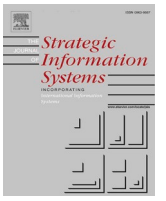
Journal of Strategic Information Systems 34 (2025) 101885

Contents lists available at ScienceDirect

Journal of Strategic Information Systems

journal homepage: www.elsevier.com/locate/jsis



Responsible artificial intelligence governance: A review and research framework

Emmanouil Papagiannidis [a], Patrick Mikalef [a,c,*], Kieran Conboy [b]

[a] Department of Computer Science, Norwegian University of Science and Technology, Norway
[b] School of Business & Economics, National University of Ireland, Galway, Ireland
[c] Department of Technology Management, SINTEF Digital, Norway

ARTICLE INFO

Keywords:
Artificial intelligence
Responsible AI governance
Governance practices
AI implementation
AI lifecycle

ABSTRACT

The widespread and rapid diffusion of artificial intelligence (AI) into all types of organizational activities necessitates the ethical and responsible deployment of these technologies. Various national and international policies, regulations, and guidelines aim to address this issue, and several organizations have developed frameworks detailing the principles of responsible AI. Nevertheless, the understanding of how such principles can be operationalized in designing, executing, monitoring, and evaluating AI applications is limited. The literature is disparate and lacks cohesion, clarity, and, in some cases, depth. Subsequently, this scoping review aims to synthesize and critically reflect on the research on responsible AI. Based on this synthesis, we developed a conceptual framework for responsible AI governance (defined through structural, relational, and procedural practices), its antecedents, and its effects. The framework serves as the foundation for developing an agenda for future research and critically reflects on the notion of responsible AI governance.

Introduction

Following a surge in data and computational capability, companies have increasingly turned to artificial intelligence (AI) to achieve a competitive edge (Arrieta et al., 2020; Holmström & Hällgren, 2021; Ransbotham et al., 2018). This has precipitated an explosion of academic research in this area, with numerous papers, special issues, conferences, and tracks emerging. Although the literature has noted several benefits of AI adoption (Arrieta et al., 2020; Ransbotham et al., 2018), for most organizations, AI has several potential ramifications (Schneider et al., 2022) and unexpected and unwanted outcomes (Di Vaio et al., 2020). This study defines AI as "the ability of a system to identify, interpret, make inferences, and learn from data to achieve predetermined organizational and societal goals" (Mikalef & Gupta, 2021).

The development of responsible principles to minimize AI's negative and unintended consequences has been a central point of discussion over the past years (Council of Europe, 2018; European Commission, 2019; Floridi et al., 2021; Hagendorff, 2020; Mökander & Floridi, 2021). Generally, these principles provide a guide and set of targets for designing and deploying AI to ensure that the technology is fair, equitable, ethical, and generally "good" for all those affected by it (Ghallab, 2019; Rakova et al., 2021). Despite extensive efforts to define the dimensions of responsible AI principles, research has largely focused on high-level guidelines, notably

* Corresponding author.
E-mail addresses: emmanouil.papagiannidis@ntnu.no (E. Papagiannidis), patrick.mikalef@ntnu.no (P. Mikalef), kieran.conboy@nuigalway.ie (K. Conboy).

https://doi.org/10.1016/j.jsis.2024.101885
Received 1 February 2023; Received in revised form 19 December 2024; Accepted 19 December 2024
Available online 3 January 2025

lacking the elements of governance (Butcher & Beridze, 2019; Schiff et al., 2021). Consequently, a gap exists in understanding how AI technologies are governed responsibly throughout their life cycles (Meske et al., 2022).

This review is motivated by the fact that adhering to responsible AI principles is generally deprioritized or considered an ancillary task during the actual implementation and management of AI projects (Mäntymäki et al., 2022). For instance, Meske et al. (2022) and Mannes (2020) demonstrate that organizations need certain trade-offs to find the right equilibrium between performance, transparency, and ethical conduct. While doing so provides some insight into the choices that organizations must make in response to developing and deploying AI applications, it does not provide a holistic or comprehensive understanding of how responsible AI governance practices are formulated and enacted. Hence, a significant challenge exists in translating theoretical principles into practical implementation approaches (Mäntymäki et al., 2022).

We aim to develop a more coherent understanding of how responsible AI governance can be comprehended and implemented in research and practice. Specifically, we address the lack of guidance needed to translate high-level abstract principles into deployable practices throughout the AI project lifecycle. We argue that bridging this gap facilitates a more comprehensive approach to responsibly developing and deploying AI technologies. Additionally, we understand how responsible AI governance is shaped depending on the context and its effects on organizations and their environments.

We achieve this by: (i) synthesizing prior research on the definition and principles of responsible AI; (ii) proposing a concept of responsible AI governance based on seven key principles of responsible AI spanning three types of organizational practices (structural, procedural, and relational); and (iii) discussing the broader context wherein responsible AI governance is developed and utilized, highlighting its key antecedents and effects, both internal and external. Placing responsible AI governance in a framework that highlights the broader context wherein it is developed and deployed enables us to uncover key assumptions and highlights important areas for future research. This paper concludes with important issues underpinning research and practice and develops a set of research questions to help guide future studies.

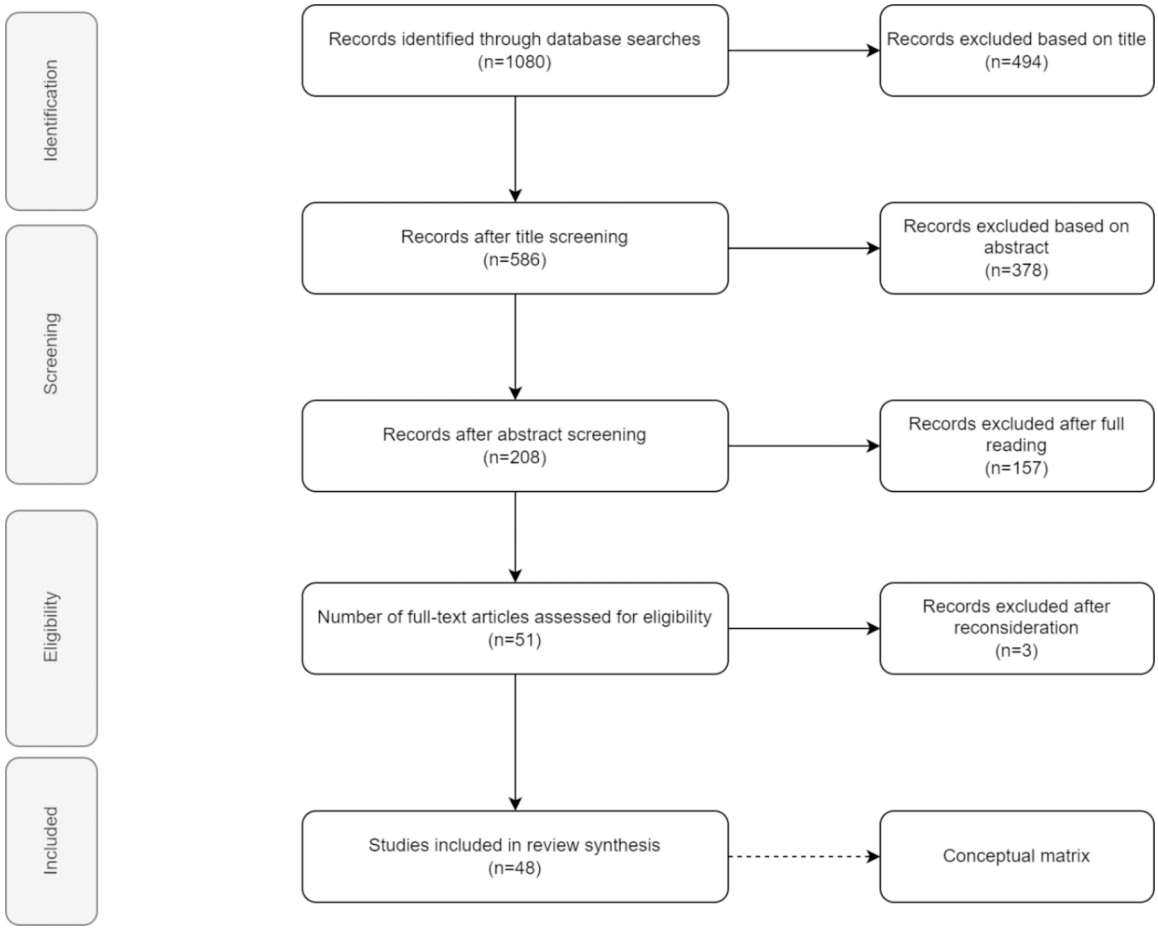


Fig. 1. Stages of literature search.

Research methodology

We conducted a systematic literature review comprising a sequence of steps to identify the relevant research work. In the following sub-sections, we present the protocol development process, the inclusion and exclusion criteria, the data sources and strategy for searching articles, and how we quality-checked the pool of papers and extracted data from them.

