# RAI Guidelines: Method for Generating Responsible AI Guidelines Grounded in Regulations and Usable by (Non-)Technical Roles

ANONYMOUS AUTHOR(S)*

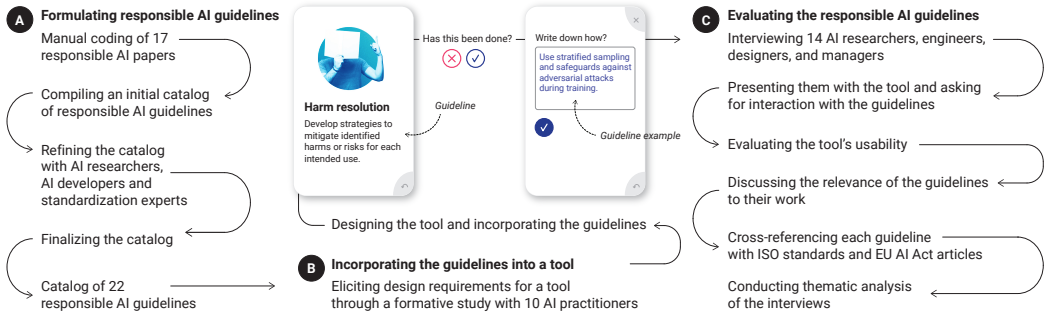


Fig. 1. Overview of our method for generating responsible AI guidelines and evaluating them: (A) formulating responsible AI guidelines that are grounded in regulations and are usable by different roles; (B) incorporating the guidelines into a tool; and (C) evaluating them.

Many guidelines for responsible AI have been suggested to help AI practitioners in the development of ethical and responsible AI systems. However, these guidelines are often neither grounded in regulation nor usable by different roles. To bridge this gap, we developed a four-step method to generate a list of responsible AI guidelines; these steps are: (1) manual coding of 17 papers on responsible AI; (2) compiling an initial catalog of responsible AI guidelines; (3) refining the catalog through interviews and expert panels; and (4) finalizing the catalog. To evaluate the resulting 22 guidelines, we incorporated them into an interactive tool and assessed them in a user study with 14 AI researchers, engineers, designers, and managers from a large technology company. Through interviews with these practitioners, we found that the guidelines were grounded in current regulations and usable across roles, encouraging self-reflection on ethical considerations at early stages of development. This significantly contributes to the concept of 'Responsible AI by Design'—a design-first approach that embeds responsible AI values throughout the development lifecycle and across various business roles.

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**; **Interactive systems and tools**; • **Computing methodologies** → **Machine learning**; **Artificial intelligence**.

Additional Key Words and Phrases: responsible AI, AI ethics, AI guidelines, system development, co-design

**ACM Reference Format:**
Anonymous Author(s). 2024. RAI Guidelines: Method for Generating Responsible AI Guidelines Grounded in Regulations and Usable by (Non-)Technical Roles. In *Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '24)*. ACM, New York, NY, USA, 29 pages.


*CSCW '24, November 09–13, San José, Costa Rica*

## 1 INTRODUCTION

The development of responsible AI systems [50, 87, 92] has become a significant concern as AI technologies continue to permeate various aspects of society [88]. While AI holds the potential to benefit humanity, concerns regarding biases [7, 15, 19] and the lack of transparency and accountability [66, 79] hinder its ability to unlock human capabilities on a large scale. In response, AI practitioners[1] are actively exploring ways to enhance responsible AI development and deployment. One popular approach is the use of tools such as checklists [59] or guideline cards [4, 26, 54] that are designed to promote AI fairness, transparency, and sustainability. These tools provide practical frameworks that enable practitioners to systematically assess and address ethical considerations throughout the AI development lifecycle. By incorporating checklists and guideline cards into their workflows, practitioners can evaluate key aspects such as data sources, model training, and decision-making processes to mitigate potential biases, ensure transparency, and promote the long-term sustainability of AI. However, these tools face two main challenges, creating a mismatch between their potential to support ethical AI development and their current design.

The first challenge is that these tools often exhibit a static nature, lacking the ability to dynamically incorporate the latest advancements in responsible AI literature and international standards [30, 73]. In the rapidly evolving field of responsible AI, new ethical considerations and regulatory guidelines constantly emerge (e.g., the EU AI Act [2]). It is therefore crucial for AI practitioners to stay updated of these developments to ensure their AI systems align with the current ethical and responsible AI practices. While checklists and guideline cards are increasingly used to assist and enhance the development of responsible AI systems, they are rarely grounded in current regulations. For example, Vakkuri et al. [94] proposed the ECCOLA cards that are based on AI ethics guidelines (e.g., IEEE Ethically Aligned Design and EU Trustworthy AI), which are not meant to be grounded on any specific regulations. Additionally, guidelines can quickly become outdated (e.g., the AI Blindspots deck has undergone several iterations [51, 52]), limiting their effectiveness in addressing evolving concerns related to fairness, transparency, and accountability.

The second challenge is that, while these tools emphasize the importance of involving stakeholders from diverse roles and backgrounds, they are often designed for specific AI practitioners (e.g., ML engineers), neglecting a broader spectrum of stakeholders (e.g., non-technical roles). Balayn et al. [8] found that less experienced practitioners in machine learning tend to use a limited set of metrics and methods from toolkits. Similarly, Deng et al. [20] stressed the lack of standardized guidelines in toolkits like AIF360 for introducing fairness issues to non-technical collaborators. Therefore, it is important that toolkits enhance communication, provide comprehensive guidance and support for cross-functional collaboration [101].

To overcome these challenges, we developed a four-step method to generate a list of responsible AI guidelines which we then incorporated in a tool to evaluate them (Figure 1). With this method, we aim to equip different roles with actionable guidelines that are grounded in regulations. To achieve this, we focused on answering this main research question: *How to generate responsible AI guidelines that are grounded in regulations and are usable by different roles?* In addressing this question, we made two main contributions:

(1) We proposed a four-step method for generating Responsible AI guidelines; these steps are: (1) manual coding of 17 papers on responsible AI; (2) compiling an initial catalog of responsible AI guidelines; (3) refining the catalog through interviews with 10 AI researchers and engineers, and workshops with 4 standardization experts; and (4) finalizing the catalog. This procedure resulted into a set of 22 Responsible AI guidelines (§4).

[1]We use the term practitioners to cover a wide range of stakeholders including AI engineers, developers, researchers, designers, ethics experts.

(2) We evaluated the 22 guidelines in a user study with 14 AI researchers, engineers, designers, product managers from a large technology company (§5) by designing and deploying a tool incorporating the guidelines. To develop the tool, we conducted a formative study with 10 AI practitioners to determine key design requirements. Using these requirements, we populated the tool with the guidelines and conducted the case study. Interviews with the 14 AI researchers, engineers, designers, and managers revealed that the guidelines were grounded to current regulations and were effectively usable across different roles, promoting self-reflection on ethical considerations in early development stages.

In light of these findings, we discuss how our method contributes to the idea of "Responsible AI by Design" by contextualizing the guidelines, informing existing or new theories, and offering practical recommendations for incorporating responsible AI guidelines into toolkits and recommendations for technical and non-technical roles in enabling orgnanizational accountability (§6).

## 2 RELATED WORK

We surveyed various lines of research that our work draws upon, and grouped them into two main areas: *i)* AI regulation and governance (§2.1), and *ii)* responsible AI practices and toolkits (§2.2).

### 2.1 AI Regulation and Governance

The landscape of AI regulation and governance is constantly evolving [47, 68]. At the time of writing, the European Union (EU) has endorsed new transparency and risk-management rules for AI systems known as the EU AI Act [2], which is expected to become law in 2024. Similarly, the United States (US) has recently passed a blueprint of the AI Bill of Rights in late 2022 [44]. This bill comprises *"five principles and associated practices to help guide the design, use, and deployment of automated systems to protect the rights of the American public in the age of AI."* While both the EU and US share a conceptual alignment on key principles of responsible AI, such as fairness and explainability, as well as the importance of international standards (e.g., ISO 24028 for Trustworthiness), the specific AI risk management regimes they are developing are potentially diverging, creating an "artificial divide" [32]. The EU aims to become the leading regulator for AI globally, while the US takes the view that excessive regulation may impede innovation.

Notable predecessors to AI regulations include the EU GDPR law on data protection and privacy [28], the US Anti-discrimination Act [27], and the UK Equality Act 2010 [38]. GDPR's Article 25 mandates that data controllers must implement appropriate technical and organizational measures during the design and implementation stages of data processing to safeguard the rights of data subjects. The Anti-discrimination Act prohibits employment decisions based on an individual's race, color, religion, sex (including gender identity, sexual orientation, and pregnancy), national origin, age (40 or older), disability, or genetic information. This legislation ensures fairness in AI-assisted hiring systems. Similarly, the UK Equality Act provides legal protection against discrimination in the workplace and wider society.

The National Institute of Standards and Technology (NIST), a renowned organization for developing frameworks and standards, recently published an AI risk management framework [73]. According to the NIST framework, an AI system is defined as *"an engineered or machine-based system capable of generating outputs such as predictions, recommendations, or decisions that influence real or virtual environments, based on a given set of objectives. These systems are designed to operate with varying levels of autonomy."* Similarly, the Principled Artificial Intelligence white paper from the Berkman Klein Center [30] highlights eight key thematic trends that represent a growing consensus on responsible AI. These themes include privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology,

professional responsibility, and the promotion of human values. Building on these themes, previous works have proposed a set of guidelines involving specific groups of AI practitioners. Saleema et al. [4] proposed 168 guidelines on how to design AI tailored to HCI practitioners. Similarly, Vakkuri et al. [94] formulated AI ethics guidelines tailored to researchers and technologists. No subsequent work has associated these guidelines with current international standards or regulations.

**Research Gaps.** As AI regulation and governance continue to evolve, AI practitioners are faced with the challenge of staying updated not only with the changing guidelines, but also with the regulations, requiring significant time and effort. Because prior guidelines lacked alignment with regulations, standards, and the input of experts in those fields, this work aims to create a methodology for crafting responsible AI guidelines that adhere to regulations and standards.

### 2.2 Responsible AI Practices and Toolkits

**Responsible AI Toolkits.** At the time of writing, the OECD's website lists 613 toolkits dedicated to fostering the development and deployment of responsible AI systems [74]. These toolkits are essential for operationalizing guidelines and regulations to assist AI practitioners such as engineers and researchers in addressing algorithmic bias [11, 34], explaining algorithmic decisions [6], and ensuring privacy in AI systems [30]. For addressing algorithmic bias, Google's Fairness Indicators toolkit allows developers to assess data distribution and model performance across user-defined groups [37]. IBM's AI Fairness 360 offers fairness metrics for bias mitigation [45]. Microsoft's Fairlearn assesses model impact on specific groups (e.g., under-represented populations) in terms of fairness and accuracy [29]. For explaining algorithmic decisions, IBM's AI Explainability 360 provides metrics and guidance for explainability, and new visualization techniques to enhance transparency [12, 36, 70]. Finally, for ensuring privacy in AI systems, IBM's AI Privacy 360 helps assess and mitigate privacy risks through data anonymization and minimization [17, 30, 84].

**Toolkits Used in Practice.** Developing toolkits specialized for certain audiences such as AI developers can lead to techno-solutionism, focusing exclusively on technical fixes. However, responsible AI entails broader socio-technical challenges (e.g., diversity and inclusion in decision-making) that require involvement of different roles with diverse expertise and background [85], typically discussed in venues with a long-standing commitment to human-centered design such as CHI, CSCW, AIES, and FAccT.

