AI and Ethics (2025) 5:1787–1801
https://doi.org/10.1007/s43681-024-00524-4

ORIGINAL RESEARCH

Operationalizing responsible AI principles through responsible AI capabilities

Pouria Akbarighatar[1]

Received: 4 May 2024 / Accepted: 11 July 2024 / Published online: 24 July 2024

Abstract

Responsible artificial intelligence (RAI) has emerged in response to growing concerns about the impact of AI. While high-level principles have been provided, operationalizing these principles poses challenges. This study, grounded in recent RAI literature in organizational contexts and dynamic capability theory, and informed by literature on RAI principles and expert interviews in organizations deploying AI systems, (1) problematizes the high-level principles and low-level requirements and underscores the need for mid-level norms by adopting dynamic capability as a theoretical lens, and (2) develops five themes to capture firms' RAI capability, including (i) understandable AI model, (ii) bias remediation, (iii) responsiveness, (iv) harmless, and vi) common good. As our contribution to the field of information systems (IS), this study extends the emerging literature on operationalizing RAI and dynamic capabilities, empirically elucidating the capabilities needed by firms. For IS practice, we provide organizations deploying AI with novel insights to aid in the responsible implementation of AI.

Keywords Responsible artificial intelligence (RAI) · Principles · Dynamic capability · Responsible AI capabilities · Operationalizing RAI principles

1 The emergent of responsible AI

Over the last decade, artificial intelligence (AI) has advanced significantly, with various applications and methods emerging under this broad umbrella term. Key areas of progress include cognitive functions such as perception, learning, reasoning, problem-solving, planning, decision-making, natural language processing, and interaction with the environment [1]. AI can be defined as the field of study aimed at developing technologies that exhibit intelligent behavior, as Nils J. Nilsson explained in stating that "*Artificial Intelligence is actively devoted to making machines intelligent, and intelligence is that which enables an entity to function appropriately*, *and with foresight in its environment* [2]." This definition encompasses both technical aspects, such as functioning appropriately, and social dimensions, such as foresight in navigating its surroundings. This holistic view of AI aligns well with IS Research, which aims to offer socially relevant insights into Responsible AI. By exploring the balance between instrumental and humanistic AI outcomes, information systems (IS) research can contribute valuable knowledge to the sociotechnical landscape of AI development [3].

From a sociotechnical perspective, harmonizing interactions between multiple interconnected human and technological actors holds promise but also raises concerns. These include potential reinforcement of biases, privacy issues, public opinion manipulation, job displacement, and ethical implications like mass surveillance and autonomous weapons [4–6]. Addressing these responsibilities requires understanding how to balance the needs of diverse stakeholders and govern AI development responsibly [7].

Organizations implementing AI must navigate these complex dynamics by developing responsible practices for engaging with data [8], overseeing algorithmic performance, and collaborating on AI governance. Achieving this balance is crucial for realizing AI's potential while minimizing risks, especially given recent legislative developments like the approval of the AI Act by parliament. This legislation aims to ensure AI systems' safety, compliance with

✉ Pouria Akbarighatar
Pouriaa@uia.no

[1] Department of Information Systems, University of Agder, Universitetsveien 25, Kristiansand 4604, Norway

fundamental rights, and fostering innovation [9]. Recent regulatory actions, such as the European Parliament's approval of the AI Act, highlight the importance of responsible AI usage by setting penalties for organizations that do not fulfill their obligations. Compliance with regulations of this nature requires competencies in privacy, fairness, transparency, and other important aspects of RAI in firms. By cultivating these competencies and implementing responsible practices, firms can leverage AI's benefits safely and foster responsible AI.

2 Responsible AI principles

Over the years, many private and public organizations have published guidelines for developing AI responsibly. These guidelines, such as those proposed by AI4People and the European Commission, emphasize key principles like autonomy, beneficence, and fairness, while also stressing the significance of transparency and explainability in AI systems [10, 11]. These principles have been tailored to address the distinct challenges posed by AI across various sectors, including healthcare and human resource management, and have garnered recognition in both academic literature and practical applications. In addition to influential reports from organizations like the OECD and ISO, such as ISO 22,989 and ISO 24,038, which provide definitions and frameworks, there's a focus on principles like transparency, justice, and non-maleficence, aimed at avoiding harm, safeguarding human well-being, and ensuring comprehensibility of technologies and their impacts [12–14].

Barredo Arrieta et al. (2020) categorized these principles into two groups: AI-specific principles, which concentrate on aspects unique to AI, and end-to-end principles, covering all facets involved in AI development and implementation [15]. The distinctions between these principles are elaborated further in the results section. By adopting this classification, I summarized the most frequent principles in literature in Table 1, categorizing them. Nonetheless, the insights gathered from diverse sources contribute to a comprehensive understanding of these pivotal principles, underlining the importance of ethical considerations in AI initiatives and projects worldwide.

3 Problem statement

3.1 Operationalizing the principles

While increasingly more organizations are publishing AI principles to declare that they care about avoiding unintended negative consequences, there is much less experience on how to actually implement the principles into an organization. Principles are typically utilized to provide guidance and advice on achieving certain objectives, but they do not necessarily offer specific recommendations on how to operationalize and attain those objectives. This issue is also prevalent in the context of RAI within organizational settings. The literature highlights difficulties in translating abstract RAI principles into tangible actions within organizational contexts. Organizations face hurdles in implementing RAI principles effectively. Consequently, it can be

Table 1 RAI principles

Group	Principle	Literature descriptions	Refs
End-to-end principles	Benevolence and Non-maleficence	Indicate that AI technology is designed to promote good and maximize benefits, all the while avoiding harm and minimizing risks.	[10, 11, 16, 17]
	Reliability and Safety	AI systems should aim to prevent failures and accidents ensuring intended performance.	[12, 16, 17]
	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.	[12, 17]
	Security	Security refers to protecting data and controlling access based on authorization levels.	[12, 13, 16]
AI-specific principles	Accountability	Accountability refers to taking responsibility, providing justifications for actions, responding to inquiries, and being liable.	[12, 13, 16]
	Explainability	Explainability refers to providing comprehensive information about AI's inner workings.	[13, 14, 16, 17]
	Intelligibility	Intelligibility refers to enabling humans who use or manage AI to understand the reasoning of an AI system.	[12, 17]
	Transparency	Transparency entails disclosing AI system details, like performance, limitations, components, measures, design goals, data sources, for a decision, prediction, or recommendation.	[11, 13, 14, 16]
	Inclusiveness	Inclusiveness refers to involving diverse individuals and perspectives, regardless of their unique circumstances.	[13, 17]
	Fairness	AI systems must be designed to ensure impartial treatment, and prevention of discriminatory outcomes.	[11, 13, 16, 17]

argued that the current body of RAI principles now suffers from a number of conceptual problems as follows:

3.2 High-level and operational requirements challenges

One of the significant challenges encountered in the translation of RAI principles lies in the substantial disparity between high-level principles and the practical techniques essential for designing and developing responsible AI systems [18]. This gap, highlighted in the literature, poses a critical obstacle in effectively implementing RAI principles. The primary reason contributing to this gap is the lack of alignment between high-level principles and specific operational requirements. This discrepancy complicates the process of translating overarching principles into actionable steps within organizations, creating additional hurdles. Furthermore, the absence of well-defined methods for translating principles into practices exacerbates this issue. [19] emphasized that this translation process involves detailing high-level principles into mid-level norms and operational requirements, underscoring the complexity involved in bridging this gap.