Protocol development

The systematic literature review was based on well-established procedures to ensure all relevant publications' inclusion (Kitchenham, 2004; Okoli, 2015; Rowe, 2014; Templier & Paré, 2015). We followed a scoping review approach (Paré et al., 2015), outlining how the primary research would be conducted, the search phrases used, and the sites to consult when gathering literature (Boell & Cecez-Kecmanovic, 2015). The review was conducted in five steps. Two researchers collaboratively examined the papers to mitigate bias during the selection process. Step one involved gathering information, resulting in the identification of 1,080 documents. The documents were iteratively filtered according to their relevance. All titles were reviewed, and 494 papers were excluded because of irrelevance or because the same paper appeared twice. Step two involved reviewing the abstracts, which resulted in the exclusion of 378 papers after carefully reading each to determine the inclusion or exclusion criteria. Step three, which involved critically viewing the approach in the studies, excluded 157 papers. Step four involved the extraction of data, which were then organized into a spreadsheet. Three papers did not conform to the inclusion criteria, resulting in a final sample of 48 papers. Step five included data synthesis, which involved using the concept matrix to structure the content of the articles. A concept matrix was used to draw connections between the different research articles. Fig. 1 depicts the steps for conducting this review.

Inclusion and exclusion criteria

To define this systematic literature review's scope, various inclusion and exclusion criteria were used to ensure that the selected studies aligned closely with the research objectives. Our criteria were designed to encompass a broad spectrum of research on AI's integration and impact in business and organizational contexts. These criteria include studies examining how AI contributes to digital transformation within businesses and how organizations leverage AI to address operational challenges. First, we examined whether there was a focus on AI in the organizational context. Second, we checked the publication dates. Considering AI's rapid evolution and increasing adoption in recent years, only studies published from 2017 onwards were considered. Third, we included only papers published in English to ensure accessibility and comprehension for a wider audience. Fourth, we filtered based on publication type, ensuring the inclusion of peer-reviewed journal articles and providing a substantial depth of analysis.

For the exclusion criteria, we first examined the technical focus of the articles. Studies that primarily focused on AI's technical aspects, such as architectural infrastructure or model benchmarking, were excluded. Second, we excluded certain publications – such as book series, disseminated articles, and webpages – from the selection process. Third, the articles were examined based on their publication status; pre-publication or under-review studies were excluded to maintain the included studies' integrity and quality.

Data sources and search strategy

In the first phase, a series of search strings were developed. The first group of terms (see Table 1) contained keywords linked to AI and related technologies, whereas the second set focused on organizational viewpoints. The terms used were treated as exact keywords; thus, to increase the number of search strings, keywords from both sets were concatenated to generate a search string using wildcard symbols. Thereafter, the search phrases were used in Scopus, Business Source Complete, Emerald, Taylor & Francis, Springer, Web of Knowledge, ABI/Inform Complete, IEEE Xplore, and the Association of Information Systems (AIS) libraries – as well as in other electronic databases, including ScienceDirect, JSTOR, Digital Bibliography & Library Project, and Google Scholar. This was done to ensure that the index contained all the relevant items. Data collection commenced in November 2022, followed by further refinement in December 2023.

Quality assessment

All documents were subjected to quality evaluation to improve their internal and external validity and eliminate bias. Each article was evaluated for relevance using three fundamental values. The first was to examine the biases in the data collection process of the identified articles; we assessed the validity of the research methodology employed therein. Second, we checked for internal validity,

Table 1
Keywords for literature search.

Category	Keywords
Responsible artificial intelligence (AI)	responsible AI, trustworthy AI, beneficial AI, ethical AI, explainable AI, AI ecosystem, AI dark side, design responsible AI, principled AI, fair AI
Context	governance, challenges, business value, business digitization, organizational challenges, public value, corporate values, risk management

which refers to how well the study's design and execution prevented systematic mistakes, suggesting that the examination would yield positive findings (Porritt et al., 2014). Nevertheless, it may lack intrinsic validity; therefore, internal validity was further investigated by examining the research's discussion and conclusion (Porritt et al., 2014). Following the eligibility check, two coauthors independently reviewed the papers and evaluated their quality using multiple criteria. These criteria include study design and methodology, sample size and representativeness, data quality, ethical considerations, data interpretation, results, and reporting.

When the two coders disagreed, a third researcher provided additional perspective and expertise to help mediate and resolve disagreements. The third researcher contributed by offering insights, conducting further analyses, and facilitating discussions between the first two researchers to reach a consensus on the papers' quality. This collaborative approach ensured a more robust and unbiased evaluation. The scientific rigor and relevance of these studies were also reviewed. Specifically, we examined the credibility and impact of the retained studies. These criteria helped us produce reliable findings with less bias that can be generalized to broader populations or contexts.

Data extraction and synthesis of findings

The first step in data extraction was identifying all the relevant materials within the final pool of papers. A concept matrix was developed to classify the investigations and combine data. The matrix includes information regarding the data extraction date, title, authors, journal, publishing details, and topic-specific information. The extraction was conducted following the guidelines of Kitchenham (2004). Furthermore, arranging the data on a spreadsheet made comparing the data from each article easier. The studies were analyzed based on 15 distinct values. Table 2 presents each of these values and the key categories and themes they belong to. The need for sub-views arose because each responsible principle may encompass diverse viewpoints and techniques. These perspectives were used as labels to depict the perspective of each article. Nevertheless, some labels may not apply to all responsible principles, resulting in missing values.

The data were synthesized after inserting and analyzing 48 entries into the concept matrix. During this stage, we systematically combined and summarized data from multiple sources to draw meaningful conclusions and generate new knowledge (Grimshaw et al., 2001). This study aimed to fulfill the standards of scoping synthesis and provide an in-depth overview of the available evidence (Arksey & O'Malley, 2005). Additionally, we utilized a descriptive synthesis that helped map materials from several studies and presented them to research streams (Sandelowski & Barroso, 2006). Furthermore, we examined the consistencies and inconsistencies of responsible AI governance and compared different studies on the same topic.

Definitional aspects of responsible AI

In the last decade, how AI should be developed and AI deployment based on responsible principles have received significant attention (Dignum, 2019b; Ghallab, 2019). Questions such as what responsible AI is and what governance practices should be applied to enact it remain unanswered. In recent years, reports have provided conceptualizations and descriptions of responsible AI principles (de Almeida et al., 2021; Freiman, 2023; IBM, 2019; Singapore Government, 2020; Smuha, 2021). This trend toward an increased number of articles on responsible AI is largely attributable to the growing number of incidents in which the use of AI leads to unforeseen or undesirable repercussions (Fuchs, 2018). For instance, Amazon has been developing AI applications to automate the process of analyzing resumes to identify top vacancy candidates (Kodiyan, 2019). In 2015, Amazon's machine learning (ML) experts found that its AI-powered recruitment tool discriminated against women when recruiting technical professionals, such as software developers. These algorithms were partially trained on resumes submitted to the corporation over the previous 10 years, during which, most successful resumes were disproportionately from male applicants (Dastin, 2022). Such cases have spurred policymakers,

Table 2
Themes extracted from the concept matrix.

Category	Value
Artificial intelligence (AI)	Types of AI
	AI capabilities
	Definitions of AI
Responsible principles	Accountability
	Human agency and oversight
	Technical robustness and safety
	Privacy and data governance
	Transparency
	Diversity, non-discrimination, and fairness
	Societal and environmental well-being
	Responsible AI
AI governance	Definition of AI governance
	Governance capabilities
	Organizational level outcomes
	Business values achieved through governance

researchers, and practitioners to consider how AI development and use should adhere to "responsible" norms. Therefore, discussing the core principles of responsible AI is critical.

Responsible AI principles

Governments, researchers, and corporations are increasingly focusing on AI-related ethical standards and principles. Numerous reports have been published outlining the significance of these principles and the reasons behind their importance. However, determining what constitutes responsible AI is a work in progress, with several organizational bodies and researchers aiming to provide complete and coherent conceptualizations (Wu et al., 2020). While previous research has focused on AI's specific aspects – such as bias elimination (Brighton & Gigerenzer, 2015), the explainability of AI outcomes (Arrieta et al., 2020), and safety and security (Hernández-Orallo et al., 2020) – recent years have observed a shift toward a more holistic understanding of responsible AI's constituents (Theodorou & Dignum, 2020). The European Commission recently requested an independent expert body – the High-Level Expert Group on Artificial Intelligence (AI HLEG) – to develop an integrated framework for responsible and trustworthy AI (European Commission, 2019). Meanwhile, the Singapore government (2020) recognized the forthcoming AI difficulties regarding discrimination, biased outcomes, and concerns linked to consumer awareness and understanding of AI engagement in decision outcomes in the ethical, legal, and governance sectors (Trocin et al., 2021). Simultaneously, there is a push for independent bodies to certify responsible AI best-practice advocacy from corporations, such as Google.