Different roles (e.g., data scientists, ML engineers and developers, UX designers) use toolkits in various ways. Data scientists often struggle to fully grasp visualizations of interpretable tools (e.g., InterpretML [69] and SHAP [58]), hindering their ability to understand datasets and underlying models [48]. Experienced ML developers and engineers often go beyond what fairness toolkits offer to tackle algorithmic unfairness, while those with less experience typically use only a few metrics and methods from these toolkits [8]. UX designers often rely on custom prototypes and their own past experiences to help contextualize responsible AI issues for non-technical colleagues [20] due to communication gaps [101].

Major communication gaps between technical and non-technical roles typically arise because these roles are involved in different stages of a project, which is likely to create fragmentation in communication [77]. By exploring how data science teams collaborate, Zhang et al. [102] found that non-technical roles play more prominent roles in the early and late stages of projects, while technical roles primarily handle the core data and modeling tasks. However, this disparity in involvement at various project stages is likely to create fragmentation. In fact, Organizational Science research reinforces the notion that effective communication and collaboration is crucial for overcoming the "silo mentality" [35]. Consequently, due to this fragmentation and a lack of robust organizational support, practitioners often take on "bridging" roles to help the communication

between the technical and non-technical project members [21]. One way of doing so is through "leaky abstractions" [91]. These are representations that are meant to communicate the inner workings and technical aspects of an AI system to these roles. Similarly, Nahar et al. [71] highlighted the extreme difficulty faced by non-technical practitioners in eliciting requirements due to the absence of suitable tools and the involvement of diverse stakeholders, highlighting the need for integrating communication features into toolkits. The design of such features was explored by Elsayed-Ali et al. [26] who developed question cards to facilitate stakeholder group discussions. These cards included built-in mechanisms for the automatic and cyclical assignment of cards to different participants, ensuring that everyone had the opportunity to share their opinions during the discussion.

**Research Gaps.** While many toolkits emphasize the importance of involving stakeholders from diverse roles and backgrounds, it becomes evident that they are frequently designed for specific stakeholders (e.g., ML engineers), thereby neglecting a broader spectrum of roles, especially non-technical roles. To address this gap, we aim to develop a set of actionable guidelines that are usable by a diverse range of stakeholders.

## 3 AUTHOR POSITIONALITY STATEMENT

Understanding researcher positionality is crucial for transparently examining our perspectives on methodology, data collection, and analyses [33, 42]. In this paper, we situate ourselves in a Western country[2] during the 21$^{st}$ century, writing as authors primarily engaged in academic and industry research. Our team comprises three males and two females from Southern, Eastern, and North Europe, and Middle East with diverse ethnic and religious backgrounds. Our collective expertise spans various fields, including human-computer interaction (HCI), ubiquitous computing, software engineering, artificial intelligence, data visualization, and digital humanities.

It is important to recognize that our backgrounds and experiences have shaped our positionality. As HCI researchers affiliated with a predominantly Western organization, we acknowledge the need to expand the understanding of the research questions and methodology presented in this paper. Consequently, our positionality may have influenced the subjectivity inherent in framing our research questions, selecting our methodology, designing our study, and interpreting and analyzing our data.

## 4 METHOD FOR GENERATING RESPONSIBLE AI GUIDELINES

To generate a list of responsible AI guidelines, we followed a four-step process (Figure 2), based on the methodology proposed by Michie et al. [63]. This process allowed us to identify the essential element of a guideline, referred to as the "active ingredient," focusing on the "what" rather than the "how" [62]. A similar parallel can be drawn in software engineering, where the "what" represents the software requirements and the "how" represents the software design, both of which are important for a successful software product [3]. However, by shifting the focus to the "what," AI practitioners can develop a clearer understanding of the objectives and goals they need to achieve, fostering a deeper comprehension of complex underlying ethical concepts. Throughout this process, we actively engaged a diverse group of stakeholders, including AI engineers, researchers, designers, product managers, and experts in law and standardization. As a result, we were able to formulate a total of **22** responsible AI guidelines (Panel A of Figure 1).

[2]REDACTED FOR BLIND REVIEW.

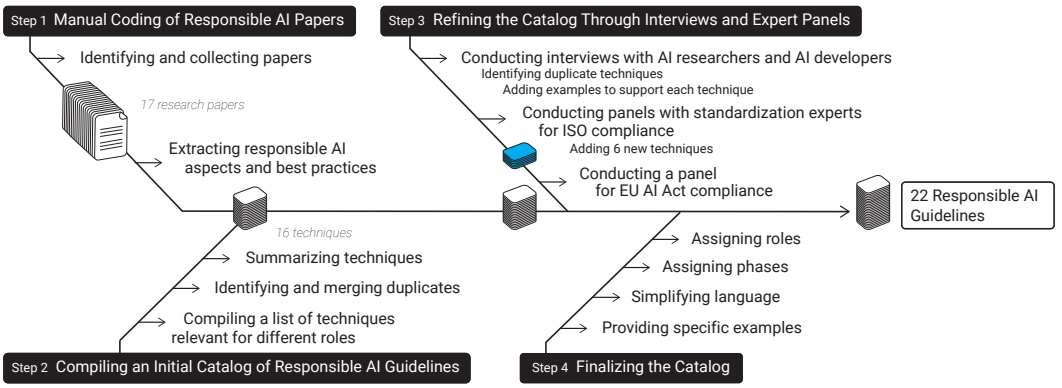


Fig. 2. Four-step method for generating responsible AI guidelines. These guidelines were derived from research papers, and are in line with ISO standards and the EU AI Act [2].

### 4.1 Manual Coding of Responsible AI papers

In the first step, we compiled a list of key scientific articles focusing on responsible AI guidelines applicable to a diverse set of roles, and manually coded them. We created this list by targeting papers published in renowned computer science conferences, such as the ACM CHI, CSCW, FAccT, AAAI/ACM Conference on AI, Ethics, and Society (AIES), and scientific literature from the medical domain (e.g., the Annals of Internal Medicine). Note that we did not conduct a systematic literature. Instead, we identified 17 papers that served as a starting point to compile an initial catalogue of techniques covering a broad range of responsible AI aspects, including fairness, explainability, sustainability, and best practices for data and model documentation and evaluation. These are foundational papers in responsible AI, and, as we shall see in a subsequent step of our methodology (§4.3), we refined the techniques identified from those papers through interviews and expert panels as well as cross-referencing them with the EU AI Act and ISO standards.

These papers encompass a growing body of research focusing on the work practices (e.g., ensuring fairness or models' explainable outputs) of AI practitioners in addressing responsible AI issues. This strand of research covers various aspects of responsible AI, including fairness, explainability, sustainability, and best practices for data and model documentation and evaluation. Fairness is a fundamental value in responsible AI, but its definition is complex and multifaceted [72]. To assess bias in classification outputs, various research efforts have introduced quantitative metrics such as disparate impact and equalized odds, as discussed by Dixon et al. [23]. Another concept explored in the literature is "equality of opportunity," advocated by Hardt et al.[75], which ensures that predictive models are equally accurate across different groups defined by protected attributes like race or gender. Equally important is the development of dedicated checklists for fairness [59]. Explainable AI (XAI) is another aspect of responsible AI. XAI involves tools and frameworks that assist end users and stakeholders in understanding and interpreting predictions made by machine learning models [5, 25, 40, 53, 56, 70]. Furthermore, the environmental impact of training AI models should also be considered. Numerous reports have highlighted the significant carbon footprint associated with deep learning and large language models [41, 86, 90]. Best practices for data documentation and model evaluation have also been developed to promote fairness in AI systems. Gebru et al. [34] proposed "Datasheets for Datasets" as a comprehensive means of providing information about a dataset, including data provenance, key characteristics, relevant regulations, test results, and potential biases. Similarly, Bender et al.[10] introduced "data statements" as qualitative summaries

that offer crucial context about a dataset's population, aiding in identifying biases and understanding generalizability. For model evaluation, Mitchell et al. [66] suggested the use of model cards, which provide standardized information about machine learning models, including their intended use, performance metrics, potential biases, and data limitations. Transparent reporting practices, such as the TRIPOD statement by Collins et al. [18] in the medical domain, emphasize standardized and comprehensive reporting to enhance credibility and reproducibility of AI prediction models.

### 4.2 Compiling an Initial Catalog of Responsible AI Guidelines

For each source, we compiled a list of techniques that could be employed to create responsible AI guidelines, focusing on the actions different roles (i.e., designers, researchers, developers, product managers) should consider during AI development. Following the methodology proposed by Michie et al. [63] (which was also used to identify community engagement techniques by Dittus et al. [22]), we sought techniques that describe the "active ingredient" of what needs to be done. This means that the phrasing of the technique should focus on *what* needs to do be done, rather than the specific implementation details of *how* it should be done. For example, a recommended practice for ensuring fairness involves evaluating an AI system across different demographic groups [23, 59, 75]. In this case, the technique specifies "what" needs to be done rather than "how" it should be implemented (e.g., using common fairness metrics such as demographic parity or equalized odds). In total, we formulated a set of 16 techniques based on relevant literature sources [5, 10, 18, 23, 30, 34, 41, 43, 53, 59, 66, 67, 75, 86, 95, 96].

We then conducted an iterative review of the collection of techniques to identify duplicates, which were instances where multiple sources referred to the same technique. For example, four sources indicated that data biases could affect the model [10, 34, 43, 66], emphasizing the need to report the characteristics of training and testing datasets. We consolidated such instances by retaining the specific actions to be taken (e.g., reporting dataset characteristics). This process resulted in an initial list of 16 distinct techniques. We provided a concise summary sentence for each technique, utilizing active verbs to emphasize the recommended actions.

### 4.3 Refining the Catalog Through Interviews and Expert Panels

The catalog of techniques underwent eleven iterations to ensure clarity and comprehensive thematic coverage. The iterations were carried out by two authors, with the first author conducting interviews with five AI researchers and developers. During the interviews, the participants were asked to consider their current AI projects and provide insights on the implementation of each technique, focusing on the "how" aspect. This served two purposes: firstly, to identify any statements that were unclear or vague, prompting suggestions for alternative phrasing; and secondly, to expand the catalog further. The interviews yielded two main recommendations for improvement: (1) mapping duplicate techniques to the same underlying action(s); and (2) adding examples to support each technique.