Moreover, the current literature predominantly focuses on high-level principles aimed at end users and those affected by AI technologies, rather than on the deployer of AI systems [20]. As a result, organizations often find themselves lacking comprehensive guidance tailored to their specific needs and operational contexts. To address this gap, organizations must develop guidelines from an organizational perspective that facilitate the effective translation of principles into practical implementation. These guidelines should be formulated by taking into account the operational conditions, available resources, and organizational objectives. Additionally, it is crucial to acknowledge the potential for irresponsible behavior that may emerge as AI systems evolve over time. By recognizing this potential for irresponsible behavior, organizations can proactively safeguard against it. They need to ensure that these guidelines are not only clearly understood but also seamlessly integrated into their operational processes from an organizational standpoint. This approach promotes responsible AI practices and helps mitigate the risks associated with the long-term deployment and evolution of AI systems.

Furthermore, RAI principles have failed due to the lack of consequences in existing the organizations business ethical principles and the phenomenon known as ethical washing. This term, referred to as “toothless principles” in the literature, underscores the issue of principles lacking enforceability or real-world consequences [21, 22]. Addressing this challenge requires organizations to establish robust conditions and competencies for enforcement to ensure that RAI principles are adhered to in practice. Business as usual suggests a gulf between principles and practical implementation, as researchers have identified gaps between high-level ideals and concrete design requirements. This disconnect indicates that a new perspective is needed to bridge theory and practice. Zimmer and colleagues argue this point directly in another study. They note that for companies to translate principles into effective RAI systems, the value of responsibility must be clear and compelling from a business perspective. Without a viable value proposition integrated into a company’s overall goals and model, responsible practices are unlikely to be properly incentivized or prioritized against other competing demands [23].

Addressing these challenges necessitates a comprehensive approach that integrates high-level principles with practical operational requirements to ensure the effective implementation of RAI principles. Such an approach involves utilizing “theoretical glue”—a strong underlying logic and rationale that ties together disparate concepts, as Whetten (1989) describes [24]. By providing coherent connections, theoretical glue can help organizations navigate these barriers. One way to establish this theoretical glue is through the resource-based view and dynamic capabilities framework. As Whetten notes, conceptual theories should unite otherwise separate elements. The resource-based view conceptualizes how internal resources and competencies shape organizational dynamics and decision-making and gain competitive advantages.

3.3 Towards a responsible AI capability

In resource-based view (RBV) argues that organizations possess resources, which enable them to achieve competitive advantages and long-term performance. They make it clear and provide a clear interpretation of the meaning of the resources. Wade & Hulland (2004), defined it as “assets and capabilities that are available and useful in detecting and responding to market opportunities” [25]. Assets are defined as anything tangible or intangible the firm can use in its process for creating, producing, or offering goods or services. Assets can serve as inputs to a process or as the outcomes of a process. In contrast, capabilities transform inputs into output of greater worth. Capabilities can include managerial or technical abilities, or processes, such as knowledge management, system development, or integration. We adopt this definition for both capabilities and assets. In the context of AI, assets may include hardware, algorithms, data storage, and data. However, capability specifically refers to “*a firm’s strategic ability to select*, *orchestrate*, *and effectively leverage its AI-specific assets throughout various stages of the AI lifecycle*”. RBV provides a useful theoretical lens for IS researchers to think about how information systems relate to firm strategy and performance. Specifically, this theory provides a structured approach to assess the strategic significance of information systems resources and their individual impacts on performance. In the AI context, IS researchers have explored AI capabilities and argued that the strength of a firm's capabilities is contingent upon the resources on which they are built [26].

Moreover, various studies, academic publications, and business reports underline the diverse array of requirements and practices vital for organizations to tackle challenges related to the irresponsible use of AI [27]. For instance, Markus has introduced two concepts: responsibility in digital transformations (DT) and responsibilities in AI [23]. Responsibility in digital transformations pertains to the "how" of directing DT practices and operations to achieve responsible DT outcomes in a responsible manner. On the other hand, responsibilities in AI focus on the "what" of responsible DT outcomes, encompassing principles, guidelines, and responsibilities in DT. However, it is acknowledged by researchers that agreement on principles alone cannot guarantee their achievement. Consequently, the first concept can aid in integrating them into organizational activities and processes [28]. While this conceptualization helps delineate what RAI entails, there remains a lack of empirical and theoretical grounding regarding the implementation of these principles within organizational contexts.

Despite the widespread recognition among researchers of the pivotal role played by responsible AI in organizational success, there remains a dearth of understanding regarding the specific competencies and repeatable practices required for its effective integration into AI initiatives [29]. Acknowledging responsible AI as a critical resource and capability is essential for organizations aiming to navigate the responsible and practical complexities of AI adoption. However, the absence of a solid theoretical foundation to comprehend the requisite organizational competencies and capabilities underscores the pressing need for further exploration and development of frameworks and strategies aimed at fostering responsible AI practices within organizations.

Applying the Resource-Based View (RBV) theory to integrate responsible AI principles into their strategic framework enables organizations to enhance their capacity to leverage AI effectively while ensuring adherence to ethical principles and minimizing risks (responsibility in/of). RBV theory emphasizes the strategic importance of recognizing and utilizing internal resources and capabilities that contribute to a company's competitive advantage. In the realm of responsible AI, RBV theory assists organizations in identifying and developing specific competencies and repeatable activities—referred to as capabilities in this research—to effectively implement responsible AI principles. The RBV has been criticized for overlooking factors related to resources, often assuming their mere existence. This oversight becomes particularly crucial in the context of AI, especially when considering responsibility concerns. The ever-changing environmental and contextual factors necessitate organizations to continually adjust their resources to align with the evolving business landscape [30]. Dynamic capabilities emerge as a solution to address these challenges. While the RBV emphasizes resource choice or the selection of appropriate resources, dynamic capabilities focus on resource development and renewal. Therefore, organizations require dynamic capabilities to effectively navigate the complexities of the AI landscape.

Dynamic capabilities are pertinent to our study because they enable organizations to adapt and innovate in response to changes in the AI environment, ensuring they remain competitive and capable of addressing responsibility concerns. Through this lens, I define Responsible AI (RAI) capability as "*the ability of a firm to strategically select*, *orchestrate*, *and effectively implement responsible AI throughout the AI lifecycle*, *with a commitment to continuous adherence to RAI principles and readiness to act responsibly at any critical decision point*." This definition underscores the need to configure, integrate, and reconfigure resources and processes end-to-end as part of organizational dynamic capabilities, considering that responsibility may change as technologies impact individuals over time. This approach enables organizations to maintain responsibility as a key strategic differentiator amid disruptions in AI and associated markets. Consequently, organizations must identify the complementary competencies or capabilities they need to develop and adapt their strategies in response to timely requests and changes. Thus, it is imperative to explore how organizations can build RAI capability effectively.

3.4 Research methods

In developing the concept of RAI capability, various sources were drawn upon. These include the literature on RAI principles, which are summarized in Table 1, as well as professional literature such as the guidelines for implementing RAI from Microsoft, ISO/IEC 24028:2022, and the National Institute of Standards and Technology (NIST) risk management report [12, 17, 31]. Additionally, a series of expert interviews focusing on the deployment of AI systems were conducted. By combining insights from these resources and applying the Thematic Analysis (TA) and Reflexive TA approaches [32], first-order themes (sub-themes) were identified and subsequently aggregated into second-order themes (capabilities). Furthermore, I adopted the categorization provided by Barredo Arrieta et al. (2020), which comprises two dimensions: AI-specific principles and end-to-end principles, as the domain summaries to conceptualize the RAI capability.