Regarding responsible AI practices, the word *responsible* can be interpreted in various ways. For example, the AI HLEG promotes trustworthy AI with three main necessary components: the system in question should be (1) *lawful*, complying with all applicable laws and regulations; (2) *ethical*, ensuring adherence to ethical principles and values; and (3) *robust*, having the ability to withstand and adapt to different challenges and disruptions in the environment that encompasses both social and technical elements. Similarly, Singapore Government (2020) framework is based on two high-level guiding concepts that foster AI trust. The first concerns companies that use AI to make decisions and ensure transparent, explainable, and fair processes (Nishant et al., 2023). Although absolute explainability, transparency, and fairness are difficult to achieve, companies should invest every effort to ensure these values, thereby contributing to AI development (Feuerriegel et al., 2020). The second category comprises human-centered AI solutions (Shneiderman, 2020b). Human interests, including well-being and safety, should be key considerations in designing, developing, and deploying AI, as it augments human skills (Yerlikaya & Erzurumlu, 2021). Therefore, businesses should ensure that human-centric decision-making processes adhere to ethical norms (Brendel et al., 2021).

A recent Harvard report highlighted 38 similar corporate and group efforts (Fjeld et al., 2020). An underlying agreement exists that responsible AI represents a set of principles assuring ethical, transparent, and accountable usage of AI technology per user expectations, corporate values, and societal laws and conventions based on responsible AI's emerging consensus (Flavián & Casaló, 2021). In this sense, responsible AI encompasses a wide range of standards that must be satisfied throughout the lifecycle of AI applications (European Commission, 2019; Telia, 2019). Winfield and Jirotka (2018) describe responsible principles as a collection of processes, procedures, cultures, and beliefs that ensure the highest levels of conduct. They emphasize AI governance's ethical side, suggesting that responsible AI transcends principles and instills ethical behaviors in individuals and companies. Winfield and Jirotka (2018) claim that these are critical components of responsible research and development, which "entails an approach, rather than a mechanism; hence, they seek to tackle ethical issues before they arise in a principled manner rather than waiting until a problem surfaces and dealing with it in an ad-hoc way".

Building on these concentrated efforts, responsible AI principles can be divided into *accountability, diversity, non-discrimination and fairness, human agency and oversight, privacy and data governance, technical robustness and safety, transparency, and social and environmental well-being* (European Commission, 2019; Janssen et al., 2020). Several of these principles appear in different reports using various terms. For instance, "transparency" might appear as "transparency and explainability," while "human agency and oversight" might be noted as "human control of technology." The terms describe the same principles, with no major differences, signifying an ongoing conceptualization process. Through an open consultation process, these guidelines describe the key components of responsible AI principles (Mikalef et al., 2022).

The discussion pertains to the need to establish a set of responsible principles emerging from challenges specific to AI technologies. These include effectively governing and controlling autonomous intelligent systems, establishing responsibility and accountability for algorithms, and ensuring privacy and data security in opaque and multilayered systems (Wirtz et al., 2020). Responsible AI has gained traction at the policymaking level as a testament to its importance, with several countries defining responsible AI's fundamental principles (Jobin et al., 2019). At the national level, the AI readiness index measures the degree to which countries are implementing AI technologies; they now include a new sub-index that quantifies the degree to which responsible AI principles are adopted (Nzobonimpa & Savard, 2023).

When considering more in-depth the issues potentially arising when using AI, several situations may be preempted using responsible AI principles. The complexity of AI renders it difficult to comprehend and interpret the final outcomes, frequently rendering the results opaque. This phenomenon is commonly called a "black box," whereby AI may implement unforeseeable actions or suggest outcomes that are difficult to trace. This may be exacerbated when AI gains the autonomy to pursue its own objectives, even if it unintentionally harms others (Nath & Levinson, 2014). Such instances raise concerns regarding AI transparency in decision-making processes and accountability for the outcomes of AI use. Consequently, responsibility and accountability are important concepts in governance and regulation. Operators or developers of AI systems cannot accurately predict with exact certainty all actions and results generated by the self-learning ability of AI algorithms at any given time. Therefore, carefully assessing the actors and regulation of transparent and explainable AI systems is necessary (Helbing, 2019).

Mass data reuse and ubiquitous digitalization have become global drivers of competitiveness. Furthermore, AI has been widely described as having the potential to vastly improve efficiency across all domains and sectors and help resolve humanity's greatest challenges, such as the United Nations Sustainable Development Goals (Jelinek et al., 2021). Nevertheless, the swift and extensive adoption of narrow AI – with notable attributes such as efficiency, scalability, performance, decision automation, and speculative progression toward general AI – has raised significant concerns, which have prompted extensive research into the concepts of human dignity and existence. (Jelinek et al., 2021). Human dignity can be jeopardized in situations involving job displacement, loss of privacy and surveillance, and loss of control or autonomy over AI systems. This landscape of direct threats and structural imbalances increases the urgency to develop appropriate governance solutions (Jelinek et al., 2021). The coordination of the development and implementation of responsible AI principles is not mutually exclusive. They complement each other in building trustworthy and ethical AI systems (Lubberink et al., 2019). Therefore, understanding how the design, development, and implementation of AI applications can be infused with the key principles of responsibility is necessary.

Responsible AI governance

Responsible AI governance has been conceptualized as a framework that encapsulates the practices that organizations must implement in their AI design, development, and implementation to ensure AI systems' trustworthiness and safety. Responsible AI governance concerns delegating authority and control over data (Brackett & Earley, 2017) and exercising authority through data-related decision-making (Conboy et al., 2020). Awareness and understanding of AI's impact are crucial for effective governance, as AI-educated individuals are essential pillars of any successful AI system. Thus, responsible AI governance should incorporate incentives and sanctions to encourage desirable data collection, administration, and utilization behaviors (Margetts, 2022). Furthermore, responsible AI governance relies on collaboration between firms and individuals who comprise the system and extends beyond a single company (Mäntymäki et al., 2022). This multi-organizational context necessitates trusted frameworks to ensure dependable data sharing among organizations while adhering to the General Data Protection Regulation (GDPR) and other applicable laws and regulations (Mills et al., 2020). Accordingly, we define responsible AI governance as follows:

A set of practices for developing, deploying, and monitoring AI applications in a safe, trustworthy, and ethical manner that ensures appropriate functionality of AI over the entire lifecycle.

Although clear definitions of responsible AI exist (Schneider et al., 2023), the literature uses terms that are synonymous or largely overlapping (see Table 3). For example, a large stream of research refers to "trustworthy AI," while others use the term "principled AI." We argue that this divergence in naming stems from the immaturity of responsible AI as a research concept. Researchers have built on various definitions and themes that encompass responsible AI practices and dimensions. This conceptual opaqueness makes it challenging to determine what responsible AI governance should include.

The overarching objective of frameworks is to maximize the value of AI while simultaneously lowering the associated risks and

Table 3
Description of artificial intelligence (AI) governance terms.

Name	Description	References
Principled AI	Eight themes—namely, privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility, and promotion of human values—were derived from 35 papers.	(Clarke, 2019; Fjeld et al., 2020)
	An ethical framework of AI specifying five core principles—namely, beneficence, nonmaleficence, autonomy, justice, and explicability—is defined.	(Floridi & Cowls, 2021; Thiebes et al., 2021)
	More than 20 principles of beneficial AI, organized into three categories, are described as follows: 1. Research issues—Research Goals, Research Funding, Science-Policy Link, Research Culture, and Race Avoidance; 2. Ethics and values—Safety, Failure Transparency, Judicial Transparency, Responsibility, Value Alignment, Human Values, Personal Privacy, Liberty and Privacy, Shared Benefit, Shared Prosperity, Human Control, and Non-subversion; 3. Long-term issues—AI Arms Race, Capability Caution, Importance, Risks, Recursive Self-Improvement, and Common Good.	(Future of Life Institute, 2017; Pagallo et al., 2019)
Responsible AI	Frameworks for developing responsible AI based on 10 principles—namely, well-being, respect for autonomy, privacy and intimacy, solidarity, democratic participation, equity, diversity inclusion, prudence, responsibility, and sustainable development—are elucidated.	(Dignum, 2017, 2019a; Liu et al., 2022)
	Explainable AI is a suite of algorithmic techniques generating high-performance, explainable, and trustworthy models.	(Adadi & Berrada, 2018; Kaur et al., 2022; Li et al., 2021; Zou & Schiebinger, 2018)
Trustworthy AI	The principles of trustworthy AI (TAI) are defined, and based on these, seven key requirements for achieving TAI are derived. Further, an assessment list is provided for operationalizing the seven key requirements. These include accountability, human agency and oversight, technical robustness and safety, privacy and data governance, transparency, fairness, and societal and environmental well-being.	(Chatila et al., 2021; European Commission, 2019; Mora-Cantallops et al., 2021; Theodorou & Dignum, 2020; Wu et al., 2020; Zicari et al., 2021)

unintended consequences (Abraham et al., 2019; Mikalef et al., 2019). However, those aiming to implement responsible AI governance risks are undermined in two ways. First, the diversity and breadth of responsible AI principles render it challenging for organizations to implement practices that cover a variety of goals. Thus, a principles-first approach may prove counterproductive and incompatible with the organizational *modus operandi*. Second, responsible AI principles are in a continuous state of flux, whereby every new type of emergent AI technology comes into a new set of conditions that must be considered. An indicative example is the recent release of Open AI's ChatGPT, which poses a novel dilemma regarding human creativity and innovation.[1] Collectively, these issues suggest that we must reconsider how we conceptualize and approach responsible AI governance, with the caveat that the notion entails actionable practices for AI applications' design, deployment, and oversight throughout their lifecycle.