In addition to the interviews, the two authors who developed the initial catalog conducted a series of six 1-hour expert panels with two standardization experts from a large organization. The purpose of these panels was to review the initial catalog for ISO compliance. The standardization experts examined six AI-related ISOs, including ISO 38507, ISO 23894, ISO 5338, ISO 24028, ISO 24027, and ISO 24368, which were developed at the time of writing. Then experts provided input on any missing techniques and mapped each technique in the initial catalog to the corresponding ISO that covers it. As a result of this exercise, six new techniques (#2, #7, #12, #13, #14, #21) were

added to the catalog, resulting in a total of 22 guidelines. Next, we provide we provide a high-level summary of each ISO.[3]

**ISO 38507 (Governance, 28 pages).** It offers guidance on responsible AI use (e.g., identify potential harms and risks for each intended use(s) of the systems) and recommendations about current and future AI uses to governing bodies and various stakeholders such as managers and auditors.
**ISO 23894 (Risk Management, 26 pages).** It provides guidelines for managing AI-related risks (e.g., mechanisms for incentivizing reporting of system harms) in developing, producing, deploying, or using AI products and systems, including recommendations for integrating risk management into AI processes.
**ISO 5338 (AI Lifecycle Process, 27 pages).** It provides a framework for the life cycle of AI systems, detailing processes for managing and enhancing these systems from development to implementation (e.g., through reporting of harms and risks, obtaining approval of intended uses).
**ISO 24028 (Trustworthiness, 43 pages).** It offers guidance on trustworthiness in AI systems, focusing on transparency, explainability, controllability, and addressing potential risks with mitigation techniques. It also covers AI systems' availability, resiliency, reliability, accuracy, safety, security, and privacy.
**ISO 24027 (Bias, 39 pages).** It discusses bias in AI systems related to protected attributes such as age and gender, especially in AI-aided decision-making, providing techniques to measure and assess bias throughout the AI system lifecycle.
**ISO 24368 (Ethical and Societal Concerns, 48 pages).** It provides an introduction to ethical and societal concerns related to AI (e.g., principles, processes, and methods), targeting technologists, regulators, interest groups, and society as a whole.

As the final step of refining the catalog, the two authors reviewed the 85 articles of the EU AI Act [2] to map each of the 22 guidelines with the most relevant article(s), as shown in the last column of Table 1. They began with Article 3 of the Act, which defines the key concepts of an AI system, including its definition, intended purpose, performance, training, validation, and post-deployment monitoring. After reading all the articles and annotating them, they identified 22 unique articles corresponding to the guidelines. Articles 9, 10, and 17 were mapped to multiple guidelines. For example, Article 9 (*Risk management system*) states that "a risk management system shall be established, implemented, documented and maintained throughout the entire lifecycle of a high-risk AI system". This article aligns with guidelines #1, #3-5, and #13 as it is about the identification of harms and risks of the AI system's intended use. Article 10 (*Data and data governance*) states that *"training, validation and testing data sets shall be subject to appropriate data governance and management practices"*. This article aligns with guidelines #8 and #15-18 as it discusses the management and quality of data for training, validation, and testing, including aspects of diversity and minimizing biases. Finally, Article 17 (*Quality management system*) states that *"an AI system shall be documented in a systematic and orderly manner in the form of written policies, procedures and instructions"*. This article aligns with guidelines #6, #7, #10, and #14-18 because it is about documentation of all system components, including AI models and testing and validation procedures. The full mapping along with justifications is provided in Appendix B.

### 4.4 Finalizing the Catalog

In response to the interviews with AI developers and standardization experts, we incorporated an example for each guideline. For instance, under the guideline on system interpretability (guideline #9), the example provided reads: "output feature importance and provide human-understandable explanations." Furthermore, we simplified the language by avoiding domain-specific or technical

[3]Note that the summary provided is a brief and simplified description due to a paywall restriction.

Table 1. Responsible AI guidelines are actionable items that a developer should consider during the 3 phases of AI development lifecycle. These guidelines are grounded in the scientific literature (main sources are reported), and were checked for ISO compliance: ISO 38507 (Governance); ISO 23894 (Risk management); ISO 5338 (AI lifecycle processes); ISO 24028 (Trustworthiness); ISO 24027 (Bias); ISO 24368 (Ethical considerations). They were also cross-referenced with the EU AI Act's articles [2]. They are marked with the 'Phase' during which a guideline can be applied. There are three phases: development ($P_1$), deployment ($P_2$), and use ($P_3$). Guidelines are also marked with the job 'Role' that should consider them. There are three roles: designer ($R_D$), engineer or researcher ($R_E$), and manager or executive ($R_M$). Each guideline is followed by an example, and the guidelines are categorized thematically into six categories, concerning the *intended uses, harms, system, data, oversight, and team.*

| Number | Guideline | Phase | Role | Source(s) | ISO | AI Act |
|---|---|---|---|---|---|---|
| **INTENDED USES** | | | | | | |
| 1 | Work with relevant parties to identify intended uses.<br>(e.g., identify the system's usage, deployment, and contextual conditions) | $P_{1-3}$ | $R_{D,E,M}$ | [66] | 5338, 38507, 23894, 24027, 24368 | Art. 6, 9 |
| 2 | Obtain approval from an Ethics Committee or similar body for intended uses.<br>(e.g., Obtain Ethics Committee approval for the intended use, aligned with sustainability goals) | $P_{1-3}$ | $R_{D,E,M}$ | – | 38507, 5338, 23894 | Art. 11, 69 |
| **HARMS** | | | | | | |
| 3 | Identify potential harms and risks associated with the intended uses.<br>(e.g., prevent privacy violation, discrimination, and adversarial attacks, provide interpretable output) | $P_{1-3}$ | $R_{D,E,M}$ | [59] | 23894, 24028, 38507, 24368 | Art. 9, 65 |
| 4 | Provide mechanism(s) for incentivizing reporting of system harms.<br>(e.g., provide contact emails and feedback form to raise concerns) | $P_1$ | $R_{D,E}$ | [59] | 38507, 23894 | Art. 9, 60-63 |
| 5 | Develop strategies to mitigate identified harms or risks for each intended use.<br>(e.g., use stratified sampling and safeguards against adversarial attacks during training) | $P_{1-3}$ | $R_{D,M}$ | [66] | 24368, 23894 | Art. 9, 67 |
| **SYSTEM** | | | | | | |
| 6 | Document all system components, including the AI models, to enable reproducibility and scrutiny.<br>(e.g., create UML diagrams, flowcharts, and specify model types, versions, hardware architecture) | $P_{1-3}$ | $R_{D,E}$ | [59] | 5338, 23894, 24027 | Art. 11, 12, 16-18, 50 |
| 7 | Review the code for reliability<br>(e.g., manage version control using software.) | $P_{1-3}$ | $R_{D,E}$ | – | 5338 | Art. 17 |
| 8 | Report evaluation metrics for various groups based on factors such as age, gender, and ethnicity.<br>(e.g., evaluate false positive/negative, AUC, and feature importance across protected attributes) | $P_{1-3}$ | $R_{D,E,M}$ | [23, 59, 75] [67, 95] | 23894, 5338, 24028, 24027 | Art. 10, 13 |
| 9 | Provide mechanisms for interpretable outputs and auditing.<br>(e.g., output feature importance and provide human-understandable explanations) | $P_{1-3}$ | $R_{D,E,M}$ | [5, 53] | 38507, 24028 | Art. 12-14 |
| 10 | Document the security of all system components in consultation with experts.<br>(e.g., guard against adversarial attacks and unauthorized access) | $P_{1-3}$ | $R_{E,M}$ | [30] | 24028, 24368 | Art. 12, 13, 15, 17 |
| 11 | Provide an environmental assessment of the system.<br>(e.g., report the number of GPU hours used in training and deployment) | $P_{1-3}$ | $R_E$ | [41, 86] | 38507, 23894, 5338, 24368 | Art. 69 |
| 12 | Develop feedback mechanisms to update the system.<br>(e.g., provide contact email, feedback form, and notification of new knowledge extracted) | $P_{1-3}$ | $R_{D,E}$ | – | 24028 | Art. 61 |
| 13 | Ensure safe system decommissioning.<br>(e.g., ensure decommissioned data is either deleted or restricted to authorized personnel.) | $P_3$ | $R_E$ | – | 38507, 24368 | Art. 9 |
| 14 | Redocument model information and contractual requirements at every system update.<br>(e.g., update the model information when re-training the system or using datasets with new contractual requirements) | $P_3$ | $R_E$ | – | 23894, 5338, 24368 | Art. 11, 12, 17, 61 |
| **DATA** | | | | | | |
| 15 | Ensure compliance with agreements and legal requirements when handling data.<br>(e.g., create data sharing and non-disclosure agreements and secure servers) | $P_{1-3}$ | $R_{D,E,M}$ | – | 38507, 23894, 5338 | Art. 10, 17, 61 |
| 16 | Compare the quality, representativeness, and fit of training and testing datasets with the intended uses.<br>(e.g., report dataset details such as public/private, personal information, demographics, and data provenance) | $P_{1-3}$ | $R_E$ | [10, 34, 43, 96] [59, 67, 95] | 38507, 5338, 24028, 24027 | Art. 10, 13, 17, 64 |
| 17 | Identify any measurement errors in input data and their associated assumptions.<br>(e.g., account for potential input errors in the input device, text data, audio, and video) | $P_{1-3}$ | $R_E$ | [18] | 38507 | Art. 10, 13, 17, 64 |
| 18 | Protect sensitive variables in training/testing datasets.<br>(e.g., protect sensitive data and use techniques such as k-anonymity and differential privacy) | $P_{1-3}$ | $R_{D,E,M}$ | [24] | 38507, 24028 | Art. 10, 13, 17 |
| **OVERSIGHT** | | | | | | |
| 19 | Continuously monitor metrics and utilize guardrails or rollbacks to ensure the system's output stays within a desired range.<br>(e.g., validate against concept drift and test with diverse testers and compliance and adversarial cases) | $P_{1-3}$ | $R_{D,E}$ | [30] | 38507, 5338, 24028, 24027, 24368 | Art. 12, 20, 29, 61 |
| 20 | Ensure human control over the system, particularly for designers, developers, and end-users.<br>(e.g., include human in the loop with the ability to inspect data, models, and training methods) | $P_{1-3}$ | $R_{D,E,M}$ | – | 38507, 5338, 24028, 24368 | Art. 13, 14 |
| **TEAM** | | | | | | |
| 21 | Ensure team diversity.<br>(e.g., consider diversity in gender, neurotypes, personality traits, and thinking styles) | $P_{1-3}$ | $R_{D,E,M}$ | – | 38507, 5338, 24028, 24368 | Art. 69 |
| 22 | Train team members on ethical values and regulations.<br>(e.g., train on privacy regulations, ethical issues, and raising concerns) | $P_{1-3}$ | $R_{D,E,M}$ | [30] | 38507, 24368 | Art. 69 |

jargon. We also categorized each guideline into six thematically distinct categories, namely *intended uses*, *harms*, *system*, *data*, *oversight*, and *team*.

Recognizing that certain guidelines may only be applicable at specific stages (e.g., monitoring AI after deployment) by specific roles (e.g., developers, managers), we went through two steps. First, we assigned the guidelines to three phases based on previous research (e.g., [59, 66]). These phases are development (designing and coding the system), deployment (transferring the system into the production stage), and use (actual usage of the system). For example, guidelines like identifying the

system's intended uses (guideline #1) are relevant to all three phases, while those related to system updates (guideline #14) or decommissioning (guideline #13) are applicable during the use phase.

Second, we assigned the guidelines to three types of role (Table 1) based on previous literature. Wang et al. [97] interviewed UX practitioners and responsible AI experts to understand their work practices. UX practitioners included designers, researchers, and engineers, while responsible AI experts included ethics advisors and specialists. Wong et al. [100] analyzed 27 ethics toolkits to identify the intended audience of these toolkits, specifically those who are expected to engage in AI ethics work. The intended audience roles identified included software engineers, data scientists, designers, members of cross-functional or cross-disciplinary teams, risk or internal governance teams, C-level executives, and board members. Additionally, Madaio et al. [59] co-designed a fairness checklist with a diverse set of stakeholders, including product managers, data scientists and AI/ML engineers, designers, software engineers, researchers, and consultants. Following guidance therefore from these studies [59, 97, 100], we formulated three roles as follows:

(1) Designer: This role includes interaction designers and UX designers.
(2) Engineer or Researcher: This role includes AI/ML engineers, AI/ML researchers, data scientists, software engineers, UX engineers, and UX researchers.
(3) Manager or Executive: This role includes product managers, C-suite executives, ethics advisors/responsible AI consultants, and ethical board members.

