Spring 7-10-2024

Responsible AI Principles: Findings from an Empirical Study on Practitioners' Perceptions

Pouria Akbarighatar
Department of Information Systems, University of Agder, Pouriaa@uia.no

Ilias O. Pappas
Department of Computer Science, Norwegian University of Science and Technology Department of Information Systems, University of Agder, ilias.pappas@uia.no

Polyxeni Vassilakopoulou
Department of Information Systems, University of Agder, polyxeni.vasilakopoulou@uia.no

Sandeep Purao
Department of Computer Information Systems, Bentley University Department of Information Systems, University of Agder, spurao@bentley.edu

Responsible AI Principles: Findings from an Empirical Study on Practitioners' Perceptions

Pouria Akbarighatar

Department of Information Systems, University of Agder

Ilias O. Pappas

Department of Computer Science, Norwegian University of Science and Technology

Department of Information Systems, University of Agder

Polyxeni Vassilakopoulou

Department of Information Systems, University of Agder

Sandeep Purao

Department of Computer Information Systems, Bentley University

Department of Information Systems, University of Agder

Abstract

As artificial intelligence (AI) continues its rapid evolution, ethical considerations become increasingly critical. This study presents an analytical approach to assessing the perceived importance, alignment, and implementation of Responsible AI (RAI) principles within organizations. An extensive survey collected insights from 82 AI experts across industries. The findings reveal clear patterns in how RAI principles are prioritized. Principles like privacy, security, reliability, and safety received the highest importance ratings, reflecting their status as foundational elements. Principles such as benevolence and non-maleficence were viewed as moderately important, while transparency, fairness, and inclusiveness were relatively lower priorities. This prioritization is also reflected in perceptions of alignment and implementation, with the higher-rated principles demonstrating stronger organizational alignment and operationalization. The results suggest that organizations may face challenges in effectively addressing certain RAI principles, potentially due to factors like varying expertise, resource constraints, and the complexity of translating text-based principles into concrete algorithmic implementations.

Keywords: Responsible AI, Principles, Operationalized RAI, Expert Perceptions

1. Introduction

Industries all over the world are adopting AI applications, which are extending into diverse fields like transportation, agriculture, healthcare, and security. For instance, AI is assisting with crop yield optimisation, diagnosing and treating illnesses, catching distracted drivers to improve road safety, detecting credit card fraud to protect finances, and identifying at-risk children to help provide support (Amugongo et al., 2023; Ho et al., 2019; Stilgoe, 2018; Van Esch et al., 2019; Wall, 2018). It is crucial to think about the ramifications and make sure that the development and application of AI proceed in a way that benefits both individuals and

society as these technologies continue to advance and become more integrated into our daily lives and key services and establish long-term sustainability (Clarke, 2019; Pappas et al., 2023; Vassilakopoulou et al., 2022).

While AI offers potential benefits, significant challenges also exist that must be addressed. There are risks of reinforcing unjust biases, violating privacy rights, and propagating false information online. Job displacement and reduced skills demand are also concerning (Mikalef et al., 2022). Additionally, mass surveillance, critical system failures involving autonomous technologies, and weapons applications pose risks. Even AI meant to help, like improving cybersecurity, could enable malicious uses such as cyberattacks if misapplied. These types of issues understandably cause public unease and raise valid questions about ensuring AI systems are responsible and appropriately managed (Akbarighatar et al., 2023b).

To prevent unintended negative consequences and foster positive outcomes in the deployment of AI systems and services to various stakeholders, both public and private sectors, as well as researchers, have proposed ethical and Responsible AI (RAI) principles (Clarke, 2019; Ess, 2009; Sojer et al., 2014). These principles, such as benevolence, non-malfunction, safety, and well-being, can guide organizations in their decision-making processes when implementing AI-driven technologies to achieve their strategic objectives (Mirbabaie et al., 2022). Incorporating these principles into strategic management and operationalizing them requires considering two perspectives. Firstly, it is imperative for managers and key personnel within organizations engaged in the development and deployment of AI systems to not only acknowledge the importance but also understand the synergistic nature of implementing and operationalizing these principles as a comprehensive system (which) (Akbari Ghatar et al., 2023a). Secondly, there is a need for clear and effective mechanisms (hows) that enable the practical application of these principles within the organization's processes and practices (Akbari Ghatar et al., 2023a; Whittlestone et al., 2019). Hence, in order to achieve the intended operationalization of the principles in the AI efforts, it is crucial to promote and operationalize the RAI principles to ensure alignment with these principles throughout the AI lifecycle and translate these principles into practices (Mittelstadt, 2019).