Synthesis of responsible AI principles

Based on the current discussion around frameworks for responsible AI, the following sub-sections present a synthesis of research findings based on the seven pillars of key principles and their underlying sub-dimensions (Table 4).

Accountability

Accountability involves the implementation of mechanisms and processes to ensure responsibility and auditability during and after AI development and deployment. Auditability is fundamental to every project and establishes the foundation for data selection and system architecture. Vollmer et al. (2020) propose that auditability should be assessed during data accumulation, focusing on data inspection concerning its usage and addressing questions regarding dataset distribution and sample representation. Establishing error-reporting mechanisms is critical because it allows data comparison at different stages. Similar approaches have been proposed for data collection in autonomous vehicles (Shneiderman, 2020a). These procedures may not guarantee ongoing accountability but ensure that the data trace possible deviations and those responsible for them.

Although AI can produce accurate analytical findings, implementing accountability procedures is typically challenging because of its "black-box" nature (Caner & Bhatti, 2020). Hence, technologies for detecting weaknesses and critically evaluating AI systems are essential (Matthews, 2020). Screening CVs as part of the recruitment process exemplifies the challenge of embedding accountability principles in AI systems. Unlike human recruiters, AI systems cannot be held personally responsible for filtering candidates. The question of who is accountable for a judgment made by an AI system is difficult to resolve and multilayered (Ayling & Chapman, 2021). Some businesses early in implementing AI for such processes have found it challenging to account for the decisions taken. Following these cases, organizations have been criticized for lacking accountability in using AI systems (Schlögl et al., 2019). Specifically, accurately capturing where responsibility lies and auditing such systems transparently and easily have been a key focus.

Diversity, Non-Discrimination, and fairness

Another fundamental principle of responsible AI is ensuring that systems do not reproduce discrimination or unfairness (Korinek, 2020). Recent real-life applications have highlighted how AI systems produce discriminating results based on their training data (Varona & Suárez, 2022). Two popular areas are credit ratings and criminal sentences (Taeihagh, 2021). The European Commission (2019) recommends that AI systems use appropriate mathematical and statistical methodologies to uncover unintended behavior. According to Korinek (2020), the European Commission has developed the only method for eliminating algorithmic bias whereby fairness primarily entails accessibility and the absence of unjust prejudice. Accordingly, systems should be user-centric, allowing all individuals to utilize AI products irrespective of age, gender, abilities, and characteristics.

Other forms of prejudice and discrimination concern language because it is highly complicated and includes features such as word grouping and ordering rules. Using natural language datasets to train models precipitates various biases, and detecting this prejudice may be challenging. For example, word groups of men and women, adjectives associated with them, and the frequency or order in which they appear in a list may all encourage bias in the dataset and alter the model (Leavy, 2018). Hence, one can comprehend the potential of placing a biased dataset into a "black-box" system, causing unpredictable and unjust consequences (Jakesch et al., 2022). Bias awareness can be realized by addressing record distribution and ensuring relevant updated data. Furthermore, data processing and analysis should be based on core ethical values to reduce any disparities or prejudice as much as possible (Ayling & Chapman, 2021; Gasser & Almeida, 2017).

Human agency and oversight

The principles of human agency and oversight guarantee that AI systems adhere to democratic, prosperous, and equitable societal values (European Commission, 2019). The user's knowledge and interpretation of AI system outcomes are the focus of human agency. By contrast, human review refers to the presence of humans in an AI's decision-making process. The most widely utilized strategies for supervision are planning oversight, continuous monitoring, and retrospective disaster analysis. Planning oversight entails assessing proposals ahead of time, allowing for an examination of the chosen technologies and an understanding of their influence in the context

[1] https://medium.com/@jonbello/introducing-chatgpt-a-threat-to-human-creativity-and-innovation-42903307065f.

Table 4
Responsible artificial intelligence (AI) principles.

Principle	Sub-dimensions	References
Accountability	**Auditability:** ability to assess AI applications concerning the algorithms, data, and design processes.	(de Almeida et al., 2021; European Commission, 2019; Mikalef et al., 2022)
	Responsibility: oversight of the various stages and activities involved in AI deployment and how it should be allocated to people, roles, or departments	
Diversity, non-discrimination and fairness	**Accessibility:** design of systems in a manner that makes them accessible and usable for everyone, regardless of age, gender, abilities, and characteristics	(Fjeld et al., 2020; Singapore Government, 2020)
	No unfair bias: inclination of prejudice toward or against people, objects, or positions, as well as inherent biases in datasets, which can precipitate undesirable outcomes	
Human agency and oversight	**Human review:** right of a person to challenge a decision made by an AI	(European Commission, 2019; Singapore Government, 2020)
	Human well-being: the notion that AI must include human well-being as a primary success factor for development	
Privacy and data governance	**Data quality:** accuracy of values in a dataset, matching the true characteristics of the entities described by the dataset	(Matthews, 2020; Singapore Government, 2020)
	Data privacy: AI systems' development and operation in a manner that considers data privacy throughout the data lifecycle	
	Data Access: national and international rights laws during the design of an AI for data access permissions	
Technical robustness and safety	**Accuracy:** AI system's ability to make correct judgments, such as correctly classifying information into the appropriate categories or being able to predict, recommend, or make intelligent decisions based on data or data models	(European Commission, 2019; Singapore Government, 2020)
	Reliability: AI system's ability to work properly when subjected to a range of inputs or situational contexts	
	General Safety: safety rules and fallback plans that should be established for AI systems in the event of problems	
	Resilience: AI systems that should be protected against vulnerabilities that adversaries can exploit, e.g., hacking	
Transparency	**Explainability:** ability to explain the technical processes of an AI system and related human decisions (e.g., application areas of a system)	(Fjeld et al., 2020; Mikalef et al., 2022; Singapore Government, 2020)
	Communication: human right to be informed in advance when interacting with an AI agent	
	Traceability: ability to track data and processes that yield the AI system's decision, including data gathering, labeling, and algorithms.	
Social and environmental well-being	**Social well-being:** ubiquitous exposure to social AI systems in all areas of society, such as work and education.	(European Commission, 2019; Singapore Government, 2020)
	Environmental well-being: most pressing environmental and climate concerns facing the planet	

of use before they are adopted. Continuous monitoring refers to the ongoing inspection and correction of a system at regular intervals, which is beneficial for systems that operate in dynamic and uncertain environments. Finally, retrospective disaster analysis refers to a detailed system examination following a serious incident (Shneiderman, 2020a).

The extent to which humans govern and monitor AI systems is a contention among enterprises. According to Tolmeijer (2022), two ways of engagement exist. First, "human in the loop" refers to AI suggesting recommendations while a human makes the final choice—also termed "assisted intelligence." Second, "human out of the loop" follows the logic whereby the system makes the final decisions, with developers tweaking the models to obtain the desired result. Autonomous intelligence is a term used to characterize such systems (e.g., self-driving cars and automated stock market trading systems) (Korinek, 2020). Overseeing systems is a proposed mitigation approach for ensuring system anticipation – accomplished by creating a system wherein stakeholders voice ethical concerns (Winfield & Jirotka, 2018) and such concerns are incorporated into the revised models of the AI system (Vollmer et al., 2020). However, using AI as a support system might precipitate a moral dilemma because people may exploit the system to absolve their moral responsibility for their actions. This tendency to place responsibility for system malfunctions is particularly evident in systems lacking transparency in their inner workings and with limited auditability.

Privacy and data governance

Privacy and data governance refers to the principles for managing the availability and usability (data access), integrity (data quality), and security (data privacy) of enterprise data when developing AI systems. Building on formalized processes is one possible means to ensure privacy concerns regarding data gathering that can be accomplished in different ways. First, developers who create algorithms should provide documentation about the data life cycle; moreover, a continuous exploration and sensitivity evaluation of both the data-gathering methods and the data itself is critical (Matthews, 2020). Second, without a legal compliance assessment, firms might struggle to adhere to various laws regarding data collection and processing, as they are typically subject to different national and international directives (Google, 2019). This can have significant repercussions on the maturation and use of AI models. Finally, cultural variations should be considered during data collection. This is especially important when considering country variations; in certain contexts, a stronger or weaker link exists between private enterprises and government organizations, which has ripple effects on what data are collected and how they are used (Ayling & Chapman, 2021).