The revised and final catalog, consisting of 22 unique guidelines, is presented in Table 1.

## 5 EVALUATION OF THE 22 RESPONSIBLE AI GUIDELINES

We first conducted a formative study with 10 AI practitioners from a large technology company to elicit design requirements for a tool that would allow us to evaluate the guidelines and then implemented this tool (Panel B in Figure 1 and §5.1). To evaluate the guidelines, we conducted a user study with 14 AI researchers, engineers, designers, and product managers from the same company (Panel C in Figure 1 and §5.2).

### 5.1 Incorporating the guidelines into a tool

**Eliciting design requirements for a tool through a formative study.** We conducted a formative study that included semi-structured interviews with 10 participants. These participants, comprising 6 males and 4 females, were AI practitioners in their 30s and 40s employed at a large technology company. The participants had a range of work experience, spanning from 1 to 8 years, and were skilled in areas such as data science, data visualization, UX design, natural language processing, and machine learning. The interview study took place online and consisted of three parts. In the first part, we encouraged participants to share information about their ongoing AI projects. In the second part, we presented them with the table containing the 22 guidelines and asked them to think about how each guideline could apply to their projects. Finally, in the third part, we conducted semi-structured interviews to discuss how these guidelines could be incorporated into an interactive responsible AI tool.

Each study lasted about half an hour. Two authors took notes during the interviews, and afterward, they analyzed the interview transcripts using inductive thematic analysis [13, 61, 64, 80]. This analysis then resulted in the following four design requirements (participant quotes are marked with FP):

*R1: Simplify the guidelines by breaking them into smaller visual components.* Participants found it challenging to reflect on guidelines and examples because of their quantity. According to FP5, "*the sheer number of the guidelines is the main difficulty [...] they should be separated in bite-sized*

*questions"*. Additionally, participants requested to visually separate the guidelines from the examples.

*R2: Implement clear navigation features to systematically guide users through the guidelines.* Participants were unsure about the best way to navigate through the guidelines. FP9 suggested that *"the system should provide clear navigation [...] for example, using a progress bar"*. FP5 further emphasized that the design of the progress bar could facilitate *"gaining insights while engaging with the 22 guidelines"*.

*R3: Track how guidelines are applied and share progress among team members.* Participants faced difficulty in tracking their responses on how to apply the guidelines to their projects and share progress among team members. To address this challenge, FP5 suggested implementing a feature that would save user responses as they progress through the guidelines: *"there should be some functionality there that captures the answers I gave, so it'd allow me to track progress and share it among team members"*. These responses would then be transformed into comprehensive documentation and made accessible to users for download.

*R4: Develop a mechanism for post-hoc reflections on how the project aligns with responsible AI guidelines.* Participants found it challenging to envision how well their AI systems aligned with the guidelines. Therefore, FP8 suggested developing *"visual feedback or a score that shows how responsible [their] AI system is."*. However, FP2 cautioned that this mechanism *"should not make me anxious and feel like I have not done enough"*. Instead, it should create a positive learning experience and encourage users to generate ideas for improving their AI systems.

**Designing the tool and incorporating the guidelines.** From these requirements, we designed an interactive web-based tool[4] (Figure 3) and populated it with the 22 guidelines in Table 1.

To meet the design requirement R1 (*Simplify the guidelines*), each guideline is presented as a digital card [57] with interactive boxes on both the front and back sides. The front side includes a symbolic graphic collage representing the guideline, followed by its name and full text. The back side includes an input box for users to write their thoughts on implementing each guideline in their project [82]. We also used this box to showcase an example for each guideline (refer to Figure 1). Initially, the example in the box is visible, but it disappears once the user inputs their specific implementation details. Users can view the guideline from both sides by using the flip buttons at the bottom-left corner of each side.

Each guideline is paired with two guiding questions [101] that help users think about the relevance of the guideline to their specific AI system and context (Figure 4). The first question asks the user whether the guideline has been successfully implemented in their AI system. For example, for an engineer addressing fairness, the question asks if they have reported evaluation metrics for various groups based on factors like age, gender, and ethnicity (technique #8 in Table 1). If the engineer answers "yes," they are then prompted to provide specific details on how fairness was implemented in the input box on the card's back. After sharing this information, the tool moves the guideline to the "successfully implemented" stack. In contrast, if the engineer answers "no," the tool asks a second follow-up question regarding whether the guideline should be implemented in a future iteration of the AI system. If the engineer answers "yes," they are prompted to provide specific details on how to implement it. The tool then moves the guideline to the "should be considered" stack. However, if the engineer answers "no" to both questions, indicating that the guideline is not applicable to their AI system, the tool moves the guideline to the "inapplicable" stack.

[4]anonymous-url


Fig. 3. Interactive Responsible AI Tool with 22 guidelines. The first part (A) allows for entering information about the developed AI system and (B) selecting the applicable user role. The second part (C) enables interaction with the guidelines. The third part (D) presents a summary of user responses for post-hoc reflections. Guidelines for other project phase can be viewed through the phase selectors (E/A).

To meet the design requirement R2 (*Implement clear navigation*), we explored different layout options and considered previous research that involved swiping [99], scrolling, or organizing guidelines into different groups [22]. Due to the limited screen size and the repetition of guidelines for each phase and role, we chose to organize the guidelines into nine groups. These groups were derived from three phases of the AI system: *i)* development (designing and coding), *ii)* deployment (transitioning into production), and *iii)* use (actual usage of the system), as well as from three user roles: *i)* designer, *ii)* engineer or researcher, and *iii)* manager or executive. The number of guidelines in each group varied and accommodated the specific requirements of each phase and

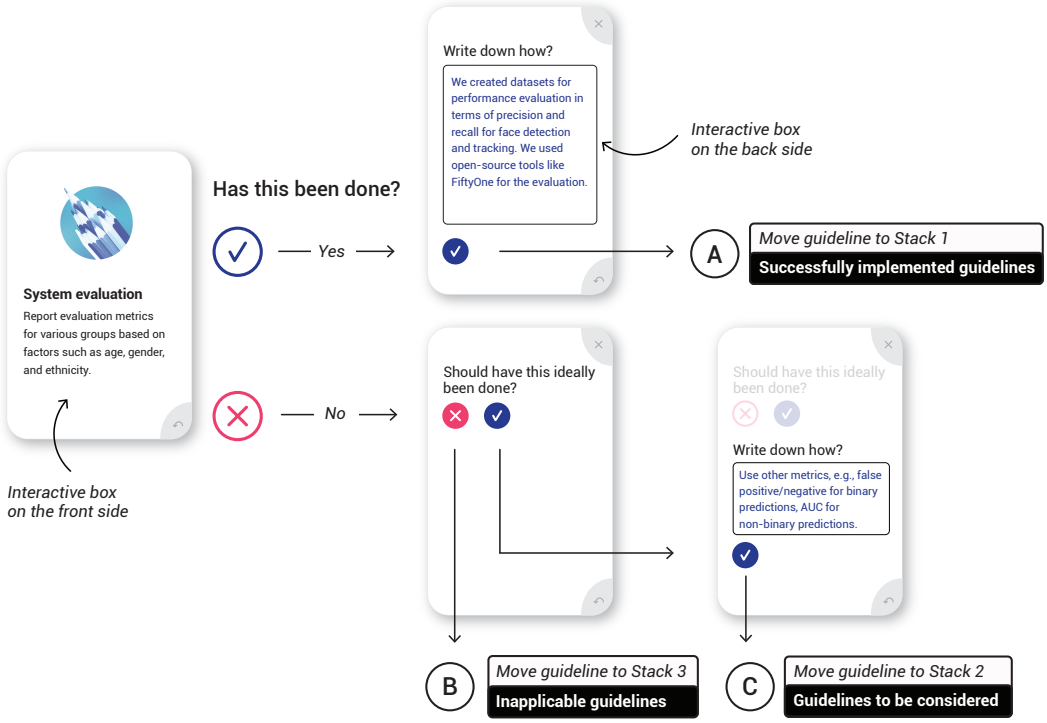


Fig. 4. Guideline sorting procedure. Users can place a guideline in any of the three stacks (i.e., successfully implemented, should be considered, inapplicable) by: (1) considering two guiding questions, and (2) using the Yes/No buttons located next to the card and on the back side of it.

role. For example, engineers or researchers needed to go through 20 guidelines for development, 18 for deployment, and 20 for use (§4, Step 4).

To facilitate the browsing of guidelines and addressing them in the user's preferred order, we introduced two arrow buttons on both the left and right side of the card group. We then introduced a graphical progress bar next to the group that not only displayed the number of remaining guidelines but also color-coded them to indicate their assignment to the three stacks. Blue leaves in the bar represented successfully implemented guidelines, magenta leaves represented guidelines for future consideration, and empty leaves represented inapplicable guidelines. Finally, we added an "Exit" button that becomes available as soon as the user goes through minimum two guidelines. In that way user can quit the experience at a preferable moment.

To meet the design requirement R3 (*Track how guidelines are applied and share progress among team members*), we added a feature to store user responses locally in the browser session. Users can download their responses as a structured PDF report at any time using the floating download button.

To fulfill the last requirement, R4 (*Develop a mechanism for post-hoc reflections*), after completing the sorting process, we display to the user a summary page. The summary is divided into three sections, one for each stack of cards (i.e., successfully implemented, should be considered, and inapplicable), with in-text counters indicating the number of guidelines in each stack. To read the responses for each guideline, hover-over functionality is provided.

Table 2. User study participants' demographics, including their job 'Role' (designer ($R_D$), engineer or researcher ($R_E$), and manager or executive ($R_M$)).

| ID | Gender | Yrs of expr. in AI | Education | Current continent | Expertise | Role |
|---|---|---|---|---|---|---|
| 1 | Male | 6 | Ph.D. | EU | Deep learning, computer vision | $R_M$ |
| 2 | Male | +10 | Ph.D. | North America | Machine learning, computer vision | $R_E$ |
| 3 | Male | 8 | Ph.D. | EU | Machine learning | $R_E$ |
| 4 | Male | 4 | Ph.D. | North America | Deep learning, IoT, computer vision | $R_E$ |
| 5 | Female | 5 | Ph.D. | EU | Machine learning | $R_D$ |
| 6 | Female | 8 | Ph.D. | EU | Computer vision | $R_D$ |
| 7 | Male | 2 | Ph.D. | North America | Computer vision | $R_E$ |
| 8 | Male | 10 | Ph.D. | EU | Machine learning | $R_M$ |
| 9 | Male | 4 | Ph.D. | North America | Computer vision | $R_E$ |
| 10 | Male | +10 | M.S. | EU | Machine learning, natural language processing | $R_E$ |
| 11 | Male | +10 | Ph.D. | EU | Machine learning | $R_M$ |
| 12 | Male | 6 | Ph.D. | EU | Machine learning | $R_E$ |
| 13 | Male | 4 | Ph.D. | EU | Reinforcement learning, decision making | $R_E$ |
| 14 | Male | 8 | Ph.D. | EU | Computer vision, robotics | $R_D$ |

Figure 3 shows the tool with its three parts that meet the four design requirements. The first part enables users to enter the name of the developed AI system (Figure 3A), select the phase it belongs to and specify user's role (Figure 3B). Once the phase and role are selected, the second part displays the guidelines (Figure 3C) and allows to address them one by one. The third part presents the user with the summary for post-hoc reflections (Figure 3D). If desired, the user can repeat the experience and generate documentation for other phases (Figure 3E).