There appear to be three gaps in the research on operationalizing responsible or ethical AI, despite recent excellent research on operationalizing and translating the concepts into practices. First, the existing literature does not sufficiently explore or provide insights into how experts within organizations assess the relative importance of RAI principles (Vakkuri et al., 2019). The gap is related to limited empirical studies that directly investigate the views of these experts

of the relative importance of different principles across organizational contexts. Second, there is a gap in the understanding of how experts perceive the alignment of AI-infused initiatives with RAI principles (Munn, 2022). Previous research has not extensively examined the specific criteria or indicators that experts consider when evaluating this alignment. Additionally, there is limited research that delves into the potential challenges or barriers encountered in achieving alignment. Here, the focus is on assessing alignment from a higher, strategic standpoint, considering broader strategic factors. Finally, a gap in the literature exists regarding how experts perceive the operationalization of RAI principles within organizations (Morley et al., 2020). This gap is related to a scarcity of studies that provide in-depth insights into the practical implementation and integration of RAI principles into the daily operations and decision-making processes of organizations.

Our research is structured to provide evidence-based insights by conducting a survey gathered from AI experts who are actively involved in contributing to, managing, or consulting on AI initiatives. This approach allows us to gain a better understanding of how experts perceive the operationalization of RAI principles. To address this goal, we have framed three key research questions:

RQ1: How do experts perceive the relative importance of RAI principles in their organizations?
RQ2: How do experts perceive the alignment of AI-infused initiatives with RAI principles?
RQ3: How do experts perceive the operationalization of RAI principles in their organizations?

We expect to contribute to research in three areas. First, we extend knowledge of the existing literature by providing insights into how experts within organizations perceive the importance of RAI principles. Second, our research seeks to advance the understanding of how experts perceive the alignment of AI-infused initiatives with RAI principles. Third, we intend to contribute to the literature by shedding light on how experts perceive the operationalization of RAI principles within organizations.

The subsequent sections of the paper are structured as follows. In Section 2, we delve into the existing literature concerning responsible AI principles and discuss the journey from principles to practical application. Within this section, we also present a summary of the most important principles. In the upcoming sections, we delve into key aspects of our study. Section 3 details our data collection methods and analysis. In Section 4, we present empirical findings that directly address our research questions. Following these sections, our discussion section offers an extensive analysis of these results. It not only examines expert perspectives on RAI principles but also explores their theoretical and practical implications. We also discuss the limitations of our study and potential areas for future research.

2. Related Literature

2.1. Responsible AI principles in practice

It has been over five years since IS scholars initiated their investment in understanding how AI should be managed (Berente et al., 2021). Also, others highlighted the unintended consequences of the unethical use of AI and proposed some principles for being responsible AI. The AI4People recommendations, rooted in bioethical principles, serve as significant ethical guidelines in Western AI development. These principles, which encompass Autonomy, Beneficence, Non-Maleficence, Justice, and Explicability, have been adapted to address AI's unique challenges in healthcare. Specifically, transparency and explainability have been integrated into these recommendations. Transparency pertains to users' understanding of AI system development and functionality, while explainability focuses on the AI system's capacity to provide clear explanations for its decisions.

In practical literature, numerous inquiries also have taken place. A notable report, published by the Organization for Economic Co-operation and Development (OECD) in early 2019, stands out. This report synthesizes insights from over 70 documents that discuss ethical AI principles across various sectors. The documents originate from a range of sources, spanning industry players like Google, IBM, and Microsoft, governmental entities such as the Montreal Declaration and the Lords Select Committee, and academic institutions including the Future of Life Institute, IEEE, and AI4People. The standard comprises five complementary value-based principles: inclusive growth, fairness, transparency, security and safety, and accountability.

In a study that reviewed 84 ethical AI documents, the prevalent themes were transparency, justice and fairness, non-maleficence, responsibility, and privacy, each appearing in over 50% of cases (Jobin et al., 2019). Moreover, a systematic analysis of the ethical technology literature by (Royakkers et al., 2018) underscored recurring themes encompassing privacy, security, autonomy, justice, human dignity, technology control, and power equilibrium. As posited by these scholars, when considered collectively, these themes collectively 'define' ethically aligned machine learning as technology that is (a) beneficial and respectful towards individuals and the environment (beneficence); (b) resilient and secure (non-maleficence); (c) reflective of human values (autonomy); (d) fair (justice); and (e) transparent, accountable, and comprehensible (explicability).

When examining the European Commission’s High-Level Expert Group report's ethical principles, a consistent pattern emerges. The report outlines four ethical principles, deeply rooted in fundamental rights, that must be upheld to ensure the trustworthy development,

deployment, and use of AI systems. The first principle prioritizes respecting human autonomy and freedom (respect for human autonomy). The second emphasizes that systems should neither cause harm nor worsen existing issues for humans (prevention of harm). The third underscores the necessity for fairness throughout AI's lifecycle (fairness). Lastly, explicability proves essential for establishing and maintaining user trust in AI systems. This mandates transparent processes, clear communication of AI system capabilities and intentions, and comprehensible decisions for those directly and indirectly impacted. The absence of such information impedes the ability to challenge decisions effectively (explicability).