Technical robustness and safety

Technical robustness – closely linked to damage prevention – involves creating AI systems that are proactive in their approach to risks and consistently perform as intended while minimizing unintentional and unanticipated harm and preventing unacceptable harm (European Commission, 2019). This should also apply to potential changes in operating environments and the presence of other agents (both human and artificial) that may interact antagonistically with AI models. Additionally, AI systems must protect (resilience to attacks) against flaws that could allow adversaries to exploit them or gain unauthorized access. Data poisoning (a malicious or adversarial attack in which an attacker manipulates the training data used to train an AI or ML model) (Wang & Chaudhuri, 2018)), model leakage, and attacks on the underlying infrastructure (both software and hardware) pose risks to AI applications' proper functioning (Hamon et al., 2020).

When AI systems are targeted, data and system behavior can be altered, precipitating significant deviations in outcomes and the underlying logic of operation. Hence, AI systems must have a contingency plan (general safety) for when such attacks occur to ensure continuous operation (Smuha, 2021). The extent to which safety precautions are necessary is determined by the scale of the risk posed by an AI system, which, in turn, is determined by the system's capabilities and the consequences' severity (Hamon et al., 2020). When the development process or system itself is anticipated to offer exceptionally high risks, establishing and testing safety measures in advance is critical; therefore, accuracy is crucial. The capacity of AI to make accurate judgments, such as classifying information into appropriate categories or making precise forecasts and suggestions, varies depending on its application. The unintended risks from faulty forecasts can be supported, mitigated, and corrected through well-formed development and review processes (Kuziemski & Misuraca, 2020). In situations where occasional incorrect predictions cannot be prevented, the system must demonstrate probability errors, particularly when human life is at risk. Thus, the AI system's results must be both reproducible and dependable (Chang et al., 2022).

Transparency

A prevalent concern regarding AI systems is that, despite their potential widespread use in everyday applications, knowledge about their functioning is limited (Gasser & Almeida, 2017). Consequently, there is a lack of understanding and, consequently, a lack of trust among users, who may be hesitant to use AI (Toreini et al., 2020). A key aspect of transparency is explainability – the ability to explain both the technological processes of an AI system and the human judgments resulting from these procedures (Larsson et al., 2019). Owing to these procedures, a new subfield called explainable AI (XAI) has emerged to provide human-comprehensible models and interpretations of complex machine-based calculations (Gillath et al., 2021). Nevertheless, trade-offs may be required to improve the explainability of a system (at the expense of accuracy) and increase its accuracy (at the expense of explainability). When an AI system substantially influences people's lives, an appropriate explanation of the system's decision-making process should be available on demand. Such explanations should be timely and tailored to stakeholders' knowledge (e.g., laypersons, regulators, or researchers).

Transparency is necessary to address any subsequent need for traceability. The datasets and procedures that lead to the AI system's conclusion, such as data collection and labeling and the algorithms utilized, should be documented to the highest degree of feasibility (Reddy et al., 2020). This also holds for AI system decisions and allows the discovery of the logic driving an incorrect AI judgment, which may aid in preventing future errors (Mezgár, 2021). Additionally, users should be aware and clearly informed when interacting with an AI agent rather than a human actor. This necessitates that AI systems be identified explicitly. Furthermore, the option to opt out of AI interaction in favor of human interaction should be offered to respect the fundamental rights of communication (Felzmann et al., 2020). Additionally, the AI system's capabilities and limits should be conveyed to end users suitably, including sharing advanced information on the AI system's accuracy and limitations. Collectively, these form the principles of communication.

Social and environmental Well-Being

The sustainability and ecological responsibility of AI systems are also elements of responsible AI use that have been highlighted frequently. Furthermore, AI applications should be built on the logic of addressing global challenges, such as ensuring social well-being and protecting the environment (European Commission, 2019), and can potentially solve some of society's most pressing problems. However, they must be designed in an environmentally friendly manner (Pan & Nishant, 2023). In this regard, a system's development,

deployment, and usage processes should be examined through critical studies on resource use and energy consumption (Venkataramanan et al., 2019). Concerns have been raised regarding how AI may lead to job displacement and how the replacement of jobs by AI systems may precipitate the emergence of new ways of social organization (Gasser & Almeida, 2017). These issues are sensitive and complex, as different societies may have different views on how to solve them; thus, these issues should be addressed locally and as needed (Donati et al., 2022).

Additionally, exposure to social AI systems in various contexts may alter beliefs regarding social agency and affect social interactions (social well-being). For example, AI can replace people in hazardous occupations, such as mining and quarrying (Zhang et al., 2021). This may be perceived as a threat to certain professions. However, it may also increase workplace safety if such tasks are performed by robotic agents built on AI. Furthermore, some solutions may result in "cold care" when people are replaced by AI agents. Hence, AI systems can help to provide care for those in need, but simultaneously, they may also degrade the level of social interaction (Siala & Wang, 2022). Moreover, new threats might emerge because AI could be misused in democratic processes, such as political and electoral decision-making contexts (Winfield & Jirotka, 2018; Wirtz et al., 2020). Consequently, these systems' effects must be thoroughly considered and planned.

Critical reflection & research framework

In the previous sections, we briefly reviewed the key principles on which responsible AI should be built. Nevertheless, a consolidated framework for understanding how these principles are implemented and how they shape and are shaped by society is lacking (Seppälä et al., 2021). Therefore, we build on the notion of *responsible AI governance* as a central concept for the effective infusion of AI systems with responsibility principles (Hilb, 2020). The research agenda in this section discusses the antecedents, key components, and effects of responsible AI governance (Fig. 2).

The key components of responsible AI governance are based on the works of Van Grembergen et al. (2004) and Tallon et al. (2013), who highlight the structural, procedural, and relational practices that organizations must consider as part of their governance. Structural practices include assigning roles and responsibilities for decision-making around AI. Procedural practices concern how organizations execute responsible AI governance and include aspects that concern a series of actions at different levels. Relational practices concern the different links among employees within and outside the organization, as well as the means for developing the skills and knowledge of human capital. This distinction is made because it enables the identification of different types of practices that are relevant at various levels within the organization. Table 5 summarizes the future research questions on responsible AI governance.

Antecedents of responsible AI governance

The antecedents of responsible AI governance can be divided into three broad, interdependent themes. The first refers to societal expectations and norms, which are unwritten rules of behavior shared by society and adopted by organizations, as well as more formalized regulations and directives. The second group concerns the organizational values that predate the use of AI and characterize corporate culture. Finally, the third group of antecedents involves the responsible AI principles of organizations. As these are contextual, fluid, and constantly changing, they represent key directives toward which responsible AI governance must be aligned. In Fig. 2, we also visually depict the relationship between these groups of antecedents, where broader societal and normative beliefs indirectly impact the prioritized responsible AI principles that firms decide to adhere to. These factors are mediated by how organizational values and culture adopt external stimuli and the importance they assign to adhering to responsible AI principles.

Societal expectations and norms can be considered the starting point of the antecedents of responsible AI governance. Social norms shape the AI principles that should be followed, especially those perceived as ethical and acceptable in a social context. Organizations have shifted their operating paradigms through various top-down and bottom-up efforts (Ji-fan Ren et al., 2017). This reevaluation of core beliefs aims to project or maintain a good public image, as organizations need to foster a good reputation within their operational context. Additionally, evolving regulations and directives such as the AI Act influence which aspects of governance are prioritized and how practices are designed and implemented.[2] Nevertheless, an issue that aligns with societal norms and expectations is that they are in constant flux. Therefore, organizations must develop appropriate mechanisms to identify external stimuli quickly and accurately and interpret how they affect their AI applications. An example of such a case is the controversy surrounding Google's Gemini image generator, which faced backlash from many users who accused the AI service of being "woke".[3] Such societal signals are often challenging to identify and require different mechanisms to act upon formalized regulations and directives. Organizations must develop different mechanisms to identify emerging external pressures and create appropriate channels for filtering and adaptation.

Apart from external signals and feedback, an essential part of what organizations implement in their responsible AI governance frameworks relates to internal corporate values and capabilities (Jöhnk et al., 2021; Papagiannidis et al., 2022). Organizations must foster a culture where individuals are encouraged to acknowledge and respond to external stimuli that are perceived as significant. Several internal factors mediate the extent to which such signals are leveraged – including the style of decision-making, concentration of power, level of democratic and inclusive organization in the firm, and other contextual and industry-specific aspects. These contingencies mediate how effective and adaptive organizations are in identifying changing external stimuli and incorporating them into their principles of AI governance. An example of this is an organization's capability to capture data influenced by norms and redefine

[2] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.
[3] https://www.vox.com/future-perfect/2024/2/28/24083814/google-gemini-ai-bias-ethics.

