doi.org/10.14763/2021.4.1598

McGregor, S. (2020). Preventing repeated real world ai failures by cataloging incidents: The ai incident database. *arXiv preprint* arXiv:2011.08512

Meek, T., Barham, H., Beltaif, N., Kaadoor, A., & Akhter, T. (2016). Managing the ethical and risk implications of rapid advances in artificial intelligence: a literature review. In *2016 Portland International Conference on Management of Engineering and Technology (PICMET)* (pp. 682–693). https://doi.org/10.1109/PICMET.2016.7806752. IEEE.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical ai. *Nature Machine Intelligence, 1*, 501–507. https://doi.org/10.1038/s42256-019-0114-4

Mökander, J., Axente, M., Casolari, F., & Floridi, L. (2021). Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed european ai regulation. *Minds and Machines*, (pp. 1–28). https://doi.org/10.1007/s11023-021-09577-4

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. nyu Press. https://doi.org/10.18574/9781479833641

Pandey, S. K. (2012). A comparative study of risk assessment methodologies for information systems. Bulletin of Electrical Engineering and Informatics, 1, 111–122. https://doi.org/10.12928/eei.v1i2.231

Purdy, G. (2010). Iso 31000: 2009 setting a new standard for risk management. *Risk Analysis: An International Journal, 30*, 881–886. https://doi.org/10.1111/j.1539-6924.2010.01442.x

Rakova, B., Yang, J., Cramer, H., & Chowdhury, R. (2021). Where responsible ai meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction, 5*, 1–23. https://doi.org/10.1145/3449081

Rodrigues, R. (2020). Legal and human rights issues of ai: Gaps, challenges and vulnerabilities. *Journal of Responsible Technology, 4*. https://doi.org/10.1016/j.jrt.2020.100005

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., & Sculley, D. (2017). No classification without representation: Assessing geodiversity issues in open data sets for the developing world. https://doi.org/10.48550/arXiv.1711.08536

Theodorou, A., & Dignum, V. (2020). Towards ethical and socio-legal governance in ai. *Nature Machine Intelligence, 2*. https://doi.org/10.1038/s42256-019-0136-y

Tiganoaia, B., Niculescu, A., Negoita, O., & Popescu, M. (2019). A new sustainable model for risk managementrimm. *Sustainability, 11*, 1178. https://doi.org/10.3390/su11041178

Tseng, M.-L., Tran, T. P. T., Ha, H. M., Bui, T.-D., & Lim, M. K. (2021). Sustainable industrial and operation engineering trends and challenges toward industry 4.0: A data driven analysis. *Journal of Industrial and Production Engineering, 38*, 581–598. https://doi.org/10.1080/21681015.2021.1950227

Whittaker, M., Alper, M., Bennett, C. L., Hendren, S., Kaziunas, L., Mills, M., Morris, M. R., Rankin, J., Rogers, E., Salas, M. et al. (2019). Disability, bias, and ai. *AI Now Institute, November*.

Wolf, M., Miller, K., Grodzinsky, F. (2017). Why we should have seen that coming: Comments on microsoft tay "experiment", and wider implications. The ORBIT Journal, 1, 1–12. https://doi.org/10.29297/orbit.v1i2.49

Wright, S. A. (2020). Ai in the law: Towards assessing ethical risks. In *2020 IEEE International Conference on Big Data (Big Data)* (pp. 2160–2169). https://doi.org/10.1109/BigData50022.2020.9377950