ISO 22989:2022 and ISO 24038 provide definitions and detailed explanations of the concept of trustworthiness, encompassing elements such as robustness, reliability, transparency, explainability, interpretability, accountability, safety, privacy, and fairness. All these concepts align with the categories established by OCED and the European Commission (2019). For instance, transparency, interpretability, expandability, and accountability, share a common goal from varying perspectives, reinforcing each other. Collectively, these principles advance AI systems' understandability. Additionally, principles connected to avoiding harm and positive impacts, such as safety, privacy, benevolence, and non-maleficence, uphold AI's beneficence nature. Similarly, fairness and inclusiveness aim to eradicate disparities, ensure equal opportunities, and prevent marginalization. The harmonious combination of Responsible AI principles contributes to a better understanding of RAI and how they synergistically work together (Akbarighatar et al., 2023c). By sharing common objectives, these principles support and reinforce each other, forming a cohesive framework for RAI. This means that the various principles of RAI complement and enhance one another, resulting in an integrated approach to RAI development and deployment concisely presented in Table 1, offering a holistic grasp of these pivotal principles.

While recent research has made significant contributions to the field of AI ethics, particularly in the exploration of duty ethics and virtue ethics within sociotechnical systems, there remains a need to further elucidate the interconnectedness of these ethical viewpoints. (Heyder et al., 2023) have provided a theoretical framework in this regard. In our research, our emphasis is on duty ethics, which involves establishing ethical principles to guide human behavior, specifically in our context—experts. While virtue ethics cultivate character duty ethics better suit the governance needs of organizations through organizational principles and policies. Duty and virtue ethics complement each other. Organizational principles and rules aimed at duties/obligations (duty ethics) can help shape an ethical culture and virtuous behavior

over time (virtue ethics). Our research, focusing on duty ethics, aims to contribute to the ongoing discourse on AI ethics in practice.

Principle	Literature descriptions	Refs
Benevolence and Non-maleficence	Indicate that AI technology is designed to promote good and maximize benefits, all the while avoiding harm and minimizing risks.	(European Commission., 2019; Microsoft AI, 2022; Clarke., 2019; Floridi et al., 2018).
Reliability and Safety	AI systems should aim to prevent failures and accidents ensuring intended performance.	(ISO:24028, 2020; Microsoft AI, 2022; Clarke., 2019)
Privacy	Freedom from intrusion into an individual's private life or affairs when it happens due to improper or illegal collection and use of their data.	(ISO:24028, 2020; Microsoft AI., 2022).
Security	Security refers to protecting data and controlling access based on authorization levels.	(ISO:24028, 2020; Microsoft AI., 2020).
Accountability	Accountability refers to taking responsibility, providing justifications for actions, responding to inquiries, and being liable.	(ISO:24028., 2020; Microsoft AI., 2022; Clarke., 2019)
Explainability	Explainability refers to providing comprehensive information about AI's inner workings.	(ISO:22989., 2020; Microsoft AI., 2022; Clarke., 2019)
Intelligibility	Intelligibility refers to enabling humans who use or manage AI to understand the reasoning of an AI system.	(ISO:24028., 2020)
Transparency	Transparency entails disclosing AI system details, like performance, limitations, components, measures, design goals, data sources, for a decision, prediction, or recommendation.	(IS:22989., 2020; Microsoft AI., 2022; Clarke., 2019; Floridi et al., 2018)
Inclusiveness	Inclusiveness refers to involving diverse individuals and perspectives, regardless of their unique circumstances.	(OECD., 2018; Microsoft AI., 2022)
Fairness	AI systems must be designed to ensure impartial treatment, and prevention of discriminatory outcomes.	(OECD., 2018; 2020; Microsoft AI., 2022; Clarke., 2019; Floridi et al., 2018)

Table 1. Responsible AI principles and their descriptions

3. Data and Research Methodology

3.1. Instrument development

To ensure the validity and robustness of the developed survey instrument, we followed the guidelines recommended by (MacKenzie et al., 2011). Our process began with the conceptualization of the constructs representing RAI principles in our study, as outlined in

Table 1. To evaluate the content validity of these principles, we engaged a panel of six experts with substantial academic and practical experience in responsible AI. Four of these experts had over 15 years of industry experience in data science and AI, while the remaining two were senior academics specializing in Information Systems in organizations. We provided the experts with definitions of each principle and asked them to answer the survey questions. Additionally, we sought their recommendations for improving or refining questions. Their feedback led to minor modifications and clarifications in the definitions, reinforcing the content validity of our instrument.

To assess convergent, discriminant, and nomological validity, we distributed the revised survey instrument to four C-level technology managers. These managers were selected from companies that specialize in the responsible development and deployment of AI and possess extensive experience in implementing RAI principles. Taking their valuable input into account, we carefully revised and refined the definitions of the principles to ensure they were more concise and understandable.

3.2. Data Collection

A 'survey' is a research method where experts in a specific field are queried about their views on relevant organizational factors (Rungtusanatham et al., 2003). Surveys enable a stronger connection between academia and the real world by testing conceptual models with real-world data (Flynn et al., 1990), making it a suitable approach for our current research.

Our survey targeted AI and machine-learning experts involved in AI solution development and integration as business enabler. The participants consisted of CEOs, managers, AI governance experts, and other relevant positions within these organizations. We identified and contacted potential respondents through professional groups on LinkedIn, such as the “Artificial Intelligence and Business Analytics” group, and the website “Ethical AI Database” to search for responsible AI or ethical AI companies and in general AI companies. This approach ensured a robust and representative sample for our study.