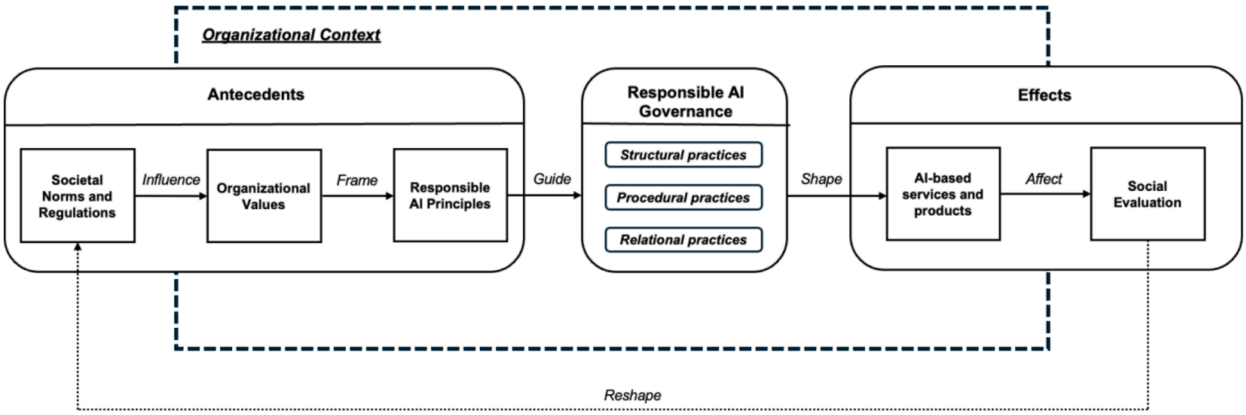


Fig. 2. Framework of antecedents, practices, and effects of responsible AI governance.

its responsible AI principles in a relatively short period. Nevertheless, we still have limited knowledge concerning how organizational path dependencies, culture, and other internal factors influence an organization's ability to be receptive and adaptive to such signals and, therefore, to effectively roll out appropriate responsible AI governance schemes.

Another important issue that organizations must consider when formulating responsible AI governance practices is cultural and ethical variations when deploying systems or services to different countries or populations. One of the generally held assumptions of many responsible AI frameworks is that there is uniformity in the elements and importance of aspects noted as key pillars. However, a challenge with such an approach is the considerable diversity among cultures and even segments of user groups concerning what is ethical and appropriate. Thus, responsible AI governance should consider the requirement to accommodate such variation. Key decision-makers should be conscious of potential user groups and how they may perceive the elements of use and design of AI systems with which they will interact.

Responsible AI governance

Structural practices

Structural practices in the context of responsible AI governance describe the key decision-makers and the rights and responsibilities of individuals and user groups within an organization. Such structural practices, in turn, impact various phases of AI development and deployment, as well as different levels within the organization. Within the context of structural practices, AI governance committees are responsible for overseeing AI initiatives within an organization (Radu, 2021). Structural practices should outline the criteria for selecting committee members, ensuring that key decision-makers with relevant domain knowledge and strategic insights are included (Salunke et al., 2011). Furthermore, practitioners should clarify roles, responsibilities, and decision-making by establishing clear approaches for effective oversight and fostering alignment with organizational objectives and ethical standards. Simultaneously, structural practices include the implementation of decision-making protocols within AI governance structures that provide consistency in the decision-making process (Janssen et al., 2020). Such practices should also specify the roles and responsibilities of individuals and user groups involved in decision-making to ensure accountability and transparency (Shneiderman, 2020a). By establishing clear decision-making structures, organizations can enhance agility, responsiveness, and integrity in their AI governance, thus driving better outcomes by mitigating risks (Felzmann et al., 2020).

Structural practices articulate the rights and responsibilities of stakeholders engaged in AI development and deployment (Yeung et al., 2020). These practices should provide a clear and concise framework for understanding the roles, obligations, and expectations of the individuals and user groups involved in AI initiatives. Within this set of practices, documentation should be provided for drafting, reviewing, and approving rights and responsibilities, ensuring alignment with organizational values, regulatory requirements, and ethical guidelines (Too & Weaver, 2014). Moreover, the procedures should establish mechanisms for regular reviews and update the documentation of rights and responsibilities to accommodate evolving organizational needs and industry standards (Van Grembergen et al., 2004). By formulating robust documentation on rights and responsibilities, organizations can promote transparency, accountability, and trust in AI-related activities, thus fostering a culture of ethical and responsible AI governance (Ashok et al., 2022). Nevertheless, there is limited knowledge of how organizations should define their structural practices within and outside firm boundaries. Within organizational boundaries, there is a lack of research identifying how rights and responsibilities should be assigned for vertical decision-making and how control and power dynamics influence the effectiveness of AI projects. Additionally, there is a limited understanding of how horizontal coordination and decision-making are enacted and how different departments within organizations coordinate actions at the respective levels. Similarly, the extended stakeholder ecosystem for formulating responsible AI governance has received limited attention. Understanding how different stakeholders can optimally provide input and exert influence on AI projects is critical to ensuring that AI applications meet the requirements and are ethically aligned.

Table 5
Issues and research agenda on responsible artificial intelligence (AI) governance.

Theme	Issue	Description	Research questions
Antecedents	Dynamic nature of societal norms and values	Ethical norms and expectations from society are constantly evolving. Designing and developing AI systems that adhere to these norms and regulations becomes a continuous process.	• How can organizations sense changing cultural norms and values affecting AI? • What mechanisms are effective in gauging societal norms and values vis-à-vis AI applications? • How do evolving regulations and directives influence the uptake and interpretation of responsible AI governance?
	Prevailing organizational values and path dependencies	Organizations typically filter and assimilate current societal norms and values through the perspectives of prevailing organizational values. Prior path dependencies may condition how such signals are perceived and which ones are acted upon.	• What role do organizational values have in filtering and acting upon societal norms and values in AI applications? • How do path dependencies influence the adaptability of organizations to assimilate emerging societal norms and values related to AI?
	Cultural and Ethical Variations	Differences in cultural and ethical norms make applying responsible AI principles universally challenging.	• What do cultural and ethical variations mean for designing and implementing AI applications? • How can responsible AI governance practices be developed to facilitate cultural and ethical variation?
Structural practices	Multilevel structures	Responsible AI governance concerns different organizational levels that must be aligned and coordinated.	• How should organizations develop vertical and horizontal structures to attain responsible AI outcomes? • What types of structures are best suited for vertical governance of responsible AI within organizations? • What are the key actors in horizontal structures that should be included in decision-making bodies related to responsible AI? • How can responsible AI governance practices be scaled up in organizations by gradually involving more levels and departments?
	Inter- and extra-organizational structures	Concerning AI applications, organizations must collaborate, communicate, and engage with different actors outside of their organizational boundaries; hence, appropriate structures must be established.	• How should organizations develop appropriate structures to involve end users and relevant stakeholders in their responsible AI governance? • How can organizations develop responsible AI governance structures that facilitate coopetition? • What are the potential risks or drawbacks of involving external parties in responsible AI governance practices?
Procedural practices	Strategic planning	Responsible AI governance requires strategic planning for how it will be implemented and how it will be consistent with the organization's competitive strategy.	• What processes should organizations adopt when designing and implementing responsible AI governance? • How can responsible AI practices incorporate elements of competitive strategies? • How are tensions/contradictions between being responsible with AI and attaining a competitive edge over rivals resolved?
	Mitigating negative or unintended effects	Negative or unintended consequences can occur during different stages of AI design, deployment, and organization use. Establishing robust processes that account for such effects and preempt them is important.	• What processes are necessary to ensure algorithmic transparency and explainability? • How should organizations design processes to ensure data quality and minimize bias in AI models? • What processes should companies implement to allow for auditability and ensure compliance with regulations and relevant guidelines? • How do ethical AI guidelines translate into specific processes that can be implemented at different organizational levels?
Relational practices	Responsible AI literacy	Implementing responsible AI practices depends on the competencies of individuals at different levels within the organization.	• How should organizations assess and develop the level of responsible AI literacy? • What are the dynamics by which collective and individual responsible AI competencies are developed? • In which key areas do different employees need to be educated regarding responsible AI practices?
	Stakeholder Involvement	Responsible AI governance requires the involvement of various stakeholders at different stages of the design and implementation of these technologies.	• How should organizations establish mechanisms to ensure that all relevant stakeholders collaborate effectively for AI development? • What tensions and competing interests develop between stakeholders, and how are they resolved? • How should organizations approach sourcing, contracting, and joint ventures with external parties

(*continued on next page*)

Table 5 (*continued*)

Theme	Issue	Description	Research questions
			regarding AI, and how can they ensure responsible governance in such arrangements? • How can organizations allocate AI responsibilities to different stakeholders, so that they can understand new risks, implications, and opportunities?
Effects	Business value	Responsible AI governance is commonly regarded as a requirement but only treated as such. However, implementing such practices may have benefits and organizational value.	• How can responsible AI governance effects be captured, and through what mechanisms do they create value for organizations? • How should organizations communicate their responsible AI governance practices to different stakeholders, and how does that affect corporate reputation and perceptions? • What spillover effects do adopting responsible AI governance practices have on internal operations and employee-level impacts? • What novel business models can organizations develop that build on the digital responsibility of AI products and services?
	Social assessment	Society continuously assesses new AI applications, which creates a dynamic relationship between what organizations deploy and how they should revise/adapt based on received signals.	• How should organizations gauge the reception of their AI services for the end users and society in general? • What elements of communication and interaction are important to improve social approval, trust, and use of AI-based services? • How does social evaluation of AI-based applications affect regulations, policymaking, and accepted practices that organizations should adopt?