### 5.2 Evaluating the Guidelines Through a User Study

To evaluate whether our guidelines are usable by different roles and whether they match the EU AI Act articles and ISO standards, we conducted a user study with 14 AI researchers, engineers, designers, and managers from a large technology company (Panel C in Figure 1).

**Participants.** We recruited 14 participants from the same large technology company.[5] The recruitment process took place in October and November 2022. We aimed for a balanced sample of participants, including a variety of roles such as researchers (5), designers (3), engineers (3), and managers (3). All participants had significant expertise in AI, including areas such as machine learning, deep learning, and computer vision. Additionally, each participant was actively involved in at least one ongoing AI project during the time of the interviews. Table 2 summarizes participants' demographics.

**Procedure.** Ahead of the interviews, we sent an email to all participants, providing a concise explanation of the study along with a brief demographics survey. The survey consisted of questions regarding participants' age, domain of expertise, role, and years of experience in AI system development. The survey is available in Appendix A. It is important to note that our organization[6] approved the study, and we adhered to established guidelines for user studies, ensuring that no personal identifiers were collected, personal information was removed, and the data remained accessible solely to the research team.

During the interview session, we presented to the participants two systems: (a) our tool with the 22 guidelines and (b) a web page with the checklist items from Microsoft's Fairness Checklist. We

[5]Participants who took part in the formative study were not eligible to participate in this evaluation study.
[6]REDACTED FOR BLIND REVIEW

used the Microsoft's AI Fairness Checklist as a baseline alternative because it is a published work in a human-computer interaction conference (CHI 2020), is freely available, and has a rigorous, transparent creation process.[7] We asked participants to interact with each system for 20 minutes (or less if finished sooner), alternating between them to avoid any learning effect. To make the scenario as realistic as possible, we encouraged participants to reflect on their ongoing AI projects and consider how the guidelines could be applied in their roles. We also presented them with excerpts from the EU AI Act articles [2] and summaries of each ISO standard (§4.3), and asked them whether the guidelines link to these articles and summaries. We further engaged participants by asking about their preferences, dislikes, and the relevance of the guidelines to their work. Subsequently, we administered the System Usability Scale (SUS) [14] to assess the usability of the guidelines and the checklist items.

We piloted our study with two researchers (1 female, 1 male), which helped us make minor changes to the study guide (e.g., clarifying question-wording and changing the order of questions for a better interview flow). These pilot interviews were not included in the analysis.

**Analysis.** First, we compared the two usability scores after using each system (i.e., the guidelines and the checklist items). Second, two authors conducted an inductive thematic analysis (bottom-up) of the interview transcripts, following established coding methodologies [61, 64, 80]. The transcripts included how the guidelines could be applied in the ongoing AI projects, how they link to the EU AI Act articles and ISO standards, and any other preferences or dislikes. The authors used sticky notes on the Miro platform [65] to capture the participants' answers, and collaboratively created affinity diagrams based on these notes. They held seven meetings, totaling 14 hours, to discuss and resolve any disagreements that arose during the analysis process. Feedback from the last author was sought during these meetings. In some cases, a single note was relevant to multiple themes, leading to overlap between themes. All themes included quotes from at least two participants, indicating that data saturation had been achieved [39]. As a result, participant recruitment was concluded after the 14$^{th}$ interview.

**Results.** Participants, on average, rated the guidelines' usability with a score of 66 out of 100 in SUS, with a standard deviation of 16.01 (Figure 5). This indicates a generally positive user experience [83]. The moderately high usability score can likely be attributed to factors such as familiarity and efficiency [55] in interactions with the guidelines, further supporting their usability across different roles. In contrast, participants, on average, rated the checklist items' usability with a score of 44 out of 100 is SUS, with a standard deviation of 21.16. Despite the comparative lower SUS score, checklist items were seen as relevant for audit, formal processes, and certification purposes—acting as a 'safeguard'. As for the thematic analysis, the resulting themes are provided in Table 3 in the Appendix. These themes pertain to how our participants saw the application of guidelines, what worked well and what could have been improved.

Guidelines were generally well-received by the participants. The majority of them (12 out of 14 participants) considered the guidelines valuable for raising awareness and facilitating self-learning about responsible AI, though to different extents. Participants found the set of guidelines to be comprehensive and aligned with their roles (10 out of 14 participants), as evidenced by P8's observation that *"There are some aspects of responsible AI in the project that I knew about, but I never faced them in such an organized manner"*. Similarly, P4 *"felt that the guidelines were concrete and well-scoped, instead of the lengthy documents of current regulations."* Participants also stated that the

[7]We decided not to compare our tool with existing card-based systems for responsible AI as they serve different purposes. Card-based systems such as the IDEO AI Ethics and the Feminist Tech card aim at providing thought-provoking activities [46] and stimulating ethical conversations [54]. However, they cannot be used as tools ensuring compliance with internal ethical procedures (like Microsoft's AI Fairness Checklist) or ISO standards.

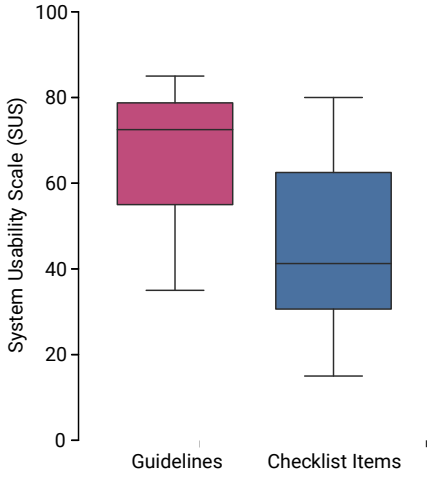


Fig. 5. SUS (usability) results. Guidelines are more usable than checklist items.

guidelines align with current regulations (10 out of 14 participants). P7 mentioned that *"he could understand the guidelines relevancy to the ISO standards and their applicability to his work."* Similarly, P11 found the excerpts from the EU AI Act *"relevant and the guidelines helped him to reflect how the current regulations will affect his project"*. Additionally, seven participants acknowledged the usefulness of the provided examples, which helped them think about potential scenarios and make the guidelines more actionable. One participant expressed that *"the guidelines made me reflect on my previous choices and how I would describe my decisions when I had to develop the system (P3)."* P3 specifically mentioned four guidelines about data (listed 15-18 in the Table 1 in the manuscript) and thought of the following improvements: *"I would collect more annotated data from diverse populations and incentivize underrepresented groups to participate in data annotation"*. An action plan was also devised by P1, who recognized that, *"I need an expert in different areas of assessments, because I am probably not in the right position to do that,"* and they intended to consult system security and sustainability experts. Finally, after becoming familiar with the guidelines, P2 felt more empowered to introduce the topic of responsible system development during group discussions with his team, stating that *"I can at least raise a few questions during team discussions—these are some additional aspects we may need to consider."*

## 6 DISCUSSION

To assist AI practitioners in navigating the rapidly evolving landscape of AI ethics, governance, and regulations, we have developed a method for generating responsible AI guidelines that are grounded in regulation and usable by different roles. We validated our method in a user study at a large technology company, where we designed and evaluated a tool that incorporates our responsible AI guidelines. We conducted a formative study involving 10 AI practitioners to design the tool, and evaluated our guidelines in a user study with an additional 14 AI practitioners. The results indicate that the guidelines were perceived as practical and actionable, promoting self-reflection and enhancing understanding of the ethical considerations associated with AI during the early stages of development. In light of these results, we discuss how our method contributes to the idea of "Responsible AI by Design", that is, a design-first approach that considers responsible AI values throughout the development lifecycle and across business roles. We now discuss the inherent problem of decontextualization in responsible AI toolkits, the concept of meta-responsibility, and

provide practical recommendations for incorporating responsible AI guidelines into toolkits and recommendations for technical and non-technical roles in enabling organizational accountability.

### 6.1 Theoretical Implications

**Decontextualization.** The inherent challenge in responsible AI toolkits lies in their attempt to reconcile the tension between scalability and context specificity [100]. Traditional approaches to toolkit development have often favored a universal, top-down approach that assumes a one-size-fits-all solution [49, 60]. However, participatory development, such as the methodology we followed in designing and populating a responsible AI tool with our guidelines, emphasizes the importance of tailoring responsible AI guidelines to specific contexts and job roles needs. It is crucial therefore to recognize that different AI practitioners, such as designers, developers, engineers, and executives, have distinct requirements and considerations that cannot be treated as identical. This highlights the complexity of developing toolkits that cater to a diverse range of practitioners while accounting for their unique roles and settings—the problem of decontextualization in responsible AI toolkits [100].

To tackle the problem of decontextualization, our proposed method incorporates two key elements: *guidelines usable by different roles* and *guiding questions*. Firstly, the integration of guidelines, tailored to different roles and projects, provides practical steps and recommendations that technical practitioners can easily implement, or C-level executives can make informed decisions upon. These guidelines serve as a starting point for ethical decision-making throughout the AI lifecycle, contributing to the vision of responsible AI by design (borrowing from the idea of 'privacy by design'[8]). Secondly, the inclusion of the two guiding questions (§5.1) enhances our toolkit's ability to capture the complexities of different social and organizational contexts. Expanding upon the concept that guiding questions are an effective means of communication [98], as they help in gaining deeper insights, clarifying responses, and uncovering underlying meanings, AI practitioners can engage with these questions to explore the ethical considerations and challenges that are unique to their deployment context.

**Meta-responsibility** Scholars have long recognized the need for a socio-technical approach that considers the contextual factors governing the use of AI systems, including social, organizational, and cultural factors [93]. In fact, Ackerman [1] introduced the concept of socio-technical gap to highlight the disparity between human requirements in technology deployment contexts (socio-requirements) and the technical solutions. This gap arises due to the flexible and nuanced nature of human activity compared to the rigid and brittle nature of computational mechanisms, resulting from necessary formalization and abstraction. Along these lines, Stahl [89] introduced the concept of meta-responsibility to stress that AI systems should be viewed as systems of systems (ecosystems) rather than single entities. To establish a regime of meta-responsibility, Stahl argued for an adaptive governance structure to effectively respond to new insights and external influences (e.g., upcoming AI regulation), and for a knowledge base that equips AI stakeholders with technical, ethical, legal, and social understanding. By integrating ethical, legal, and social knowledge into the AI development process—what Stahl referred to as adaptive governance structure, our work contributes to this line of research by providing empirical evidence to it and pushing the theoretical boundaries further.

[8]"Privacy by design" is a standard practice for incorporating data protection into the design of technology. In other words, data protection is achieved when it is already integrated into the technology during its design and development [17].

### 6.2 Practical Implications

**Recommendations for incorporating responsible AI guidelines into toolkits.** Our work identified four essential design requirements for incorporating guidelines into tools. These requirements are simplifying guidelines into smaller visual components; implementing clear navigation through the guidelines; tracking and sharing progress how guidelines are applied; and developing mechanisms for reflection on the guidelines.