We reached out to selected respondents via email, specifically targeting those in high-level technology management roles who possessed knowledge of RAI operations and practices. Following an initial invitation and three subsequent reminders, each one week apart, we sent a total of 600 email invitations from September to October 2023 to potential participants experienced in AI-infused projects. From these invitations, we received a total of 82 complete and 13 incomplete ones, primarily due to respondents' unfamiliarity with certain initiatives. These responses came from various industries, including financial services, manufacturing, and

high-tech companies. The participants held a range of job titles, including head of AI or data science, chief data governance officers, directors of IT, co-founders, and chief data scientists.

4. Empirical results

4.1 Demographic data

In this study after giving short definitions of RAI principles and understanding them, we ask about the participant's perception of the importance, alignment, and operationalisation mechanisms and we use seven Likert points to measure them. To gain insights into participant demographics, we collect additional information such as their age, gender, professional background, years of experience, and organizational affiliations, allowing us to better understand the diverse perspectives within our participant pool. Across the total sample, the gender balance was 33% women 66% men, and 1% identified as non-binary or with other gender identities. 95 percent (80) of the respondents contributed, managed, or consulted to AI projects. The remained participants were excluded from the further analysis.

The sample (N=82) comprised professionals with significant experience in AI roles. The majority held graduate degrees, with 24% possessing a PhD and 56% a master's-level qualification. Over half (51%) had accrued more than 10 years of overall work experience. Regarding AI specialty, 33% reported 1-3 years spent in AI-related duties.

Geographically, Europe was most represented at 47% of respondents. North America accounted for 22% and the Australia/New Zealand region 21%. Participant organizations ranged in size, with 34% employed by large enterprises (>500 employees), and 48% by small-to-medium businesses (1<x<100). A diversity of industries was sampled, including 36%, 10%, and 8% from technology, healthcare, and finance respectively. Additional sample characteristics are provided in Table 2. This delineation by demographics, roles, sectors, and geographies offered a breadth of expert insights across the global AI landscape. Overall, the sample comprised knowledgeable professionals well-positioned to offer informed perspectives on organizational responsible AI strategy and implementation efforts.

Category	Percentage of respondents N=80	Category	Percentage of respondents N=80
Working experience		**AI-related Working experience**	
Fewer than one year	1,2%	Fewer than one year	3,7%
1-3 years	11,0%	1-3 years	32,9%
4-6 years	15,9%	4-6 years	32,9%
7-10 years	19,5%	7-10 years	17,1%
More than 10 years	52,4%	More than 10 years	13,4%
Familiarity with the concept of RAI		**Education level**	
Expert	22,0%	PhD	24,4%
Very familiar and involved in RAI practices actively.	34,0%	Master	58,5%
Very familiar and some involvement in RAI practices	22,0%	Bachelor	17,1%
Familiar but never involved in RAI practices	14,6%		
Heard about it - slightly familiar	7,3%		
Never heard about it	0,0%		

Table 2. Sample Characteristics

4.2 Expert Perceptions on Responsible AI Implementation

To explore the three research questions, before responding to the questions, participants were given a description of these principles, which can be found in Table 1. To address all research questions a 7-point Likert scale was used to measure experts' perceptions. Regarding the questionnaire's reliability assessment, we utilized SPSS software (version 29.0). The results exhibited strong Cronbach's alpha values of 0.919, 0.928, and 0.896 for importance, alignment, and operationalization, respectively. The total Cronbach's alpha value of 0.962 confirms the reliability of our questionnaire data.

- RAI Principles' Perceived Importance

To address the first research question, we inquired about the extent to which participants perceived their organization's adherence to responsible AI principles. As previously mentioned, the surveys used a 7-point Likert scale ranging from 1 to 7 to rate each principle. A value of 1 represented "Never" in terms of importance/implementation, while 7 represented "Always". Table 3 presents a summary of the observed minimum, maximum, average, and standard

deviation scores for various RAI principles. These metrics provide insights into the participants' perceptions of each principle's importance within their organizations.

Participants' responses revealed that principles like Reliability and Safety (average score: 5.976) and Privacy (6.167) were considered more important than principles such as Intelligibility (4.643) and Inclusiveness (4.506). The standard deviation scores further illustrate the variation in experts' responses, indicating how consistently each principle was rated. For example, while Benevolence and Non-maleficence received a high average importance rating of 5.548, it also had a relatively wide standard deviation of 1.5, suggesting diverse views on their importance. In contrast, Reliability and Safety had a narrower standard deviation of 1.202, indicating more consensus among experts. Overall, Table 2 provides a nuanced understanding of participants' prioritization of each RAI principle.