Procedural practices

Procedural practices encompass processes crucial for data and model management, pipeline evaluation, and human-AI interaction (Papagiannidis et al., 2022; Tallon et al., 2013). In data management, these practices ensure the organization, security, and accessibility of data through classification and encryption protocols in a manner that minimizes bias and checks for errors. Procedural practices streamline workflows, optimize efficiency, and facilitate the continuous improvement of models and data throughout their lifecycles (Tallon et al., 2013). Furthermore, procedural practices in human-AI interaction focus on designing ethical and transparent AI systems, emphasizing user trust and accountability. These practices serve as essential frameworks across diverse domains – ensuring operational efficiency, reliability, and ethical integrity by aligning AI activities with applicable laws, regulations, and industry guidelines (Raji et al., 2020; Ryan, 2020). By upholding responsible AI governance processes through robust compliance monitoring and enforcement, organizations can mitigate legal and ethical risks, build trust with stakeholders, and safeguard themselves against the potential harm arising from AI-related activities (Min et al., 2023). Furthermore, organizations can develop robust incident responses and crisis management procedures to address AI-related issues promptly and effectively (Lee et al., 2022). These procedures should encompass protocols for detecting and assessing incidents, defining escalation pathways, coordinating response efforts, and communicating with stakeholders (Ahmad et al., 2021). Organizations can mitigate risks, minimize potential damage, and maintain trust and confidence in their AI governance frameworks by implementing comprehensive incident response and crisis management procedures.

Despite a broad set of practices and processes for implementing responsible AI governance, knowledge concerning strategic planning is limited. Organizations often struggle to harmonize their competitive strategies with the aims and ambitions of responsibly managing AI applications. A common assumption is that being responsible for AI deployment is incompatible with gaining a competitive edge when leveraging such technology. Thus, it is crucial to understand the processes at the strategic level that facilitate the responsible deployment of AI technologies to be utilized as a competitive strategy, and *vice versa*. At the same time, many of the defined processes around responsible AI governance concern the operational levels of decision-making, with limited well-defined practices concerning strategic and tactical levels. Exploring how responsible AI principles can be translated into concrete processes at these levels and how they can be utilized to proactively mitigate negative and unintended consequences is critical for ensuring that organizations remain competitive by leveraging AI and doing so responsibly and ethically.

Relational practices

Relational practices within the context of responsible AI governance involve establishing links within and beyond the organization, training and educating stakeholders, and ensuring that appropriate links are formed to facilitate the responsible development and use of AI. To facilitate cross-functional collaboration, organizations establish systematic procedures to encourage and enhance communication and teamwork among various departments (Rakova et al., 2021). These procedures outline clear communication channels, establish regular collaboration meetings, and promote knowledge-sharing and collaboration platforms to facilitate the exchange of ideas and expertise. By fostering a culture of collaboration, organizations can ensure that AI initiatives are aligned with overarching organizational values and norms, leveraging diverse perspectives and skill sets to drive innovation and achieve strategic goals (Aldoseri et al., 2023). Relational practices define the modes of stakeholder participation in the development process. For example,

participation can be inclusive throughout the entire data collection process – from strategic planning through identifying data needs, choosing and testing a suitable collection approach, collecting the data, and finally storing, disseminating, analyzing, and interpreting it (Edwards & Veale, 2018). Such approaches are useful for avoiding conflicts of interest, unethical uses of data (where data collected for one purpose are employed for an entirely different purpose), unauthorized data ownership or access (determining the true owner of the data), and misuse of data (sharing information with third parties that the user is not aware of). To this end, organizations must identify all relevant stakeholder groups and envision ways to complement the design, development, and use of AI.

Nevertheless, an inclusive and expansive view of stakeholders in AI development can potentially slow assimilation. To prevent risk management from stifling innovation, organizations must carefully consider their internal and external relational practices. Interest in responsible AI literacy has been growing as it becomes evident that much of what organizations do regarding responsible AI practices hinges on employees' knowledge. Thus, it becomes increasingly important to understand what skills and competencies employees at different roles and levels need to be equipped with, as well as how individual and collectively responsible AI literacy is developed. Doing so requires organizations to empower their employees to raise questions or concerns about AI systems to effectively control technology without limiting innovation. Additionally, research should focus on facilitating a greater sense of belonging by creating inclusion and diversity strategies. Hence, research on responsible AI practices should identify opportunities to act in areas where AI meets human ingenuity. Doing so relies as much on formal control mechanisms from the organization as on informal and self-organizing mechanisms in stakeholder groups. Thus, relational practices in responsible AI governance should also be examined through the informal channels through which they are diffused and executed. Additionally, ensuring that external stakeholders are involved through appropriate formal and informal approaches and that there are effective control mechanisms to orchestrate and govern such relationships is critical for project outcomes.

Effects of responsible AI governance

The diffusion of AI applications and services has opened new opportunities for organizations, promising a competitive edge over their rivals. However, the lack of comprehensive responsible AI governance practices in organizations and limited knowledge about implementing such principles in practice can be problematic (Mikalef et al., 2022). The rationale for adopting responsible AI governance is tied to avoiding possible negative repercussions and formulating competitive strategies that build on key pillars. Similar to how corporate social responsibility has been linked to positive returns for firms that practice such approaches, so can responsible AI governance be tied to corresponding outcomes. A recent study by Minkkinen et al. (2024) indicated that responsible AI governance practices are increasingly important in the context of environmental, social, and governance (ESG) assessments, which are key indicators of the corporate reputation of external actors and investors.

Recent commentaries have suggested that adopting responsible AI governance can generate value for companies by legitimizing their presence in a broader societal context and improving morale, productivity, and employee loyalty (de Laat, 2021; Martin & Waldman, 2023). At the organizational level, deploying and using responsible AI governance practices can enhance organizations' sense of external legitimacy in broader ecosystems. Organizations that apply responsible AI governance across their various applications are often perceived as more reliable and trustworthy business partners, enhancing their external image among customers. Knowing the secure, transparent, and reliable use of data and AI applications can result in positive customer trust and an overall corporate image return. With the prevalence of digital responsibility, cultivating a positive image and fostering appropriate practices to support it can increase the likelihood of other stakeholders engaging in business partnerships. This positive image of responsibility for using AI agents can lead to higher levels of profitability, greater customer retention and satisfaction, and greater ease of entering into strategic alliances and partnerships that benefit the focal organization (Enholm et al., 2021).

Conversely, a lack of responsible AI practices has been associated with low job performance in different occupations and industries (Rana et al., 2021) and AI-induced anxiety (Johnson & Verdicchio, 2017). When employees perceive that AI applications have been introduced in a manner that diminishes their autonomy and threatens their occupations, it can lead to lower levels of career satisfaction and organizational commitment (Johnson & Verdicchio, 2017). Nevertheless, effective, responsible AI governance can limit the feelings of threat that individuals in organizations perceive. Additionally, it can enhance synergy and seamless cooperation, eventually resulting in higher job satisfaction, productivity, and overall well-being. Nevertheless, we still lack empirical evidence on the effects of adopting responsible AI governance on internal operations and how it affects organizations' competitive positioning in broader ecosystems. A generally held assumption is that responsible AI is implemented primarily to adhere to societal expectations and norms; however, the role of the individual within an organization has been largely overlooked. Recent advancements in Generative AI technologies have shown that deploying AI applications in organizational operations can significantly impact the work life and well-being of employees (Caporusso, 2023). Ethical considerations for deploying AI in the workplace are manifold and raise questions about human identity, autonomy, and anxiety. In this regard, responsible AI governance must incorporate practices that prevent such occurrences and emphasize the well-being of employees.

As AI-based services and products are increasingly diffused in everyday life, their adoption by end users will affect the social perspective and shape the current norm around AI use (Akter et al., 2021). During this process, different AI service offerings are gauged by society, not only in terms of functionality and use but also in terms of their fit to societal norms and expectations. Organizations will, therefore, need to find suitable approaches to communicate their AI offerings and develop mechanisms to develop trust to facilitate

adoption by end users. Research to date has provided limited insight into how organizations can foster a sense of trust and effective communication with the end users of their AI services. Examples, such as Meta's attempts to leverage user data to train AI models, have shown that insufficient communication practices and opaque approaches to data handling can lead to backlash from end users.[4] Simultaneously, the relationship between AI offerings and societal perceptions and expectations is dynamic, with what is considered responsible and normatively aligned continuously evolving. Thus, not only do social norms and expectations shape what AI applications are deployed, but the way we interact with such technologies' daily shapes are also considered normative and ethical. While users are aware of data privacy concerns related to AI applications, in numerous circumstances, they are willing to overlook such problems when the perceived value they realize through such applications is sufficiently high. Consequently, deploying more sophisticated and complex AI applications that often go unnoticed in our interactions can lead us to re-evaluate what we consider to be the ethical and responsible use of AI. Thus, future research must examine the relationship between AI offerings from organizations and societal norms and expectations, as well as how developing tensions influence the evolution of responsible AI governance.