3.5 Data collection

Experts with technical and process expertise in RAI were targeted, including C-level executives, RAI project managers, and researchers. Initially, I identified 89 experts involved in RAI projects in academia and industry through various sources such as LinkedIn profiles, literature research, practitioner reports, and Google Scholar profiles. Twenty experts participated in our study (Table 2), representing diverse educational backgrounds, including Information Systems (IS), Computer Science, and Ethics. Nine of the participants were from European organizations, and the other eleven were from North American organizations. Interviews were conducted in English, employing a semi-structured approach via online video conferencing between August 2023 and March 2024, with each session lasting between 45 min and 1 h.

The interviews explored how RAI principles were applied in organizational practices, processes, and roles. I developed an interview guide based on an extensive literature review and industry reports (Appendix A). Interviewees received a summary of responsible AI principles before the interviews. The questions were open-ended, encouraging participants to share their perspectives freely. Probing questions were used to delve deeper into specific topics. The interview guide included 25 questions organized into four themes: Demographics, AI projects and RAI implementation, RAI practices and actions, and reasons for operationalizing RAI principles in organizations.

3.6 Data analysis

The interview transcripts underwent analysis using both the 'codebook' school of reliability Thematic Analysis (TA) and Reflexive TA approaches. In the former, themes are considered 'domain summaries' predefined before analysis, while in the latter, themes emerge organically through an open and iterative coding process, focusing on identifying meaning-based patterns in the data [32]. Initially, I conducted a comprehensive coding of the entire dataset to identify prevailing sub-themes. This preliminary round of analysis enabled the identification of key themes and patterns within the data. In the subsequent phase, the outputs from the initial reliability text analysis were examined. To interpret this interaction, literature on ethics theories and the operationalization of RAI was consulted. However, the existing theories offered limited capacity to elucidate the complexities and challenges in operationalizing RAI that emerged prominently in our dataset. For example, attempts to apply deontological and consequential ethics theories revealed their inadequacies in addressing the dynamic nature of RAI principles and their practical implementation at the organizational level. This analysis highlighted a significant gap in the literature regarding the integration of high-level ethical principles with operational requirements. Consequently, it underscored the necessity for a more nuanced approach

Table 2 Interviews information

ID	Experts Role	Industry sector	Duration
E01	Researcher	Working on large-scale research project	40 min
E02	Researcher	Working on large-scale research project	40 min
E03	Senior Data Scientist	Providing services for public and private clients	30 min
E04	Senior Data Scientist	IT services and IT consulting	50 min
E05	C-level manager	Developing AI by considering responsible AI principles.	55 min
E06	Senior Data Scientist	Developing applications using machine learning (ML)	70 min
E07	CEO & Co-Founder	Providing AI ethics services to companies	60 min
E08	CEO & Co-Founder	Providing AI ethics services to companies	50 min
E09	C-level manager	Providing AI ethics services to companies	45 min
E10	C-level manager	Providing the AI ethics services to SMEs and startups	35 min
E11	CEO & Co-Founder	Developing AI by considering responsible AI principles.	30 min
E12	C-level manager	Developing AI by considering responsible AI principles.	55 min
E13	Senior Data Scientist	Developing applications using (ML)	45 min
E14	C-level manager	Developing applications using (ML)	4o min
E15	C-level manager	Developing AI by considering responsible AI principles.	35 min
E16	Lead data scientist	Developing applications using (ML)	65 min
E17	Lead data scientist	Developing applications using (ML)	33 min
E18	C-level manager	Providing AI ethics services to companies	40 min
E19	CEO	Developing applications using AI	38 min
E20	C-level manager	Providing services for public and private clients	39 min

to bridge these theoretical and practical aspects in the context of RAI.

Engaging with the literature on dynamic capabilities and RAI principles, particularly that on categorization from Barredo Arrieta et al., (2020) for RAI principles (AI-specific & end-to-end), served as the domain summaries. Additionally, the research questions guided the open coding approach. Through a process of reviewing and reflecting on the coded data, emergent themes (capabilities) were identified. Six phases of reflexive thematic analysis, including familiarization, generating codes, constituting, revising, and defining themes, as outlined by [32], were adopted in this stage. Table 3 illustrates the relationships between the sub-themes derived from thematic analysis, the themes (capabilities), and the two dimensions of capabilities outlined in the principles for RAI.

The data was further analyzed through the lens of the dynamic capability approach. This approach helped organize our understanding of how capabilities related to responsible AI development may emerge based on insights from experts' industry experiences implementing relevant principles. The reflective nature of this approach ensured that the analysis was guided by relevant concepts established prior to interpretation, while still allowing for codes to evolve iteratively during the coding process. The consistent movement between data coding and transcription review, guided by dynamic capability theory, facilitated a deeper interpretation of the competencies and capabilities required for RAI.

4 Results

4.1 AI-specific dimensions

According to the literature on responsible AI, principles specific to AI are those directly related to AI systems, such as fairness evaluations, explainability techniques, and accountability mechanisms. While these dimensions are essential, they are not sufficient on their own to develop dynamic RAI capabilities.

4.2 Capability as bias remediation

Based on recently published studies in professional reports like by the KPMG, highlighted that the outcomes of the AI systems are assessed regularly to ensure they are free of unfair bias, and designed to be inclusive to a diversity of users [33, p 40]. Inclusiveness and fairness are pivotal in bias remediation. Upholding these principles ensures that individuals from diverse backgrounds enjoy equal access to resources, benefits, and opportunities, irrespective of factors like race, gender, socioeconomic status, or other characteristics. Inclusiveness aims to engage individuals with diverse perspectives actively, acknowledging their unique circumstances. In the context of AI systems, this entails organizations involving stakeholders from diverse groups and considering their specific needs and insights, thereby ensuring equal opportunities in accessing opportunities. For an AI system to be inclusive, there must be practices and procedures that consider impacts on all user groups and continually re-evaluate effects over time [10, 15].

Table 3 Themes and dimensions

Sub-Themes	Themes (capabilities)	Aggregate Dimensions	Definition of Dimension
• Explainability methods like SHAP & LIME. • Trade-off between a model's complexity and its explainability • Document the procedures involved in data collection, and model building. • User-friendly communication.	Understandable AI Model	AI-specific	AI-specific principles concentrate on technical aspects that are unique to artificial intelligence as a technological domain. This includes principles related to aspects like: Accountability, Explainability, Intelligibility, Transparency, Inclusiveness, Fairness
• Clear definition for fairness. • Monitoring to identify proactively potential biases. • Training team engaged in AI initiatives about potential biases. • Examining historical data to recognize potential biases.	Bias Remediation		
• Expert intervention in developing the AI models. • Accountability mechanisms to take corrective actions.	Responsiveness		
• Knowledge of current laws and regulations regarding AI. • Clear accuracy definitions and reliable results. • Be aware of using sensitive data. • Data quality reviewing and auditing. • Updating AI models with recent data.	Harm Less	End-to-End	End-to-end principles are not only involved in the lifecycle of AI but are also influenced by other types of technologies or processes within organizations like: Benevolence and Non-maleficence, Reliability and Safety, Privacy, Security.
• Fulfill the end users' real requirements. • Balancing benefits and potential harms. • Analysing potential impacts to individuals and society. • Enhancing human well-being.	Common Good		

One of the experts from a company that provides responsible AI services to AI companies describes a practical use case. In this use case, a face identification company, which offers an API for liveness detection, asked them to test their models for performance on different demographic groups. They wanted to ensure that their models were inclusive and performed well across ethnic groups [E07]:

> "So, what we did with them was we collected a lot of data, and afterward, we also annotated the database on the different demographic characteristics of people. So, make sure that their data set this as inclusive as possible. So, for us, the inclusiveness of AI is very much stemming from the inclusiveness of the data".