Responsible AI principles	Observed minimum	Observed maximum	Average	Standard Deviation
Benevolence and Non-maleficence	2	7	5.548	1.5
Reliability and Safety	2	7	5.976	1.202
Privacy	2	7	6.167	1.18
Security	2	7	6.012	1.247
Accountability	2	7	5.155	1.632
Explainability	2	7	4.952	1.693
Intelligibility	2	7	4.643	1.603
Transparency	2	7	4.833	1.642
Inclusiveness	1	7	4.506	1.87
Fairness	1	7	4.843	1.858

Table 3. Perceived Importance of Observed RAI Principles

- Experts' Views on AI Initiatives Alignment with RAI Principles

To address the second research question (RQ2), respondents were asked to rate on a 7-point Likert scale the degree of alignment between their organization's AI initiatives and responsible AI principles. As previously noted, the scale ranged from 1 to 7, with 1 representing "To a very little extent" and 7 being "To a great extent". Table 3 summarizes perceptions of alignment for various principles. The averages provide insight into which principles on average are perceived to be best aligned. For instance, principles like Privacy (5.905) and Reliability and Safety (5.607) had higher average alignment scores than principles such as Intelligibility (4.512) and Inclusiveness (4.229).

The standard deviations in Table 4 also offer perspective into response variability. Fairness and Explainability exhibited wider standard deviations of 1.8 and 1.773 respectively, indicating more dispersed views on the alignment of these principles within organizations. In contrast,

principles like Privacy (1.228) and Security (1.28) had tighter standard deviations, suggesting greater agreement among participants regarding their organizational alignment.

In summary, Table 4 analyzes experts' perceptions of how well-aligned their organizations are with responsible AI principles in practice. This sheds light on relative strengths and opportunities in operationalizing ethics.

Responsible AI principles	Observed minimum	Observed maximum	Average	Standard Deviation
Benevolence and Non-maleficence	2	7	5.357	1.588
Reliability and Safety	2	7	5.607	1.336
Privacy	2	7	5.905	1.228
Security	2	7	5.690	1.28
Accountability	2	7	4.786	1.636
Explainability	1	7	4.548	1.773
Intelligibility	2	7	4.512	1.639
Transparency	2	7	4.583	1.6
Inclusiveness	1	7	4.229	1.776
Fairness	1	7	4.614	1.8

Table 4. Perceived Alignment of AI Initiatives with RAI Principles

- RAI Principles' Operationalization in Organizations

As shown in Table 5, experts were asked to assess the operationalization of RAI principles using a 7-point Likert scale. Consistent with previous questions, the scale ranged from 1 to 7, with 1 representing "Never" and 7 being "Always". The data provides insights into both average ratings and standard deviations. Certain principles such as Privacy, Security, Reliability, and Safety received above-average scores of 5.8, suggesting stronger implementation compared to others. This suggests a stronger alignment of these principles with actual practices in comparison to others. In contrast, other principles like Accountability and Inclusiveness averaged below 5, implying greater room for improvement.

Furthermore, the standard deviations within the table provide additional insights. Principles like Fairness and Inclusiveness exhibited standard deviations exceeding 1.8, signifying varying perspectives on how these concepts are put into practice. In contrast, principles like Privacy displayed tighter variability, with standard deviations near 1.3, indicating a higher level of consensus among participants regarding their operationalization. These findings highlight the varied challenges and levels of agreement in implementing responsible AI principles across organizations.

Responsible AI principles	Observed minimum	Observed maximum	Average	Standard Deviation

Benevolence and Non-maleficence	2	7	5.202	1.589
Reliability and Safety	2	7	5.798	1.315
Privacy	2	7	5.869	1.306
Security	2	7	5.869	1.17
Accountability	1	7	4.905	1.712
Explainability	1	7	4.607	1.672
Intelligibility	1	7	4.536	1.704
Transparency	2	7	4.643	1.573
Inclusiveness	1	7	4.446	1.796
Fairness	1	7	4.627	1.833

Table 5. Operationalization of RAI Principles in Organizations

5. Discussion

This section presents a discussion based on the findings of this study, offering valuable insights into practitioners' perceptions of responsible AI across importance, alignment, and operationalization dimensions. The diverse, experienced sample provides well-informed perspectives on progress and gaps in the operationalization of responsible AI principles.

- Patterns of Perceived Importance

Our survey responses revealed patterns in how RAI principles are perceived, allowing us to classify them into three distinct categories, as illustrated in Figure 1. Firstly, principles such as privacy, security, reliability, and safety received significantly higher average importance ratings, with scores often approaching or exceeding 6. This alignment with the common emphasis in AI initiatives on addressing paramount concerns related to responsible use in algorithmic decision-making reflects our findings (Ashok et al., 2022). Notably, the recognition of privacy, security, reliability, and safety as foundational elements, as outlined in ISO:24028 (2020), underscores the necessity of prioritizing these aspects in AI development.

Secondly, we observed that principles like benevolence, non-maleficence, and accountability constitute the second group, with average importance ratings around 5.5 to 5.2. While these principles are still considered significant, they fall slightly below the top-tier principles in terms of perceived importance. Conversely, the third group comprises principles such as intelligibility, transparency, explainability, fairness, and inclusiveness, which received lower average importance ratings, hovering around 4.5. These scores suggest that they may not be as highly prioritized within organizations compared to the principles in the first two groups.