For simplifying guidelines, we displayed each guideline as a digital card and accompanied it with two guiding questions. Future work could explore how to further divide guidelines into additional visual elements on the cards and how to refine the guiding questions. For example, guideline #15—*ensuring compliance with agreements and legal requirements when handling data*—could be further divided into step-by-step processes, with each one marked by a visual element like a card tab or a link to a specific ISO, or excerpts from the EU AI Act. Regarding the guiding questions, we observed that formulating them is a delicate task, requiring a balance between directness and respect for the user's autonomy. For example, a question formulated as "How did you consider the potential impact of your AI system on different user groups?" employs a proactive stance, avoiding any direct accusation or presumption of oversight. This method resonates with the experiences of our participants (e.g., P14), who found value in open-ended questions. However, guiding questions can be refined in various ways. For example, the *confront* type of question can incorporate the elements of "reminding consequences" and "providing multiple viewpoints," encouraging users to consider alternative directions and diverse perspectives [16].

For ensuring clear navigation, we organized the guidelines into a one-page layout and incorporated multiple buttons along with a counter for easy navigation. Future work could explore how to develop alternative layouts and include different navigation mechanisms. For example, complementary guidelines with related content, such as guideline #3 *identify potential harms and risks associated with the intended use* and guideline #5 *develop strategies to mitigate identified harms or risks for each intended use* can be paired side by side to improve the quality of responses. Additionally, new navigation mechanisms might include a chart to illustrate the relationships between guidelines and a search bar to enable users to quickly locate specific guidelines.

For tracking how guidelines are applied and sharing progress among team members, we introduced a feature to store user responses locally within the browser session and dynamically generate a PDF report from these responses. Future work could explore how to structure user responses in formats suitable for automated analysis and integration with other tools. For instance, using JSON format as input for machine learning algorithms and Large Language Models (LLMs) can enable the analysis of user responses and the generation of automated insights and recommendations within the PDF report.

For enabling post-hoc reflections, we created a summary page where users can view the number of guidelines they have considered and their responses to each guideline. Future work could explore how to improve this summary page, for example by adding visual elements for recognizing responsible AI champions (e.g., responsible AI badges) and fostering empathy (e.g., animations presenting the environmental impact of an AI system) or by implementing a collaborative aspect where users can share and discuss their summary pages with peers or mentors.

**Recommendations for enabling organizational accountability.** While individual adoption of responsible AI best practices is crucial, fostering effective communication between technical and non-technical roles is equally important. Many existing responsible AI toolkits prioritize individual usage [100]. However, addressing complex ethical and societal challenges associated with AI systems requires diverse perspectives. Our interactive tool populated with guidelines addresses this need by offering features (e.g., adjusting number of guidelines on the fly to different

roles and system phases) that make the guidelines usable by different roles. However, our tool can further improve communication between roles by creating a knowledge base of responses. Such a knowledge base, according to Stahl [89], empowers team members to fulfill their responsibilities and supports distributed teams in constructing a shared understanding of their AI system. Furthermore, we suggest a mechanism for keeping this knowledge base up to date and enriched with diverse perspectives. This includes regularly revisiting the guidelines through our tool and providing responses at key project milestones, such as when the AI system enters a new phase or when the team communicates with external members. This approach ensures that the knowledge base remains dynamic and reflect the evolving insights and perspectives within the team.

Our guidelines and the tool that incorporates them can also be used to enable organizational accountability. Similar to Google's five-stage internal algorithmic auditing framework [78], our guidelines serve as a practical tool for closing the AI accountability gap. The automatically generated report plays a crucial role in this process by providing a summary of the guidelines that were effectively implemented, those that should be considered for future development, and the non-applicable ones. These reports establish an additional chain of accountability that can be shared with stakeholders at various levels, including managers, senior leadership, and AI engineers. By offering more oversight and the ability to troubleshoot if needed, these reports help mitigate unintentional harm. However, it is important to note that when an organization adopts our guidelines, it should establish clear ethical guidelines for their intended uses. Our tool is not intended to discourage AI developers, engineers, designers, and managers from using it due to the fear of being held accountable for their responses. On the contrary, their responses, as documented in the report, provide an opportunity to identify potential ethical issues and address them early in the design stages. This proactive approach prevents the need for post-hoc fixes and repairs, aligning with the principle of addressing ethical considerations during the development process rather than as an afterthought [81]—the idea of *Responsible AI by Design*.

### 6.3 Limitations and Future Work

Our work has four main limitations that highlight the need for future research efforts. Firstly, although we followed a rigorous four-step process involving multiple stakeholders, the list of 22 guidelines may not be exhaustive. The rapidly evolving nature of AI ethics, governance, and regulations necessitates an ongoing effort to stay abreast of emerging developments. However, one of the strengths of our method lies in its modular design, which allows for ongoing refinement and expansion of the set of guidelines. This ensures that our guidelines maintain their relevance and stay up to date in the ever-evolving landscape of AI ethics, governance, regulations, and ISOs (e.g., functional safety (ISO 5469), data quality (ISO 5259), explainability (ISO 6254), AI management system (ISO 42001)). However, we acknowledge that there may be limitations in ensuring that all standards are accessible to everyone and that experts may not always be available to evaluate them. A partial solution would be to create forums or discussion groups where individuals can share their experiences and insights about regulations and standards. At the same time, future research could also investigate the frequency with which our method should be updated as new literature emerges. One possibility would be to create an automated system that regularly collects research articles on responsible AI best practices, pairing them with current and upcoming regulations, to extract new guidelines.

Secondly, it is important to consider the qualitative nature of our user study, which involved in-depth interviews and analysis of participants' responses. The findings from this study should be interpreted with caution though, understanding that the reported frequency of themes should be viewed in a comparative context rather than taken at face value [31]. This approach helps to avoid potential misinterpretation or overgeneralization of the results.

Thirdly, we need to acknowledge the limitations associated with the sample size and demographics of our user study. The study was conducted with a specific group of participants, and therefore, the findings may not fully represent the practices and perspectives of all AI practitioners. Our sample predominantly consisted of male participants, which aligns with the gender distribution reported in Stack Overflow's 2022 Developer Survey, where 92.85% of professional developer respondents identified as male [76]. Additionally, our participants were drawn from a large research-focused technology company. While the results may offer insights into practices within certain companies, they serve as a case study for future research.

Lastly, our qualitative data suggests indicators of ease of use for AI practitioners but does not provide direct information on the actual effectiveness of the guidelines. Understanding the impact of guidelines (or other AI toolkits [100]) requires long-term studies that consider multiple projects, with some utilizing the toolkit and others not. One potential avenue, as suggested by clinical researchers developing deep learning tools for patient care [9], is to conduct observational studies with users of the AI system to assess its performance. Another approach is to use proxies, such as measuring users' attitudes, beliefs, and mindset regarding ethical values before and after utilizing the guidelines. We intend to explore these directions in future research.

## 7 CONCLUSION

We proposed a method for generating a list of responsible AI guidelines that are grounded in regulations and are usable by different roles. The resulting 22 guidelines were integrated into an interactive tool and evaluated through a user study with 14 AI researchers, engineers, designers, and managers from a large technology company. Our participants found the guidelines well-aligned with their roles, enabling them to communicate complex ethical concepts in a structured manner. The guidelines are also grounded in ISOs and the EU AI Act articles, receiving positive feedback for being comprehensive. The usefulness of examples in guidelines was particularly noted as they enabled participants to reflect on their choices concerning ethical issues. As these guidelines are likely to become part of future responsible AI toolkits, it is important to implement features that provide users with time and space for reflection. Additionally, these toolkits should take users' reflections and roles into account to offer actionable recommendations tailored to a specific project, using, for example, large language models.

## REFERENCES

[1] Mark S. Ackerman. 2000. The Intellectual Challenge of CSCW: The Gap Between Social Requirements and Technical Feasibility. *Human–Computer Interaction* 15, 2-3 (2000), 179–203. https://doi.org/10.1207/S15327051HCI1523_5

[2] EU AI Act. 2021. *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts.* Retrieved December 2023 from https://artificialintelligenceact.eu/the-act/

[3] K.K. Aggarwal and Yogesh Singh. 2008. *Software engineering.* New Age International.

[4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI).* ACM, 1–13. https://doi.org/10.1145/3290605.3300233

[5] Alejandro B. Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[6] Vijay Arya, Rachel K.E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One Explanation Does Not Fit All: A Toolkit And Taxonomy Of AI Explainability Techniques. arXiv:1909.03012

[7] Ricardo Baeza-Yates. 2018. Bias on the Web. *Commun. ACM* 61, 6 (2018), 54–61. https://doi.org/10.1145/3209581

[8] Agathe Balayn, Mireia Yurrita, Jie Yang, and Ujwal Gadiraju. 2023. "Fairness Toolkits, A Checkbox Culture?" On the Factors that Fragment Developer Practices in Handling Algorithmic Harms. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. 482–495. https://doi.org/10.1145/3600211.3604674

[9] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 1–12. https://doi.org/10.1145/3313831.3376718

[10] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604. https://doi.org/10.1162/tacl_a_00041

[11] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. *Fairlearn: A toolkit for assessing and improving fairness in AI.* Technical Report. Microsoft. Retrieved December 2022 from https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

[12] Michael Boone, Nikki Pope, Chaowei Xiao, and Anima Anandkumar. 2022. *Enhancing AI Transparency and Ethical Considerations with Model Card++*. Nvidia. Retrieved November 2022 from https://developer.nvidia.com/blog/enhancing-ai-transparency-and-ethical-considerations-with-model-card/

[13] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[14] John Brooke. 1996. SUS: A "Quick and Dirty" Usability Scale. In *Usability Evaluation In Industry*, Patrick W. Jordan, B. Thomas, Ian L. McClelland, and Bernard Weerdmeester (Eds.). CRC Press, Chapter 12, 107–114.

[15] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. Proceedings of Machine Learning Research, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[16] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 1–15. https://doi.org/10.1145/3290605.3300733

[17] Ann Cavoukian. 2009. *Privacy by Design: The 7 Foundational Principles.* Information & Privacy Commissioner of Ontario, Canada. Retrieved November 2022 from https://iab.org/wp-content/IAB-uploads/2011/03/fred_carter.pdf

[18] Gary S. Collins, Johannes B. Reitsma, Douglas G. Altman, and Karel G.M. Moons. 2015. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Journal of British Surgery* 102, 3 (2015), 148–158. https://doi.org/10.1161/CIRCULATIONAHA.114.014508

[19] Henriette Cramer, Jean Garcia-Gathright, Sravana Reddy, Aaron Springer, and Romain Takeo Bouyer. 2019. Translation, tracks & data: an algorithmic bias effort in practice. In *Extended Abstracts of the ACM CHI Conference on Human Factors in Computing Systems (CHI)*. 1–8. https://doi.org/10.1145/3290607.3299057

[20] Wesley H. Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei S. Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring how machine learning practitioners (try to) use fairness toolkits. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. 473–484. https://doi.org/10.1145/3531146.3533113

[21] Wesley H. Deng, Nur Yildirim, Monica Chang, Motahhare Eslami, Kenneth Holstein, and Michael Madaio. 2023. Investigating Practices and Opportunities for Cross-functional Collaboration around AI Fairness in Industry Practice. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. 705–716. https://doi.org/10.1145/3593013.3594037

[22] Martin Dittus, Luca M. Aiello, and Daniele Quercia. 2017. Community Engagement Triage: Lightweight Prompts for Systematic Reviews. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–22. https://doi.org/10.1145/3134674

[23] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. 67–73. https://doi.org/10.1145/3278721.3278729

[24] Cynthia Dwork. 2008. Differential Privacy: A Survey of Results. In *Theory and Applications of Models of Computation*. Springer, 1–19. https://doi.org/10.1007/978-3-540-79228-4_1

[25] Upol Ehsan and Mark O. Riedl. 2020. Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. In *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence*, Constantine Stephanidis, Masaaki Kurosu, Helmut Degen, and Lauren Reinerman-Jones (Eds.). Springer International Publishing, 449–466.