On the other hand, fairness in AI systems must be designed to ensure impartial treatment and prevent discriminatory outcomes, especially for minorities by acknowledging differences. Scholars have discussed the concept of sociotechnical inherent fairness evaluation, recognizing the importance of considering contextual differences [34]. Algorithms can become biased based on the data they are trained on. Interestingly, the way in which the interviewees use practices for evaluating fairness also differs from expert to expert and company to company. We have found that a common definition of fairness, shared among all stakeholders is difficult to arrive at.

> "It might be the case that I define fairness and how you define it, how two different definitions and a data center were fair under your definition but unfairly under, mine definitions, and then understand how to compare your definition. In my definition, that doesn't take out our weaker or stronger definitions". [E04]

> "A group of practical actions is to look at what happens when we collect data, how we collect data, and what type of fairness construct we end up constructing. For instance, If I decide gender parity over any other definition, what will be the trade-offs in choosing differently, what are different pathways?" [E03].

Even more importantly, models that were initially assessed as fair, may become biased later as they continue learning from new datasets. So, it is important to continuously monitor model fairness.

> "How does the AI system learn after you put it on the market? Because you might create a pretty good structure, a pretty good accuracy level on all groups. But if the tool is then only used in a certain segment of the population and only learns from them then you are going to have the system move towards a bias later on". [E10]

Inclusiveness helps diversify data collection and training, mitigating bias against specific groups, as a result, inclusiveness ensures these diverse needs are considered, leading to solutions that are fair and accessible to all. One of the experts from a company that provides responsible AI services to AI companies describes a practical use case. In this use case, a face identification company, which offers an API for liveness detection, asked them to test their models for performance on different demographic groups. They wanted to ensure that their models were inclusive and performed well across ethnic groups.

I have found that assessing data in terms of inclusiveness is a focus for AI companies, and they have put in place processes to ensure it. Also, testing the behaviour of models across different subgroups is a practice that has been described.

> "Whatever model that we end up designing or we're helping our customers to design, we will make sure that the model has consistent performance on different subgroups within the data set" [E11].

Without inclusiveness, achieving a fair AI system and providing fair equal opportunities becomes challenging. However, it's important to note that having an acceptable level of inclusiveness does not guarantee fair decisions with the AI system. Therefore, to achieve equality, both inclusiveness and fairness must be simultaneously present and prioritized in the design and implementation of AI systems.

Some recommendations for achieving this capability from the literature include providing training to the team engaged in AI initiatives within the organization on recognizing potential biases in datasets and implementing strategies to evaluate them, examining historical data to recognize potential biases, ongoing monitoring and auditing of AI model outputs, and analyzing and identifying minority groups that could be impacted by AI initiatives to ensure unbiased outputs [17, 35].

4.3 Capability as responsiveness

Responsibility entails being responsive by providing justifications for actions, responding to inquiries, and accepting liability for any errors or harm caused by AI systems. When AI systems are deployed and fail, the consequences can seriously harm individuals. Examining AI through the lens of dynamic capability emphasizes the importance for organizations using AI systems to ensure they are developed and deployed by appropriately trained individuals and that decision-making processes are conducted by experts. Additionally, it is vital to establish mechanisms allowing individuals and groups adversely affected by AI-based decisions to seek redress and question outcomes. This underscores the urgent need for accountability mechanisms to properly assign responsibility, according to experts. As one expert highlighted, AI errors in high-risk domains like healthcare could negatively impact patients, providing an example:

> "For example, in healthcare, if an AI system misinterprets a patient's medical history and prescribes an incorrect treatment plan, it could harm the patient. We need mechanisms to identify responsibility and take corrective actions when such errors occur" [E09].

By having the ability to seek human intervention or explanation, individuals can gain insights into the decision-making process. The inclusion of this ability in AI initiatives ensures that individuals have agency and influence over the use of AI. Moreover, by providing avenues for questioning and redress, organizations can address potential biases, errors, or unintended consequences of AI systems and uphold the principles of fairness and accountability [12, p. 240, 17], as a result, this dimension can help detect potential biases, and raise concerns if they believe the AI system has produced unfair or undesirable outcomes.

4.4 Capability as understandable AI model

According to a recent study conducted by the National Institute of Standards and Technology (NIST), increasing levels of understanding can enhance confidence in AI systems [30, p 15]. In the literature, several terms and principles have been proposed to enhance the understandability of AI systems. These principles encompass transparency, explainability, and intelligibility. Explainability primarily focuses on the algorithm level and aims to provide comprehensive information about the inner workings of AI systems [36]. It involves offering insights into how the algorithms function, how they make decisions or predictions, and the factors considered in the process. By providing this level of explanation, users and stakeholders can gain a deeper understanding of the AI system's logic and reasoning. One expert highlighted the importance of explainability, stating:

> "Explainability is more about getting an idea of how our input of the model transforms into the output more from a functional perspective like what are the weights and biases, how the numbers flowing into the architecture, and so on." [E11].

Transparency, on the other hand, operates at the data and algorithms level. It involves disclosing details about the AI system, such as its performance, limitations, components, measures, design goals, and data sources used for decision-making, predictions, or recommendations. Intelligibility focuses more on the individual level, aiming to enable humans who use or manage AI systems to comprehend the reasoning behind the system's outputs. It involves presenting information in a way that is understandable and meaningful to human users, without requiring expertise in AI technologies. Our experts mentioned specific practices related to communicating and providing meaning full information to the users.

> "In terms of communication with users, I mean external facing; practices like giving information like what data is being collected, how it is used in the model, how users can ask their questions if they have, and the way of presentation." [E09].

> "Consistency and transparency are essential; if you are transparent but don't communicate well regularly, then people won't have that trust." [E04].

From a perspective of dynamic capability, principles, including transparency, explainability, and intelligibility, serve to support that organizations have the ability to provide understandable AI models. This access is vital in fostering an understandable AI environment, supporting better responsiveness and fair AI, as described above [37].

4.5 End-to-end dimensions

From the principles identified in the literature, several are relevant for any connected IT system, including AI systems, such as privacy, security, and benevolence. These dimensions are considerably more challenging to identify within organizations. However, despite being difficult to measure, similar to the AI-specific principles, they are interconnected with other systems and technologies. As a result, addressing these shared principles necessitates an end-to-end approach. This approach involves considering the entirety of organizational processes, policies, and cultural factors that influence the implementation and maintenance of responsible AI practices. By adopting an end-to-end perspective, organizations can effectively integrate these shared principles into their overarching strategies and workflows, ensuring comprehensive and sustainable adherence to responsible AI principles.

4.6 Capability as harm less

The ability to ensure AI systems operate safely and as intended is seen as critical for building a responsible AI system [15]. Specifically, having capabilities in place to ensure systems remain within established boundaries and prevent potential harms to individuals or society is regarded as an important factor. This perspective emphasizes that without adequate safeguards, systems could potentially behave in unexpected or undesirable ways as they continue to learn and be deployed. Privacy, security, safety and reliability are key principles that guide organizational efforts to achieve this important goal of maintaining control and oversight of AI. Reliability helps guarantee proper system performance as intended. Context-specific accuracy or reliability definition is mentioned by participants (E03, E06), pointed out that:

> "Without clear accuracy definitions, it's easy to overestimate a model's capabilities or gloss over limitations. This can obscure risks that may result in unintended harm if not addressed upfront."