- Alignment and Implementation

This pattern of prioritization based on perceived importance is also reflected in the alignment and implementation perceptions of these principles. Alignment between organizations' AI initiatives and RAI principles was assessed using the same scale. Privacy, reliability, and safety showed higher average alignment scores (approximately 5.9 and 5.6), while principles like fairness and inclusiveness had lower scores (around 4.5), implying room for improvement. The operationalization of RAI principles varied, with privacy, security, reliability, and safety demonstrating stronger implementation, while fairness and inclusiveness showed lower scores, indicating areas for enhancement.

- Highlighting Variations in Importance

The standard deviation scores, as depicted in Figure 2, offer valuable insights into the diversity of opinions among experts, highlighting the extent of variation in their views concerning the significance of specific principles. Notably, despite receiving high average importance ratings, certain principles displayed relatively wide standard deviations, signifying varying perspectives on their significance when developing and applying AI systems.

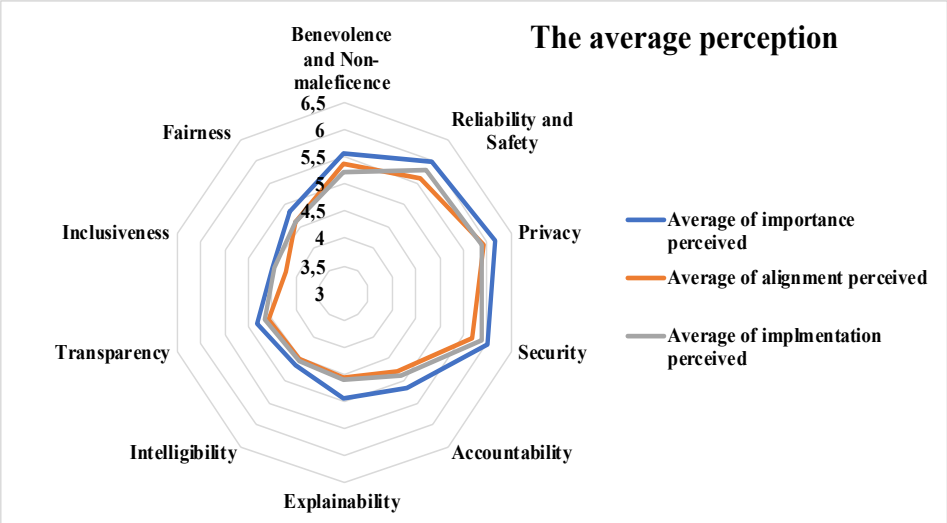

Figure 1. Average of perceived importance

For example, while benevolence and non-maleficence averaged a high importance score of 5.548, they also had a relatively wide standard deviation of 1.5. This variability can be partly explained by the subjective nature of these principles, where the definition of "benevolent" and "non-maleficent" system design may depend on contextual and cultural factors. Compared to more technical principles like reliability and safety, which received a narrower standard

deviation of 1.202, experts likely had more diverse interpretations of how benevolence should be defined and prioritized.

A similar pattern emerged for the principles of fairness and inclusiveness, which exhibited even more substantial variability with standard deviations over 1.8. This wide dispersion in views can be attributed to the contextual nature of these principles, where their operationalization and understanding often depend on the specifics of the situation, the stakeholders involved, and the broader societal context. Experts may have evaluated these principles differently given their diverse backgrounds. Effectively incorporating fairness and inclusiveness may require a more nuanced consideration of social perspectives and adaptive, context-specific approaches that can accommodate varied stakeholder needs (Díaz-Rodríguez et al., 2023).

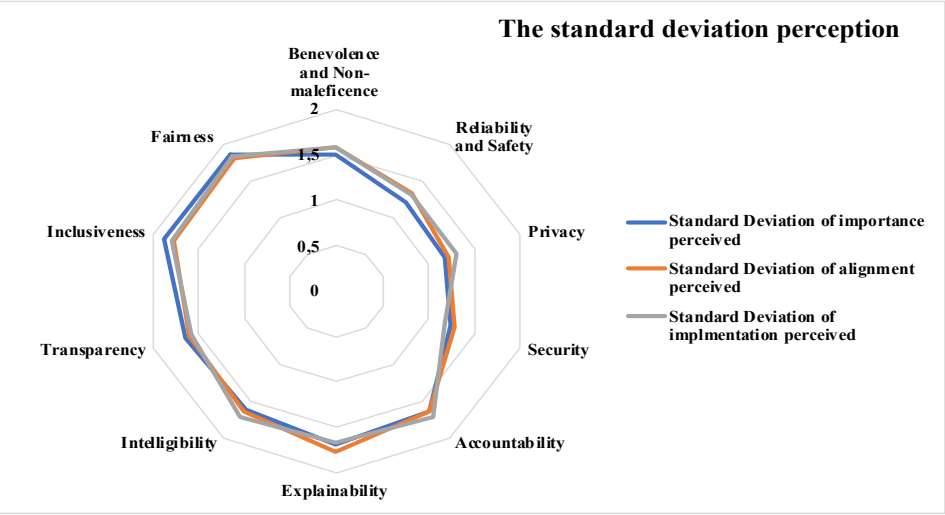

Figure 2. Standard deviation of perceived importance

- Insights and challenges in RAI principles

The results of the survey indicate a consistent prioritization of certain facets of responsible AI principles among organizations. Specifically, the principles of reliability and safety, privacy, and security were rated as high in importance. On the other hand, principles such as accountability, explainability, intelligibility, transparency, inclusiveness, and fairness were rated relatively lower.