[26] Salma Elsayed-Ali, Sara E. Berger, Vagner F. De Santana, and Juana Catalina Becerra Sandoval. 2023. Responsible & Inclusive Cards: An Online Card Tool to Promote Critical Reflection in Technology Industry Work Practices. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI)*. 1–14. https://doi.org/10.1145/3544548.3580771

[27] Equal Employment Opportunity Commission. 1977. *Prohibited Employment Policies/Practices.* Retrieved May 2023 from https://www.eeoc.gov/prohibited-employment-policiespractices

[28] European Union. 2018. *General Data Protection Regulation.* Retrieved May 2023 from https://gdpr-info.eu/

[29] Fairlearn. 2022. *Improve fairness of AI systems.* Retrieved November 2022 from https://fairlearn.org

[30] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. *Berkman Klein Center Research Publication* 2020-1 (2020). https://doi.org/10.2139/ssrn.3518482

[31] Ellie Fossey, Carol Harvey, Fiona McDermott, and Larry Davidson. 2002. Understanding and Evaluating Qualitative Research. *Australian & New Zealand Journal of Psychiatry* 36, 6 (2002), 717–732. https://doi.org/10.1046/j.1440-1614.2002.01100.x

[32] Ulrike Franke. 2021. *Artificial divide: How Europe and America could clash over AI.* European Council on Foreign Relations. Retrieved May 2023 from https://ecfr.eu/publication/artificial-divide-how-europe-and-america-could-clash-over-ai/

[33] Hana Frluckaj, Laura Dabbish, David G. Widder, Huilian Sophie Qiu, and James Herbsleb. 2022. Gender and Participation in Open Source Software Development. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–31. https://doi.org/10.1145/3555190

[34] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (2021), 86–92. https://doi.org/10.1145/3458723

[35] Brent Gleeson. 2013. *The Silo Mentality: How To Break Down The Barriers.* Retrieved December 2023 from https://www.forbes.com/sites/brentgleeson/2013/10/02/the-silo-mentality-how-to-break-down-the-barriers/

[36] Google. 2022. *AI Explorables.* Retrieved November 2022 from https://pair.withgoogle.com/explorables/

[37] Google. 2022. *Fairness Indicators.* Retrieved November 2022 from https://github.com/tensorflow/fairness-indicators

[38] Government Equalities Office and Equality and Human Rights Commission. 2010. *Equality Act 2010: guidance* . Retrieved May 2023 from https://www.gov.uk/guidance/equality-act-2010-guidance

[39] Greg Guest, Arwen Bunce, and Laura Johnson. 2006. How many interviews are enough? An experiment with data saturation and variability. *Field Methods* 18, 1 (2006), 59–82. https://doi.org/10.1177/1525822X05279903

[40] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. *Science Robotics* 4, 37 (2019). https://doi.org/10.1126/scirobotics.aay7120

[41] Karen Hao. 2019. Training a single AI model can emit as much carbon as five cars in their lifetimes. *MIT technology Review* (2019). https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/

[42] Lucy Havens, Melissa Terras, Benjamin Bach, and Beatrice Alex. 2020. Situated Data, Situated Systems: A Methodology to Engage with Power Relations in Natural Language Processing Research. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, 107–124. https://aclanthology.org/2020.gebnlp-1.10

[43] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2020. The Dataset Nutrition Label: A Framework to Drive Higher Data Quality Standards. In *Data Protection and Privacy*, Dara Hallinan, Ronald Leenes, Serge Gutwirth, and Paul De Hert (Eds.). Hart Publishing, Chapter 1, 1–26. https://doi.org/10.5040/9781509932771.ch-001

[44] The White House. 2023. *Blueprint for an AI Bill of Rights* . Retrieved May 2023 from https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[45] IBM. 2022. *AI Fairness 360.* Retrieved November 2022 from https://aif360.mybluemix.net

[46] IDEO. 2019. *AI needs ethical compass. This tool can help.* Retrieved November 2022 from https://www.ideo.com/blog/ai-needs-an-ethical-compass-this-tool-can-help

[47] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399. https://doi.org/10.1038/s42256-019-0088-2

[48] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI)*. 1–14. https://doi.org/10.1145/3313831.3376219

[49] Christopher M. Kelty. 2018. *The Participatory Development Toolkit.* Retrieved December 2023 from https://limn.it/articles/the-participatory-development-toolkit/

[50] Henry Kissinger, Eric Schmidt, and Daniel P. Huttenlocher. 2021. *The age of AI: And our human future.* John Murray London.

[51] Knowledge Centre Data and Society. 2019. AI Blindspots Card Set 1.0. Retrieved February 2023 from https://data-en-maatschappij.ai/en/tools/ai-blindspot

[52] Knowledge Centre Data and Society. 2019. AI Blindspots Card Set 2,0. Retrieved February 2023 from https://data-en-maatschappij.ai/en/tools/ai-blindspots-2.0

[53] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the ACM Conference on Intelligent User Interfaces (IUI)*. 126–137. https://doi.org/10.1145/2678025.2701399

[54] Superrr Lab. 2022. *The Feminist Tech Card Deck*. Retrieved November 2022 from https://superrr.net/feministtech/deck

[55] James R. Lewis, Brian S. Utesch, and Deborah E. Maher. 2015. Measuring Perceived Usability: The SUS, UMUX-LITE, and AltUsability. *International Journal of Human-Computer Interaction* 31, 8 (2015), 496–505. https://doi.org/10.1080/10447318.2015.1064654

[56] Q. Vera Liao and Kush R. Varshney. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. arXiv:2110.10790

[57] Qinghua Lu, Liming Zhu, Xiwei Xu, Jon Whittle, Didar Zowghi, and Aurelie Jacquet. 2023. Responsible AI Pattern Catalogue: A Collection of Best Practices for AI Governance and Engineering. *Comput. Surveys* (2023). https://doi.org/10.1145/3626234

[58] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates Inc., 4768–4777. Retrieved August 2022 from https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[59] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI)*. 1–14. https://doi.org/10.1145/3313831.3376445

[60] Shannon Mattern. 2021. *Unboxing the Toolkit*. Retrieved December 2023 from https://tool-shed.org/unboxing-the-toolkit/

[61] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-Rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019). https://doi.org/10.1145/3359174

[62] Susan Michie, Charles Abraham, Martin P. Eccles, Jill J. Francis, Wendy Hardeman, and Marie Johnston. 2011. Strengthening evaluation and implementation by specifying components of behaviour change interventions: a study protocol. *Implementation Science* 6, 1 (2011), 1–8. https://doi.org/10.1186/1748-5908-6-10

[63] Susan Michie, Michelle Richardson, Marie Johnston, Charles Abraham, Jill J. Francis, Wendy Hardeman, Martin P. Eccles, James Cane, and Caroline Wood. 2013. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques. *Annals of Behavioral Medicine* 46, 1 (2013). https://doi.org/10.1007/s12160-013-9486-6

[64] Matthew Miles and Michael Huberman. 1994. *Qualitative Data Analysis: A Methods Sourcebook*. Sage.

[65] Miro. 2022. *Miro | Online Whiteboard for Visual Collaboration*. Retrieved August 2022 from https://miro.com/

[66] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa D. Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. ACM, 220–229. https://doi.org/10.1145/3287560.3287596

[67] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2018. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. arXiv:1811.07867

[68] Brent D. Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society* 3, 2 (2016). https://doi.org/10.1177/2053951716679679

[69] Interpret ML. 2019. *Interpret ML*. Retrieved August 2022 from https://interpret.ml/

[70] Aleksandra Mojsilovic. 2019. *Introducing AI Explainability 360*. IBM. Retrieved November 2022 from https://www.ibm.com/blogs/research/2019/08/ai-explainability-360/

[71] Nadia Nahar, Shurui Zhou, Grace Lewis, and Christian Kästner. 2022. Collaboration challenges in building ML-enabled systems: communication, documentation, engineering, and process. In *Proceedings of the ACM/IEEE International Conference on Software Engineering (ICSE)*. 413–425. https://doi.org/10.1145/3510003.3510209

[72] Arvind Narayanan. 2018. 21 Fairness definitions and their politics. In *Tutorial presented at the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

[73] National Institute of Standards and Technology. 2023. *AI Risk Management Framework*. Retrieved May 2023 from https://www.nist.gov/itl/ai-risk-management-framework

[74] OECD. 2023. *Catalogue of Tools & Metrics for Trustworthy AI*. Retrieved December 2023 from https://oecd.ai/en/catalogue/tools

[75] Equality of opportunity in supervised learning. 2016. Hardt, Moritz and Price, Eric and Srebro, Nati. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates Inc., 3323–3331. Retrieved August 2022 from https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

[76] Stack Overflow. 2022. *Stack Overflow Developer Survey 2022*. Retrieved November 2022 from https://survey.stackoverflow.co/2022/

[77] David Piorkowski, Soya Park, April Yi Wang, Dakuo Wang, Michael Muller, and Felix Portnoy. 2021. How AI developers overcome communication challenges in a multidisciplinary team: A case study. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–25. https://doi.org/10.1145/3449205

[78] Inioluwa D. Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. 33–44.

[79] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction* 5 (2021), 1–23. https://doi.org/10.1145/3449081

[80] Johnny Saldaña. 2015. *The Coding Manual for Qualitative Researchers*. Sage.

[81] Nithya Sambasivan and Jess Holbrook. 2018. Toward responsible AI for the next billion users. *Interactions* 26, 1 (2018), 68–71. https://doi.org/10.1145/3298735

[82] Conrad Sanderson, David Douglas, Qinghua Lu, Emma Schleiger, Jon Whittle, Justine Lacey, Glenn Newnham, Stefan Hajkowicz, Cathy Robinson, and David Hansen. 2023. AI ethics principles in practice: Perspectives of designers and developers. *IEEE Transactions on Technology and Society* (2023), 171–187. https://doi.org/10.1109/TTS.2023.3257303

[83] Jeff Sauro. 2011. *A practical guide to the system usability scale: Background, benchmarks & best practices*. Measuring Usability LLC.

[84] Peter Schaar. 2010. Privacy by Design. *Identity in the Information Society* 3, 2 (April 2010), 267–274. https://doi.org/10.1007/s12394-010-0055-x

[85] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. ACM, 59–68. https://doi.org/10.1145/3287560.3287598

[86] Or Sharir, Barak Peleg, and Yoav Shoham. 2020. The Cost of Training NLP Models: A Concise Overview. arXiv:2004.08900

[87] Ben Shneiderman. 2021. Responsible AI: Bridging from ethics to practice. *Commun. ACM* 64, 8 (2021), 32–35. https://doi.org/10.1145/3445973

[88] Ben Shneiderman. 2022. *Human-centered AI*. Oxford University Press.