Failing to define accuracy within the context of a specific situation risks fostering misconceptions about a model's abilities and may obscure its limitations which could lead to adverse outcomes. Framing task-specific accuracy expectations beforehand helps ensure systems are evaluated and optimized responsibly according to their intended role and constraints. Establishing function-aligned accuracy metrics from the start facilitates more reliable and safe applications of AI.

The continuous learning capability of AI systems depends on the quality and relevance of their input data, making data maintenance crucial throughout the system's lifecycle. This sentiment was insightfully expressed by an AI expert who likened data to the fuel-powering AI:

> "Data is the fuel for AI. If it gets stale or incomplete, models break down over time. You have to constantly review and refresh what they're learning from." [E15].

This observation underscores the vital role that data quality and relevance play in the sustained performance of AI systems. Interviewees, drawing from their experiences, emphasized the necessity of regularly reviewing and updating data models to ensure they remain comprehensive and up to date. Participants also stressed the need for processes to regularly audit data sources, identify, and address any coverage gap, purify noisy or anomalous information, and expand datasets with new information to retain real-world applicability over the long term and finally rebuild the model. In the words of one participant:

> "If you use data from 10 years ago to build a model, it won't reflect the current risks for certain diseases. To make it more meaningful, updating it with recent data is essential." [E09].

A focus on privacy helps ensure personal data is protected while addressing security concerns and reduces risks of systems being compromised or misused. During our research interviews, participants provided perspectives on safeguarding sensitive information. They highlighted the importance of finding a balance between respecting individuals' personal data (privacy) and implementing robust technical and physical safeguards (security). As one expert noted, knowledge of AI-related laws and regulations is essential for ensuring compliance and protecting both privacy and security.

> It's really important for key decision-makers within organizationss to know about the laws and rules related to AI. If they don't keep up with these security rules, the company could face a lot of legal trouble. So, it's crucial for companies to stay updated on these rules to avoid legal problems and maintain operational integrity." [E10].

This competency enables organizations to make informed decisions and navigate compliance requirements effectively. In combination, these principles necessitate competencies and capabilities within organizations to continually monitor and identify issues, as well as to stay abreast of current laws and regulations concerning AI.

4.7 Capability as common good

According to AIUK[1], AI should be developed for the common good and benefit of humanity. Two principles, benevolence, and non-maleficence, are aligned with this vision [11]. The principle of benevolence highlights AI's potential to positively impact people's lives, aiming to enhance well-being and drive societal progress. Similarly, the principle of non-maleficence emphasizes the need to prevent harm and mitigate negative outcomes resulting from AI systems. Emphasizing these principles ensures that AI is developed and deployed with a focus on human well-being, with measures in place to prevent misuse and harmful consequences. Many participants stressed that organizations should first

[1] The "five overarching principles for an AI code" presented in paragraph 417 of the UK House of Lords Artificial Intelligence Committee's report, AI in the UK.

focus on developing systems that are intentionally engineered to fulfill important social needs and provide meaningful value to communities. Only after building and aligning core objectives with user insights and societal requirements should consideration expand to potential risks at broader societal levels. As one C-level manager succinctly stated.

> "If the AI system is designed to fulfill the real needs of the users and can provide value to them, then we will think about the potential risks that might happen during the deployment; of course, this is at a high level and not an individual level." [E16].

As a result, organizations must cultivate the capability to assess whether AI initiatives enhance well-being and add value while avoiding harm. For instance, consider a children's publishing company exploring the use of generative AI to create stories for kids. While AI-generated content could improve children's reading abilities, there's a risk it could inadvertently manipulate or harm young minds. Thus, organizations must carefully weigh the benefits against potential risks to children's and families' well-being. Participants also underscored that consideration of socioeconomic impacts is another crucial reason the practice of balancing benefits and potential harms to individuals and society emerged as a critical aspect of responsible AI development within organizations. This perspective aligns with a comment from a C-level manager who asserted,

> "AI systems are like any other product you might find in a store; they come with costs, so it's important to weigh the benefits against these costs carefully. Plus, we need to make sure they work properly and don't cause any unexpected problems when we use them." [E15].

By acknowledging AI technologies as products that must be carefully designed, tested, and deployed, this leader emphasized the importance of considering the potential socioeconomic impacts. Participants also underscored that this perspective aligns with the practice of balancing benefits and potential harms to individuals and society, which emerged as a critical aspect of responsible AI development within organizations.

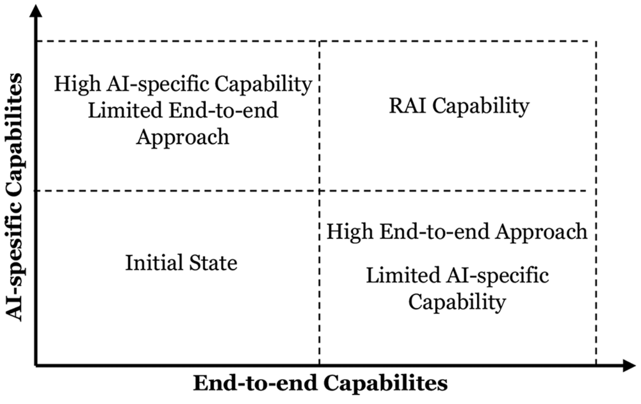

Fig. 1 RAI capability framework

To foster capabilities for maintaining a healthy, respectful, beneficial, and secure AI from the outset of AI initiatives, as highlighted by researchers Robertson et al., (2024), decision-making processes can simplify decision-making processes and reduce complexities in subsequent steps [38]. One recommended approach to foster this capability is through the formulation of business cases for the use of artificial intelligence that bring value and profitability to organizations while meeting the requirements of end users. This ensures that AI initiatives are not only ethically sound but also strategically aligned with the overarching goal of promoting human well-being and societal progress.

4.8 Responsible AI capability framework

Neither the AI-specific nor the end-to-end dimension, as proposed, fully elucidates the concept of RAI capability. Each dimension encompasses unique aspects relevant to an organization's RAI capability. This suggests that disregarding either dimension hinders the attainment of RAI capability. In other words, both dimensions are essential components that contribute to achieving RAI capability, and neglecting one would undermine the overall effectiveness of responsible AI practices within an organization. Bearing this in mind, a matrix view is proposed to depict these RAI dimensions to provide a visual representation of these dimensions and illustrate the necessity of strategically balancing both (Fig. 1). The matrix will consist of four quadrants based on the AI-specific and end-to-end dimensions, offering a comprehensive depiction of RAI capability and facilitating a better understanding of organizational approaches to responsible AI adoption.

The axes here denote the theoretical foundations of the RAI capability, and it draws from the Barredo Arrieta et al. classification of RAI principles. The vertical axis of the matrix represents the AI-specific dimension, with one end representing organizations highly focused on AI-specific principles and the other end representing organizations with minimal emphasis on this dimension. Similarly, the horizontal axis represents the end-to-end dimension, with one end representing organizations that prioritize end-to-end considerations and the other end representing organizations that overlook these broader organizational factors. The matrix divides the space into four quadrants. Each quadrant represents a different strategic profile considering the interaction between emphasis on AI-specific and end-to-end dimensions. This characterization provides insight into organizations' blended competency approaches.