The differences in the prioritization of RAI principles can be attributed to several factors. One possible explanation is that experts may have varying levels of understanding or expertise when it comes to evaluating these principles, leading to different assessments of their importance and operationalizing them (Sadek et al., 2024). Another factor could be the challenges that organizations face in effectively addressing certain principles. Some principles may require

more complex or resource-intensive measures to implement, making them more difficult to prioritize compared to more straightforward principles. The clarity and articulation of the principles themselves may also play a role. Certain principles, such as privacy and security, maybe more well-defined and have more established practices and algorithmic codes associated with them. In contrast, principles like fairness and inclusiveness may be less clearly defined, making it more challenging to translate them into concrete algorithmic implementations.

As a result, the higher-rated principles may represent more technical or tangible aspects of RAI, where the process of converting guidelines into codes and practices is more established and well-understood. For instance, when it comes to privacy and security practices, the associated algorithmic codes and practices are more accessible and mature compared to principles like fairness and inclusiveness, which may require further development and refinement to translate into effective algorithmic implementations. This distinction could contribute to the differing prioritization observed. Further research and analysis are needed to delve deeper into the factors influencing the divergent ratings and to develop a comprehensive understanding of the dynamics between these two groups of principles.

For example, previous studies have consistently supported the value of explainability and transparency in achieving fairness. Vimalkumar et al. (2021) argued that transparency makes AI mechanics visible and known, while explainability describes decisions impacting individuals in human terms, significantly contributing to fairness by enhancing the understanding of model logic and its effects (Robert et al., 2020).

However, the relatively lower prioritization of inclusiveness and fairness in our survey results diverges from views that emphasize the role of principles like transparency, expandability, intelligibility, and accountability. These principles collectively aim to make AI systems understandable, ensuring fairness and inclusiveness in AI development within organizations (Haresamudram et al., 2023). This discrepancy highlights the need for further exploration of the factors influencing the prioritization of these principles in practice. Practitioners need to better understand the importance of fairness, its benefits, and the potential risks for organizations when this principle isn’t prioritized.

6. Limitations and future research

While the study offers valuable insights, it is important to acknowledge and address inherent limitations. Firstly, a significant limitation is its reliance on subjective assessments, which may

introduce variability due to individual perceptions and biases. To mitigate this challenge, it is advisable to complement subjective assessments with objective metrics and external benchmarks whenever possible, promoting a more balanced evaluation.

Secondly, the study predominantly focuses on assessing perceptions around the importance, alignment, and operationalization of RAI principles, potentially neglecting other critical dimensions such as legal compliance or industry-specific considerations. To address this limitation, organizations can consider a broader set of aspects that are relevant to their specific context, thus providing a more comprehensive evaluation.

Lastly, the study currently only includes an analysis of quantitative data. Further work can involve incorporating qualitative methods, such as in-depth interviews or case studies, to gain a deeper understanding of nuanced contexts, especially when evaluating principles like fairness. This balanced approach would allow for a more comprehensive assessment of RAI principles and their practical implementation.

References

Akbarighatar, P., Pappas, I., & Vassilakopoulou, P. (2023a). Practices for Responsible AI: Findings from Interviews with Experts. *In Proceedings of the American conference on information systems (AMCIS 2023)*.

Akbarighatar, P., Pappas, I., & Vassilakopoulou, P. (2023b). A sociotechnical perspective for responsible AI maturity models: Findings from a mixed-method literature review. *International Journal of Information Management Data Insights*, *3*(2), p.100193.

Akbarighatar, P., Pappas, I., & Vassilakopoulou, P. (2023c). Justice as fairness: A hierarchical framework of responsible AI principles. *In Proceedings of the 31st European conference on information systems (ECIS 2023).*

Amugongo, L. M., Kriebitz, A., Boch, A., & Lütge, C. (2023). Operationalising AI ethics through the agile software development lifecycle: A case study of AI-enabled mobile health applications. *AI and Ethics*, pp.1-18.

Ashok, M., Madan, R., Joha, A., & Sivarajah, U. (2022). Ethical framework for Artificial Intelligence and Digital technologies. *International Journal of Information Management*, 62, 102433.

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

Clarke, R. (2019). *Principles for responsible AI.* https://tech.humanrights. gov.au/sites/default/files/inline-files/4A%20-%20Roger%20Clarke. pdf. Accessed 1 Nov 2020.

Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., de Prado, M.L., Herrera-Viedma, E. and Herrera, F., 2023. Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Information Fusion*, p.101896.

European Commission. (2019). Ethics guidelines for trustworthy AI. Retrieved from https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1.html.

Ess, C. (2009). Floridi's Philosophy of Information and Information Ethics: Current Perspectives, Future Directions. *The Information Society*, *25*(3), pp.159–168.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F. and Schafer, B., 2018. AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), pp. 689-5.