[89] Bernd C. Stahl. 2023. Embedding responsibility in intelligent systems: from AI ethics to responsible AI ecosystems. *Scientific Reports* 13, 1 (2023), 7586. https://doi.org/10.1038/s41598-023-34622-w

[90] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2020. Energy and Policy Considerations for Modern Deep Learning Research. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 09 (2020), 13693–13696. https://doi.org/10.1609/aaai.v34i09.7123

[91] Hariharan Subramonyam, Jane Im, Colleen Seifert, and Eytan Adar. 2022. Solving separation-of-concerns problems in collaborative design of human-AI systems through leaky abstractions. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. 1–21. https://doi.org/10.1145/3491102.3517537

[92] Mohammad Tahaei, Marios Constantinides, Daniele Quercia, Sean Kennedy, Michael Muller, Simone Stumpf, Q. Vera Liao, Ricardo Baeza-Yates, Lora Aroyo, Jess Holbrook, et al. 2023. Human-Centered Responsible Artificial Intelligence: Current & Future Trends. In *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems (CHI)*. 1–4. https://doi.org/10.1145/3544549.3583178

[93] Mohammad Tahaei, Marios Constantinides, Daniele Quercia, and Michael Muller. 2023. A Systematic Literature Review of Human-Centered, Ethical, and Responsible AI. arXiv:2302.05284

[94] Ville Vakkuri, Kai-Kristian Kemell, Marianna Jantunen, Erika Halme, and Pekka Abrahamsson. 2021. ECCOLA — A method for implementing ethically aligned AI systems. *Journal of Systems and Software* 182 (2021), 111067. https://doi.org/10.1016/j.jss.2021.111067

[95] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the IEEE/ACM International Workshop on Software Fairness (FairWare)*. Association for Computing Machinery, 1–7.

[96] Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. 2022. REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets. *International Journal of Computer Vision* 130, 7, 1790–1810. https://doi.org/10.1007/s11263-022-01625-5

[97] Qiaosi Wang, Michael Madaio, Shaun Kane, Shivani Kapania, Michael Terry, and Lauren Wilcox. 2023. Designing Responsible AI: Adaptations of UX Practice to Meet Responsible AI Challenges. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. 1–16. https://doi.org/10.1145/3544548.3581278

[98] Harry Weger Jr, Gina Castle Bell, Elizabeth M. Minei, and Melissa C. Robinson. 2014. The relative effectiveness of active listening in initial interactions. *International Journal of Listening* 28, 1 (2014), 13–31. https://doi.org/10.1080/10904018.2013.813234

[99] Stefan Werning. 2020. Making data playable: a game co-creation method to promote creative data literacy. *Journal of Media Literacy Education* 12, 3 (2020), 88–101. https://doi.org/10.23860/JMLE-2020-12-3-8

[100] Richmond Y. Wong, Michael A. Madaio, and Nick Merrill. 2023. Seeing like a toolkit: How toolkits envision the work of AI ethics. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–27. https://doi.org/10.1145/3579621

[101] Nur Yildirim, Mahima Pushkarna, Nitesh Goyal, Martin Wattenberg, and Fernanda Viégas. 2023. Investigating How Practitioners Use Human-AI Guidelines: A Case Study on the People+ AI Guidebook. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. 1–13. https://doi.org/10.1145/3544548.3580900

[102] Amy X. Zhang, Michael Muller, and Dakuo Wang. 2020. How do data science workers collaborate? roles, workflows, and tools. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–23. https://doi.org/10.1145/3392826

## A ADDITIONAL MATERIALS FOR THE USER STUDY

- How old are you?
- What is your gender? [Male, Female, Non-binary, Prefer not to say, Open-ended option]
- How many years of experience do you have in AI systems?
- What's your educational background?
- In which country do you currently reside?
- What is domain or sector of your work? (e.g., health, energy, education, finance, technology, food)
- What is your current role?
- What kinds of AI systems do you work on? (e.g., machine learning, computer vision, NLP, game theory, robotics)

Table 3. Constructed themes for the user study based on how our participants saw the application of guidelines, what worked well and what could have been improved.

| Theme | Participants |
|---|---|
| Raising awareness, facilitating self-learning | 12 |
| Aligning with roles | 10 |
| Aligning with regulations | 10 |
| Providing helpful examples | 7 |
| Engaging team members and external experts | 5 |
| Maintaining the visual simplicity of the guidelines | 3 |
| Documenting guidelines in a concise summary PDF | 3 |
| Providing a systematic flow of information and guidelines | 2 |

## B MAPPING GUIDELINES WITH EU AI ACT ARTICLES

**Article 6 (*Classification rules for high-risk AI systems*):** It states that an AI system shall be considered high-risk when *"it [the AI system] is intended to be used as a safety component of a product, or is itself a product"*. This article aligns with **guideline #1** as it mandates the identification of an AI system's intended use to determine whether its use poses a low or high risk.

**Article 9 (*Risk management system*):** It states that *"a risk management system shall be established, implemented, documented and maintained throughout the entire lifecycle of a high-risk AI system"*. This article aligns with **guidelines #1, #3-5, and #13** as it is about the identification of harms and risks of the AI system's intended use.

**Article 10 (*Data and data governance*):** It states that *"training, validation and testing data sets shall be subject to appropriate data governance and management practices"*. This article aligns with **guidelines #8 and #15-18** as it discusses the management and quality of data for training, validation, and testing, including aspects of diversity and minimizing biases.

**Article 11 (*Technical documentation*):** It states that the technical documentation of a high-risk AI system shall *"be drawn up before that system is placed on the market or put into service and shall be kept up-to date"*, and *"provide national competent authorities and notified bodies with all the necessary information to assess the compliance of the AI system"*. This article aligns with **guidelines #2, #6, #14** as it about documentation of the system and its contractual requirements, which may also be needed for obtaining ethical approvals.

**Article 12 (*Record-keeping*):** It states that high-risk AI systems shall include *"logging capabilities to enable the monitoring of the operation of the high-risk AI system with respect to the occurrence of situations that may result in the AI system presenting a risk"*. This article aligns with **guidelines #6, #9, #10, and #14** as it is about providing mechanisms for interpretable outputs and auditing, and improving the security of the system.

**Article 13 (*Transparency and provision of information to users*):** It states that *"high-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately"*. This article aligns with **guidelines #8-10, #16-18, and #20** as it is about quality, representativeness, and fit of training and testing datasets with the intended use.

**Article 14 (*Human oversight*):** It states that *"high-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which the AI system is in use"*. This article aligns with **guidelines #9 and #20** as it about ensuring human control over the system.

**Article 15 (*Accuracy, robustness and cybersecurity*):** It states that *"high-risk AI systems shall be designed and developed in such a way that they achieve, in the light of their intended purpose, an appropriate level of accuracy, robustness and cybersecurity, and perform consistently in those respects throughout their lifecycle"*. This article aligns with **guideline #10** as it is about documenting the security of all system components.

**Article 16 (*Obligations of providers of high-risk AI systems*):** It states that *"providers of high-risk AI systems shall draw-up the technical documentation of the high-risk AI system"*. This article aligns with **guideline #6** as it is about system documentation.

**Article 17 (*Quality management system*):** It states that "an AI system shall be documented in a systematic and orderly manner in the form of written policies, procedures and instructions". This article aligns with **guidelines #6, #7, #10, and #14-18** because it is about documentation of all system components, including AI models and testing and validation procedures.

**Article 18 (*Obligation to draw up technical documentation*):** It states that *"providers of high-risk AI systems shall draw up the technical documentation ".* This article aligns with **guideline #6** as it is about system documentation.

**Article 20 (*Automatically generated logs*):** It states that *"providers of high-risk AI systems shall keep the logs automatically generated by their high-risk AI systems, to the extent such logs are under their control by virtue of a contractual arrangement with the user or otherwise by law".* This article aligns with **guideline #19** as it is about monitoring of the system.

**Article 29 (*Obligations of users of high-risk AI systems*):** It states that users shall *"monitor the operation of the high-risk AI system on the basis of the instructions of use."*, and *"inform the provider or distributor when they have identified any serious incident or any malfunctioning and interrupt the use of the AI system".* This article aligns with **guideline #19** as it about monitoring of the system and utilizing guardrails or rollbacks.

**Article 50 (*Document retention*):** It states that *"the provider shall, for a period ending 10 years after the AI system has been placed on the market or put into service, keep at the disposal of the national competent authorities the technical documentation".* This article aligns with **guideline #6** as it about system documentation.

**Article 60 (*EU database for stand-alone high-risk AI systems*):** It states that information contained in the EU database shall *"be accessible to the public"* and *"include the names and contact details of natural persons who are responsible for registering the system and have the legal authority to represent the provider".* This article aligns with **guideline #4** as it is about providing mechanisms for reporting system harms.

**Article 61 (*Post-market monitoring by providers and post-market monitoring plan for high-risk AI systems*):** It states that *"the post-market monitoring system shall actively and systematically collect, document and analyse relevant data provided by users or collected through other sources on the performance of high-risk AI systems throughout their lifetime".* This article aligns with **guidelines #12, #14, #15, #19** as it is about data handling and model updates when the AI system is in use.

**Article 62 (*Reporting of serious incidents and of malfunctioning*):** It states that *"providers of high-risk AI systems placed on the Union market shall report any serious incident or any malfunctioning of those systems which constitutes a breach of obligations under Union law intended to protect fundamental rights to the market surveillance authorities of the Member States where that incident or breach occurred".* This article aligns with **guideline #4** as it is about incentivizing the reporting of system harms.

**Article 63 (*Market surveillance and control of AI systems in the Union market*):** It states that *"the national supervisory authority shall report to the Commission on a regular basis the outcomes of relevant market surveillance activities. ".* This article aligns with **guideline #4** as it about incentivizing the reporting of system harms.

**Article 64 (*Access to data and documentation*):** It states that *"access to data and documentation in the context of their activities, the market surveillance authorities shall be granted full access to the training, validation and testing datasets used by the provider, including through application programming interfaces ('API') or other appropriate technical means and tools enabling remote access".* This article aligns with **guidelines #16 and #17** as it is about data documentation.

**Article 65 (*Procedure for dealing with AI systems presenting a risk at national level*):** It states that *"AI systems presenting a risk shall be understood as a product presenting a risk defined in Article 3, point 19 of Regulation (EU) 2019/1020 insofar as risks to the health or safety or to the protection of fundamental rights of persons are concerned".* This article aligns with **guideline #3** as it is about harms and risks identification.

**Article 67 (*Compliant AI systems which present a risk*):** It states that if the AI system is compliant with the EU AI Act but still presents a risk to the health or safety of persons, the market surveillance authority *"shall require the relevant operator to take all appropriate measures to ensure that the AI system concerned, when placed on the market or put into service, no longer presents that risk, to withdraw the AI system from the market or to recall it within a reasonable period, commensurate with the nature of the risk, as it may prescribe"*. This article aligns with **guideline #5** as it is about mitigation strategies about the identified harms and risks.

**Article 69 (*Codes of conduct*):** It states that *"the Commission and the Board shall encourage and facilitate the drawing up of codes of conduct intended to foster the voluntary application to AI systems of requirements related for example to environmental sustainability, accessibility for persons with a disability, stakeholders participation in the design and development of the AI systems and diversity of development teams on the basis of clear objectives and key performance indicators to measure the achievement of those objectives"*. This article aligns with **guidelines #2, #11, #21, #22** as it is about the environmental assessment of the system, the ethical approvals obtained from ethics committees and boards, and the characteristics of the development team.