By mapping organizations onto this matrix, four distinct quadrants can be identified, each offering insights into various approaches to RAI capability. In the first quadrant, organizations neglect both dimensions. These organizations may lack a structured approach to responsible AI adoption, potentially leading to ethical, legal, or operational risks associated with AI deployment. They might struggle to harness the full potential of AI while managing associated risks and maintaining ethical standards. Moving to the second quadrant, organizations prioritizing end-to-end considerations over AI-specific principles focus on integrating AI within the organizational context but may lack in-depth expertise in AI-specific principles. While they may excel in aligning AI initiatives with end-to-end organizational principles, they might face challenges in implementing AI techniques effectively in RAI adoption like explainability or transparency.

Conversely, in the third quadrant focusing on AI-specific principles over end-to-end considerations may demonstrate excellence in areas such as explainability, transparency, accountability, and fairness evaluations. However, they may inadvertently neglect broader organizational factors such as privacy and security. While this approach may lead to advancements in models and techniques for RAI adoption, it could pose challenges in aligning with the overarching organizational principles related to responsible AI across all stages. Finally, the fourth quadrant represents organizations that emphasize both dimensions, I find a balanced approach that integrates AI-specific principles with end-to-end considerations. These organizations tend to have a comprehensive understanding of responsible AI practices and are likely to excel in implementing them across various aspects of their operations.

Another important aspect to consider is the transition of organizations from an end-to-end dimension to a more AI-specific dimension, bridging the gap between text-based governance frameworks and practical computational tools for practitioners becomes increasingly crucial. Presently, governance programs primarily offer guideline documents, which predominantly consist of explanatory texts. However, AI experts require practical tools to effectively implement these principles in their daily tasks.

> "For instance, The EU AI Act ensures AI technologies comply with legal frameworks. Governance programs focus on guidelines. However, they are mainly text-based. My role in the team is to bridge this gap between text-based governance and computational tools for engineers. This bridge is crucial". [E16].

Overall, this framework provides a nuanced understanding of organizational approaches to responsible AI adoption. By identifying their position within the matrix, organizations can assess their strengths and weaknesses, develop targeted strategies for improvement, and ultimately enhance their RAI capability as a dynamic and adaptive organizational capability over time. Additionally, capabilities extend beyond AI-related principles; other types of technologies or processes within organizations should collaborate to enhance responsible AI capability.

4.9 Discussion and conclusion

Although the interest in responsible AI is continuously growing, reports and empirical studies from early adopters indicate that many organizations are struggling to operationalize RAI principles [18, 39]. These findings are particularly notable in light of numerous articles that underscore the complexity of implementing RAI. Mittelstadt, (2019) describes this translation process, asserting that it entails elucidating high-level principles into mid-level norms and low-level requirements [19]. This highlights the intricate nature of transforming abstract principles into practical guidelines for AI development and deployment guidelines. Furthermore, technology and business consultants have documented many reports and studies regarding the operationalization of RAI principles, lacking the theoretical basis needed to consolidate findings.

4.10 Research implications

This study aims to establish a theoretical framework for defining responsible AI capabilities within organizations. Drawing upon dynamic capability theory as theoretical "glue," the primary objective is to investigate how specific competencies impact an organization's ability to successfully implement responsible AI principles over time, thereby maximizing the benefits derived from AI technologies and achieving enhanced performance gains. While numerous practitioner-focused publications have underscored the potential value of RAI, many lack a theoretical framework to elucidate the organizational prerequisites necessary for successful RAI adoption and alignment with organizational objectives. Moreover, existing academic literature predominantly concentrates on technical and business aspects, often neglecting the challenges associated with irresponsible AI usage and how these issues should be integrated into strategic planning and decision-making processes within organizations objectives [26]. Consequently, various commentaries and research studies emphasize the importance of identifying the essential competencies regarding RAI that organizations must cultivate to prepare for AI initiatives, thereby supporting their core activities [27].

This study contributes significantly to the RAI literature by introducing a theoretical framework of RAI capability, comprising five capabilities. It emphasizes the importance of examining specific capabilities and competencies tailored to emerging technologies and dynamic environments, which impose unique demands on organizations. Additionally, the study underscores the necessity of balancing tradeoffs within RAI operationalization strategies. Drawing from a recent report by NIST highlights the adverse outcomes associated with systems exhibiting high security but unfairness, reliability without understandability, and accuracy compromised by insufficient security and privacy measures. Therefore, balancing, and reflective approaches that address and reconcile these trade-offs become imperative for effective RAI implementation [30, p 13].

Secondly, I employ a dynamic capability approach and draw upon RAI principles literature and interviews with experts to identify and categorize capabilities relevant to RAI and AI initiatives within organizational settings. This study provides a definition of responsible AI capability, conceptualizing them as dynamic competencies that facilitate the translation of high-level ideals into actionable strategies at ground level. These capabilities serve as organizational prerequisites necessary to operationalize principles through requisite processes, resources, skills, and strategic adaptation. Furthermore, the identification of capabilities relies on well-established literature that utilizes a plethora of approaches to ensure an exhaustive and complete set of RAI capabilities that jointly comprise a capability. This is done by surveying business reports, practitioner-based press, research publications, and new releases concerning the operationalization of RAI at the organizational level. By performing this, a list of important principles was summarized, which was then grouped and categorized as described in the previous section.

Thirdly, by building upon the aforementioned theoretical framework, this study establishes a foundation that can be empirically applied to assess the RAI capabilities of organizations. The argument is that, theoretically, RAI capability is distinct from other digital capabilities, such as IT capabilities or soft capabilities like knowledge management capabilities. It requires a different approach to extract the necessary competencies when adopting AI initiatives, and this can affect organizations' customers, serving as an important stakeholder. The irresponsible experiences of organizations in using their AI can directly impact performance and business values over the long run [8, 40].

In summary, building upon the established theoretical framework provided by the dynamic capability lens, this study extends the existing body of research in the IS community by applying it within the context of AI and responsible AI. It seeks to elucidate the dynamic capabilities that organizations must develop to strike a balance between instrumental and humanistic objectives, both of which can generate business value for them. I follow the reasoning and argumentation of Wade & Hulland (2004), who suggest that this approach (RBV and dynamic capability) can provide benefits to the IS community as (1) the RBV provides the foundation for specifying firm-level resources, (2) it allows for distinction between the assets and capabilities which make it clear how capabilities can complement the assets, and (3) it enables researchers to systematically test the relationship between aggregate of capabilities from different perspectives, with key performance outcomes. Our adoption of the dynamic capability framework sheds light on how the constantly shifting environmental and contextual factors compel organizations to continuously adapt their capabilities required to be responsible. This adaptation is crucial for ensuring alignment with the dynamic nature of the business landscape, where strategies and resources need to evolve to remain effective and competitive over time.

4.11 Practical implications

By integrating responsible and ethical considerations into our conceptualization of RAI capabilities, the emphasis is placed on the importance of broadening perspectives to include additional ethical factors and competencies when designing AI deployment strategies. This approach underscores the need to consider a wider range of ethical dimensions, ensuring a more comprehensive and responsible AI adoption within organizations. While business and instrumental capabilities have predominantly been featured in practice-based literature for bringing AI initiatives to fruition, our study highlights the significance of more nuanced, yet equally important aspects related to AI success. In fact, prioritizing instrumental intentions and investments alone may not lead to significant performance improvements. Instead, managers and decision-makers should foster structures and practices that facilitate value generation from AI investments, including the cultivation of responsible capabilities.

Through this study, I have attempted to provide organizations with a better understanding of the development of RAI capabilities in two dimensions: AI-specific and End-to-end. Organizations first need to mobilise and utilize their existence competencies and develop and integrate them with new or required competencies in a dynamic manner. The development of AI-specific and end-to-end capabilities allows organizations to adopt a continuous view of their operations in order to manage the context, make informed decisions, and adapt to changing environments effectively.