Flynn, B. B., Sakakibara, S., Schroeder, R. G., Bates, K. A., & Flynn, E. J. (1990). Empirical research methods in operations management. *Journal of Operations Management*, *9*(2), 250–284. https://doi.org/10.1016/0272-6963(90)90098-X.

Haresamudram, K., Larsson, S. and Heintz, F., 2023. Three levels of AI transparency. *Computer*, 56(2), pp.93-100.

Heyder, T., Passlack, N., & Posegga, O. (2023). Ethical management of human-AI interaction: Theory development review. *The Journal of Strategic Information Systems*, *32*(3), pp.101772.

Ho, C. W. L., Soon, D., Caals, K., & Kapur, J. (2019). Governance of automated image analysis and artificial intelligence analytics in healthcare. *Clinical Radiology*, *74*(5), 329–337.

ISO:22989. (2022). ISO/IEC 22989:2022 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology. Retrieved from .

ISO:24028. (2020). ISO/IEC TR 24028:2020 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence . Retrieved from https://www.iso.org/standard/77608.html)

MacKenzie, Podsakoff, & Podsakoff. (2011). Construct Measurement and Validation Procedures in MIS and Behavioral Research: Integrating New and Existing Techniques. *MIS Quarterly*, pp.*35*(2), 293.

Martilla, J. A., & James, J. C. (1977). Importance-Performance Analysis. *Journal of Marketing*, *41*(1), pp.77–79.

Matzler, K., Bailom, F., Hinterhuber, H. H., Renzl, B., & Pichler, J. (2004). The asymmetric relationship between attribute-level performance and overall customer satisfaction: A reconsideration of the importance–performance analysis. *Industrial Marketing Management*, *33*(4), pp.271–277.

Microsoft, "Empowering responsible AI practices | Microsoft AI." Accessed: Mar. 26, 2024. [Online]. Available: https://www.microsoft.com/en-us/ai/responsible-ai.

Mikalef, P., Conboy, K., Lundström, J. E., & Popovič, A. (2022). Thinking responsibly about responsible AI and 'the dark side' of AI. *European Journal of Information Systems*, *31*(3), pp.257–268.

Mirbabaie, M., Brünker, F., Möllmann Frick, N. R. J., & Stieglitz, S. (2022). The rise of artificial intelligence – understanding the AI identity threat at the workplace. *Electronic Markets*, *32*(1), pp.73–99.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), pp.501–507.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics*, *26*(4), pp.2141–2168.

Munn, L. (2022). The uselessness of AI ethics. *AI and Ethics*, 1-9.

Oh, H. (2001). Revisiting importance–performance analysis. *Tourism Management*, *22*(6), pp.617–627.

OECD. (2019). The OECD AI Principles. Retrieved from https://oecd.ai/en/ai-principles.

Pappas, I. O., Mikalef, P., Dwivedi, Y. K., Jaccheri, L., & Krogstie, J. (2023a). Responsible digital transformation for a sustainable society. *Information Systems Frontiers,* 25, 945–953.

Robert, L. P., Pierce, C., Marquis, L., Kim, S., & Alahmad, R. (2020). Designing fair AI for managing employees in organizations: A review, critique, and design agenda. *Human-Computer Interaction,* 35(5-6), 545-575.

Rungtusanatham, M. J., Choi, T. Y., Hollingworth, D. G., Wu, Z., & Forza, C. (2003). Survey research in operations management: Historical analyses. *Journal of Operations Management*, *21*(4), pp.475–488.

Sadek, M., Kallina, E., Bohné, T., Mougenot, C., Calvo, R. A., & Cave, S. (2024). Challenges of responsible AI in practice: Scoping review and recommended actions. *AI & Society*

Sojer, M., Alexy, O., Kleinknecht, S., & Henkel, J. (2014). Understanding the Drivers of Unethical Programming Behavior: The Inappropriate Reuse of Internet-Accessible Code. *Journal of Management Information Systems*, *31*(3), 287–325.

Stilgoe, J. (2018). Machine learning, social learning and the governance of self-driving cars. *Social Studies of Science*, *48*(1), pp.25–56.

Vakkuri, V., Kemell, K.-K., Kultanen, J., Siponen, M., & Abrahamsson, P. (2019). *Ethically Aligned Design of Autonomous Systems: Industry viewpoint and an empirical study* (arXiv:1906.07946). arXiv.

Van Esch, P., Black, J. S., & Ferolie, J. (2019). Marketing AI recruitment: The next phase in job application and selection. *Computers in Human Behavior*, *90*, 215–222.

Vassilakopoulou, P., Parmiggiani, E., Shollo, A., & Grisot, M. (2022). Responsible AI: Concepts, critical perspectives and an Information Systems research agenda. *Scandinavian Journal of Information Systems*, *34*(2).

Wall, L. D. (2018). Some financial regulatory implications of artificial intelligence. *Journal of Economics and Business*, *100*, 55–63.

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 195–200.

Zhang, H. Q., & Chow, I. (2004). Application of importance-performance model in tour guides' performance: Evidence from mainland Chinese outbound visitors in Hong Kong. *Tourism Management*, *25*(1), 81–91.