The proposed model offers guidance for sustainably building competencies tailored to an organization's specific AI initiatives. Thus, the research provides a foundation for aligning RAI operations with a long-term perspective, with the goal of developing AI-specific and end-to-end capabilities that facilitate the continuous evolution and utilization of AI within organizations. In essence, the study aimed to enhance comprehension of RAI capabilities and aid organizations in implementing responsible, dynamically managed AI capabilities conducive to long-term sustainability within an ever-evolving landscape.

4.12 Limitations and avenues for future research

As with any research, this study has its limitations. First, while it outlines the primary types of capabilities firms should consider when designing and deploying AI initiatives, it cannot be considered a universal model entirely applicable to all organizations. This research is in the initial stages of understanding operationalized RAI in the business context, and providing an exhaustive list of capabilities driving RAI capabilities is challenging. This complexity increases due to several reasons for RAI. First, it is probable that some organizations may require additional capabilities to leverage their AI investments based on contingent aspects such as cultural factors, size-class, industry, or type of AI application. Furthermore, the AI capability construct is by no means exhaustive, so there may be additional important aspects that were not captured, which future research could examine.

This study introduces an initial theoretical framework that identifies two main categories of organizational RAI capability: AI-specific and end-to-end. Within these categories, five core capabilities are proposed. While the framework outlines the hierarchical structure of RAI capability and its constituent parts, the current research does not attempt to define or specify measurement items for each construct. Further research is necessary to empirically validate the model and determine how each capability can be rigorously assessed. Future studies could expand on this conceptual groundwork by creating measurement scales tailored to evaluate proficiency in the identified areas. Although the current study represents an initial step toward a unified understanding of RAI capability, it also presents opportunities for subsequent research to refine and apply the theoretical framework through the development and testing of quantitative instruments.

While this study identified and described the main dimensions and capabilities of RAI, it does not delve into a process-based perspective of how RAI unfolds or how to prioritize capabilities to deploy RAI more effectively. It is highly probable that organizations follow different trajectories when it comes to prioritizing dimensions for low-impact tasks versus high-impact tasks, and they may encounter different challenges and obstacles along the way. By adopting an interpretivist approach (for instance case studies), it should be possible to uncover the forces that influence choices around AI deployments and the requirements that need to be considered to prioritize the dimensions effectively. This means understanding the underlying factors that drive decision-making processes within organizations, such as application domains, the scale of the effect, and external market dynamics. Additionally, it involves recognizing that the prioritization of RAI capabilities is not a one-size-fits-all approach, but rather a nuanced and context-dependent process that requires careful consideration of various factors and stakeholders' perspectives.

Appendix A: the interview guide

- **Personal background**.

Could you tell us about your academic and professional background?

How long have you been part of the AI projects, and how long have you held your current position?

- **ML and AI projects at your organization**.

Could you please describe the main ways your organization utilizes artificial intelligence technologies?

Could you please describe the AI projects that you are involved with?

What kind of AI models are used in these projects?

- **Experience of responsible AI principles**.

How do these projects in which you have been involved do you think are following the responsible AI principles?

a) **If yes**, describe how you are using them (for what purpose); please explain your experience on how responsible AI principles are defined in your research/company.
b) **If not but planning to employ it**, why are you planning to employ such systems?
c) **If not and no plan to employ it**, why are you not using these principles in AI systems?

Based on your experiences, do you think some RAI principles are more important than others in your case? How do you balance principles like privacy and understandability?

- **Reasons for operationalizing RAI principles in organizations**:

Can you explain the primary motivations behind your organization's decision to operationalize RAI principles?

What specific benefits or goals does your organization aim to achieve by implementing RAI principles?

Have there been any external factors, such as regulatory requirements or societal expectations, influencing your organization's commitment to RAI principles.

- **Implementations and practices in organizations**.

How understandable are the decisions of the AI used in the projects involved in?

Have you encountered a case in which you needed to explain a particular AI decision?

Have these explanations been documented?

What are the practices or producers in cases explanations are requested by the users, could you give concrete examples of these practices?

How would you explain the resulting decision if requested to do so….

- By expert auditors?
- By an affected organization?
- By the general public?

To what extent do you think the decisions made by the algorithms in your projects are fair?

Have you encountered a case or cases in which the AI decision was unfair?

what were the procedures and practices to identify and solve these issues?

To what extent do you think the decisions made by the algorithms in your projects are safe?

Have you encountered a case or cases in which an AI decision put users at risk?

What kinds of risks have you encountered? Could you give concrete examples of them?

What were the procedures and practices to identify and solve these issues?

Do you need to limit your use of AI due to potential unintended consequences and harms? Can you recall any specific examples?

Acknowledgements I would like to express my gratitude to all the participants (experts), most of whom are highly experienced in the field of Responsible AI, for their invaluable contributions to this research.

Funding Open access publishing was facilitated by the University of Agder.
Open access funding provided by University of Agder

Data availability Data available on request due to privacy/ethical restrictions: The data supporting this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical limits.

Code availability Not applicable.

Declarations

Ethical approval Not applicable.

Consent of participants Not applicable.

Consent for publication Not applicable.

Conflict of interest Not applicable.

References

1. Stohr, A., Ollig, P., Keller, R., Rieger, A.: Generative mechanisms of AI implementation: A critical realist perspective on predictive maintenance. Inf. Organ. **34**(2), 100503 (Jun. 2024). https://doi.org/10.1016/j.infoandorg.2024.100503
2. Nilsson, N.J.: The Quest for Artificial Intelligence. Cambridge University Press, Cambridge (2009). https://doi.org/10.1017/CBO9780511819346
3. Sarker, S., et al.: Jan., The Sociotechnical Axis of Cohesion for the IS Discipline: Its Historical Legacy and its Continued Relevance, *MIS Q*, vol. 43, no. 3, pp. 695–719, (2019). https://doi.org/10.25300/MISQ/2019/13747
4. Fumagalli, E., Rezaei, S., Salomons, A.: OK computer: Worker perceptions of algorithmic recruitment, *Res. Policy*, vol. 51, no. 2, p. 104420, Mar. (2022). https://doi.org/10.1016/j.respol.2021.104420
5. Mateen, H.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy: Cathy O'Neil. Broadway Books, 268 Pages, *Berkeley J. Employ. Labor Law*, vol. 39, no. 1, pp. 285–292, 2018. (2016)
6. Rinta-Kahila, T., Someh, I., Gillespie, N., Indulska, M., Gregor, S.: Managing unintended consequences of algorithmic decision-making: The case of Robodebt. J. Inf. Technol. Teach. Cases. 204388692311655 (Mar. 2023). https://doi.org/10.1177/20438869231165538
7. Meske, C., Bunde, E., Schneider, J., Gersch, M.: Explainable Artificial Intelligence: Objectives, stakeholders, and Future Research opportunities. Inf. Syst. Manag. **39**(1), 53–63 (Jan. 2022). https://doi.org/10.1080/10580530.2020.1849465
8. Vassilakopoulou, P., Parmiggiani, E., Shollo, A., Grisot, M., Responsible, A.I.: Concepts, critical perspectives and an Information Systems research agenda, *Scand. J. Inf. Syst*, vol. 34, no. 2, Dec. [Online]. Available: (2022). https://aisel.aisnet.org/sjis/vol34/iss2/3
9. Parliament, E.U.: Artificial Intelligence Act: MEPs adopt landmark law| News| European Parliament. Accessed: Mar. 15, 2024. [Online]. Available: https://www.europarl.

europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law

10. European Parliament, European Parliament, Accessed: May 02, 2024. [Online]. Available: https://www.europarl.europa.eu/cms-data/196377/AI%20HLEG_Ethics%20Guidelines%20for%20Trustworthy%20AI.pdf
11. Floridi, L., et al.: AI4People—An ethical Framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds Mach. **28**(4), 689–707 (Dec. 2018). https://doi.org/10.1007/s11023-018-9482-5
12. ISO 24028: Information technology Artificial intelligence Overview of trustworthiness in artificial intelligence, Bing. Accessed: Mar. 26, [Online]. Available: (2024). https://www.bing.com/search?q=ISO+24028&cvid=87301a63b9ce4cc7808ab7b8282b8531&gs_lcrp=EgZjaHJvbWUyBggAEEUYOTIGCAEQABhAMgcIAhBFGPxV0gEINjMxMGowajmoAgCwAgA&FORM=ANAB01&PC=U531
13. OECD: AI-Principles Overview. Accessed: May 02, 2024. [Online]. Available: https://oecd.ai/en/principles
14. ISO:22989: ISO/IEC 22989:2022. Information technology — Artificial intelligence — Artificial intelligence concepts and terminology., ISO. Accessed: May 02, 2024. [Online]. Available: https://www.iso.org/standard/74296.html
15. Barredo Arrieta, A., et al.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion. **58**, 82–115 (Jun. 2020). https://doi.org/10.1016/j.inffus.2019.12.012
16. Clarke, R.: Principles for responsible AI., [Online]. Available: (2019). https://tech.humanrights.gov.au/sites/default/files/inline-files/4A%20-%20Roger%20Clarke. pdf. Accessed 1 Nov 2020
17. Microsoft: Empowering responsible AI practices| Microsoft AI. Accessed: Mar. 26, 2024. [Online]. Available: https://www.microsoft.com/en-us/ai/responsible-ai
18. Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices, *Sci. Eng. Ethics*, vol. 26, no. 4, pp. 2141–2168, Aug. (2020). https://doi.org/10.1007/s11948-019-00165-5
19. Mittelstadt, B.: Principles alone cannot guarantee ethical AI. Nat. Mach. Intell. **1**(11), 501–507 (Nov. 2019). https://doi.org/10.1038/s42256-019-0114-4
20. Sanderson, C., et al.: Jun., AI Ethics Principles in Practice: Perspectives of Designers and Developers, *IEEE Trans. Technol. Soc*, vol. 4, no. 2, pp. 171–187, (2023). https://doi.org/10.1109/TTS.2023.3257303
21. Green, B.: The Contestation of Tech Ethics: A Sociotechnical Approach to Technology Ethics in Practice, *J. Soc. Comput*, vol. 2, no. 3, pp. 209–225, Sep. (2021). https://doi.org/10.23919/JSC.2021.0018
22. Munn, L.: The uselessness of AI ethics. AI Ethics. (Aug. 2022). https://doi.org/10.1007/s43681-022-00209-w
23. Zimmer, M., Minkkinen, M., Mäntymäki, M.: Responsible Artificial Intelligence Systems Critical considerations for business model design, *Scand. J. Inf. Syst*, vol. 34, no. 2, Dec. 2022, [Online]. Available: https://aisel.aisnet.org/sjis/vol34/iss2/4
24. Whetten, D.A.: What constitutes a theoretical contribution? Acad. Manage. Rev. **14**(4), 490–495 (1989). https://doi.org/10.2307/258554
25. Wade, Hulland: Review: The resource-based View and Information Systems Research: Review, extension, and suggestions for Future Research. MIS Q. **28**(1), 107 (2004). https://doi.org/10.2307/25148626
26. Mikalef, P., Gupta, M.: Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. Inf. Manage. **58**(3), 103434 (Apr. 2021). https://doi.org/10.1016/j.im.2021.103434
27. Akbarighatar, P., Pappas, I., Vassilakopoulou, P.: A sociotechnical perspective for responsible AI maturity models: Findings from a mixed-method literature review. Int. J. Inf. Manag Data Insights. **3**(2), 100193 (Nov. 2023). https://doi.org/10.1016/j.jjimei.2023.100193
28. Zimmer, M.P., Järveläinen, J., Stahl, B.C., Mueller, B.: Responsibility of/in digital transformation. J. Responsible Technol. **16**, 100068 (Dec. 2023). https://doi.org/10.1016/j.jrt.2023.100068
29. Akbari Ghatar, P., Pappas, I., Vassilakopoulou, P.: Practices for Responsible AI: Findings from Interviews with Experts, *Proc. Am. Conf. Inf. Syst. AMCIS* Aug. 2023, [Online]. Available: (2023). https://aisel.aisnet.org/amcis2023/sig_odis/sig_odis/4
30. Daniel, E.M., Wilson, H.N.: The role of dynamic capabilities in e-business transformation, *Eur. J. Inf. Syst*, vol. 12, no. 4, pp. 282–296, Dec. (2003). https://doi.org/10.1057/palgrave.ejis.3000478
31. NIST: Artificial Intelligence Risk Management Framework (AI RMF 1.0), National Institute of Standards and Technology (U.S.), Gaithersburg, MD, NIST AI 100-1, (2023). https://doi.org/10.6028/NIST.AI.100-1
32. Braun, V., Clarke, V., Hayfield, N., Terry, G.: Thematic analysis. In: Liamputtong, P. (ed.) in Handbook of Research Methods in Health Social Sciences, pp. 843–860. Springer Singapore (2019)
33. Gillespie, N., Lockey, S., Curtis, C., Pool, J., Akbari, A.: Trust in Artificial Intelligence: A global study, The University of Queensland; KPMG Australia, Brisbane, Australia, Feb. (2023). https://doi.org/10.14264/00d3c94
34. Kazim, E., Koshiyama, A.S.: A high-level overview of AI ethics, *Patterns*, vol. 2, no. 9, p. 100314, Sep. (2021). https://doi.org/10.1016/j.patter.2021.100314
35. Someh, I., Wixom, B.H., Beath, C.M., Zutavern, A.: Building an Artificial Intelligence Explanation Capability, *MIS Q. Exec*, vol. 21, no. 2, Jun. 2022, [Online]. Available: https://aisel.aisnet.org/misqe/vol21/iss2/5
36. Haresamudram, K., Larsson, S., Heintz, F.: Three levels of AI transparency. Computer. **56**(2), 93–100 (2023). https://doi.org/10.1109/MC.2022.3213181
37. Watson, H., Nations, C.: Addressing the growing need for algorithmic transparency. Commun. Assoc. Inf. Syst. **45**(1) (Dec. 2019). https://doi.org/10.17705/1CAIS.04526
38. Robertson, J., Ferreira, C., Watson, R., McCarthy, I., Kietzmann, J., Pitt, L.: Assessing digital responsibility in a digital-first world: Revisiting the U-commerce framework. Organ. Dyn. 101044 (Mar. 2024). https://doi.org/10.1016/j.orgdyn.2024.101044
39. Heyder, T., Passlack, N., Posegga, O.: Ethical management of human-AI interaction: Theory development review. J. Strateg Inf. Syst. **32**(3), 101772 (Sep. 2023). https://doi.org/10.1016/j.jsis.2023.101772
40. Minkkinen, M., Zimmer, M.P., Mäntymäki, M.: Co-shaping an ecosystem for responsible AI: Five types of expectation work in response to a Technological Frame. Inf. Syst. Front. (Apr. 2022). https://doi.org/10.1007/s10796-022-10269-2