Conclusions

In this study, we examine the notion of responsible AI and synthesize current knowledge from academic research. Additionally, we illustrate the significance of responsible principles and their applications in organizations and society. Based on this synthesis, we developed an extended conceptual framework (Fig. 2) for responsible AI governance that illustrates how organizations can apply responsible AI principles in their AI applications. We discuss the key dimensions of responsible AI governance and identify the key antecedents and outcomes. In conclusion, we propose a set of actionable research themes that can help expand our understanding of how responsible AI governance can be deployed and its impact on organizations and society.

Our study had certain limitations. First, while we applied a thorough search strategy, the resulting papers and the corresponding AI principles identified were predominantly derived from institutions and organizations in Europe and the USA; the resulting ethical and cultural norms that are present in such frameworks may partially reflect the formulation of responsible AI governance. Nevertheless, our framework and the proposed approach to implementing responsible AI governance are principally agnostic and thus can be applied in different contexts. Second, although we have synthesized a diverse and fragmented body of literature, our arguments and assumptions must be tested empirically, and the relationships we propose require further exploration. Finally, although we aimed to provide a set of themes for future research, the proposed research questions were neither exhaustive nor complete. Many other open questions and essential themes will likely emerge as AI diffusion accelerates. Similarly, the emergence of regulations and directives, such as those of the AI Act, imposes ordinances on values that must be prioritized and conditions how responsible AI governance is perceived and thought of. How such regulations affect organizations' responsibility for AI governance implementation and how society and end users perceive their effect remain unclear.

CRediT authorship contribution statement

Emmanouil Papagiannidis: Writing – review & editing, Writing – original draft, Project administration, Methodology, Conceptualization. **Patrick Mikalef:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Conceptualization. **Kieran Conboy:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors have no conflict of interest, and the research did not receive any funding.

References

Abraham, R., Schneider, J., vom Brocke, J., 2019. Data governance: A conceptual framework, structured review, and research agenda. Int. J. Inf. Manag. 49, 424–438. https://doi.org/10.1016/j.ijinfomgt.2019.07.008.

Adadi, A., Berrada, M., 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052.

Ahmad, A., Maynard, S.B., Desouza, K.C., Kotsias, J., Whitty, M.T., Baskerville, R.L., 2021. How can organizations develop situation awareness for incident response: A case study of management practice. Comput. Sec. 101, 102122. https://doi.org/10.1016/j.cose.2020.102122.

Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y.K., D'Ambra, J., Shen, K.N., 2021. Algorithmic bias in data-driven innovation in the age of AI. Int. J. Inf. Manag. 60, 102387. https://doi.org/10.1016/j.ijinfomgt.2021.102387.

Aldoseri, A., Al-Khalifa, K., Hamouda, A., 2023. A road map for integrating automation with process optimization for AI-powered digital transformation. Preprints. 2023, 2023101055. https://doi.org/10.20944/preprints202310.1055.v1.

[4] https://www.medianama.com/2024/06/223-norwegian-consumer-council-files-complaint-against-*meta*-for-training-ai-models-on-user-data.

Arksey, H., O'Malley, L., 2005. Scoping studies: Towards a methodological framework. Int. J. Soc. Res. Methodol. 8, 19–32. https://doi.org/10.1080/1364557032000119616.
Ashok, M., Madan, R., Joha, A., Sivarajah, U., 2022. Ethical framework for Artificial Intelligence and Digital technologies. Int. J. Inf. Manag. 62, 102433. https://doi.org/10.1016/j.ijinfomgt.2021.102433.
Ayling, J., Chapman, A., 2021. Putting AI ethics to work: Are the tools fit for purpose? AI Ethics 2 (3), 405–429.
Boell, S.K., Cecez-Kecmanovic, D., 2015. On being'systematic'in literature reviews. Formulating Res. Methods Inf. Syst. 30, 161–173. https://doi.org/10.1057/9781137509888_3.
Brackett, M., Earley, P.S., 2017. DAMA-DMBOK: data management body of knowledge. Technics Publications LLC.
Brendel, A.B., Mirbabaie, M., Lembcke, T.-B., Hofeditz, L., 2021. Ethical management of artificial intelligence. Sustainability. 13. https://doi.org/10.3390/su13041974.
Brighton, H., Gigerenzer, G., 2015. The bias bias. J. Bus. Res. 68, 1772–1784. https://doi.org/10.1016/j.jbusres.2015.01.061.
Butcher, J., Beridze, I., 2019. What is the state of artificial intelligence governance globally? RUSI J. 164, 88–96. https://doi.org/10.1080/03071847.2019.1694260.
Caner, S., Bhatti, F., 2020. A conceptual framework on defining businesses strategy for artificial intelligence. Contemp. Manag. Res. 16, 175–206. https://doi.org/10.7903/cmr.19970.
Caporusso, N., 2023. Generative artificial intelligence and the emergence of creative displacement anxiety. Res. Directs Psychol. Behav. 3. https://doi.org/10.53520/rdpb2023.10795.
Chang, H., Nguyen, T.D., Murakonda, S.K., Kazemi, E., Shokri, R., 2022. On adversarial bias and the robustness of fair machine learning International Conference on Learning Representations. https://arxiv.org/pdf/2006.08669.pdf.
Chatila, R., Dignum, V., Fisher, M., Giannotti, F., Morik, K., Russell, S., Yeung, K., 2021. Reflections on artificial intelligence for humanity. Springer, Cham, pp. 13–39, 10.1007/978-3-030-69128-8.
Clarke, R., 2019. Principles and business processes for responsible AI. Comput. Law Sec. Rev. 35, 410–422. https://doi.org/10.1016/j.clsr.2019.04.007.
Conboy, K., Dennehy, D., O'Connor, M., 2020. 'Big time': An examination of temporal complexity and business value in analytics. Inf. Manag. 57, 103077. https://doi.org/10.1016/j.im.2018.05.010.
Council of Europe, 2018. European ethical charter on the use of artificial intelligence in judicial systems and their environment. https://rm.coe.int/ethical-charter-en-for-publication-4-december2018/16808f699c.
Dastin, J., 2022. Amazon scraps secret AI recruiting tool that showed bias against women. In: Ethics of Data and Analytics, 1 ed. Auerbach Publications, pp. 296–299. 10.1201/9781003278290.
de Almeida, P.G.R., dos Santos, C.D., Farias, J.S., 2021. Artificial intelligence regulation: a framework for governance. Ethics Inf. Technol. 23, 505–525. https://doi.org/10.1007/s10676-021-09593-z.
de Laat, P.B., 2021. Companies committed to responsible AI: From principles towards implementation and regulation? Philos. Technol. 34, 1135–1193. https://doi.org/10.1007/s13347-021-00474-3.
Di Vaio, A., Palladino, R., Hassan, R., Escobar, O., 2020. Artificial intelligence and business models in the sustainable development goals perspective: a systematic literature review. J. Bus. Res. 121, 283–314. https://doi.org/10.1016/j.jbusres.2020.08.019.
Dignum, V., 2017. Responsible autonomy. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia, pp. 4698–4704, 10.24963/ijcai.2017/655.
Dignum, V., 2019. Responsible artificial intelligence: how to develop and use ai in a responsible way. Springer.
Dignum, V., 2019a. Responsible artificial intelligence: How to develop and use AI in a responsible way, 1 ed. Springer Nat. 10.1007/978-3-030-30371-6.
Donati, F., Dente, S.M.R., Li, C., Vilaysouk, X., Froemelt, A., Nishant, R., Liu, G., Tukker, A., Hashimoto, S., 2022. The future of artificial intelligence in the context of industrial ecology. J. Ind. Ecol. 26, 1175–1181. https://doi.org/10.1111/jiec.13313.
Edwards, L., Veale, M., 2018. Enslaving the algorithm: From a "right to an explanation" to a "right to better decisions"? IEEE Secur Privacy. 16, 46–54. https://doi.org/10.1109/MSP.2018.2701152.
Enholm, I.M., Papagiannidis, E., Mikalef, P., Krogstie, J., 2022. Artificial intelligence and business value: a literature review. Inf. Syst. Front. 24, 1709–1734. https://doi.org/10.1007/s10796-021-10186-w.
European Commission, 2019. Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.
Felzmann, H., Fosch-Villaronga, E., Lutz, C., Tamò-Larrieux, A., 2020. Towards transparency by design for artificial intelligence. Sci. Eng. Ethics. 26, 3333–3361. https://doi.org/10.1007/s11948-020-00276-4.
Feuerriegel, S., Dolata, M., Schwabe, G., 2020. Fair AI: Challenges and opportunities. Bus. Inf. Syst. Eng. 62, 379–384. https://doi.org/10.1007/s12599-020-00650-3.
Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., Srikumar, M., 2020. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. SSRN Journal. Publication (2020–1). https://doi.org/10.2139/ssrn.3518482.
Flavián, C., Casaló, L.V., 2021. Artificial intelligence in services: current trends, benefits and challenges. Serv. Ind. J. 41, 853–859. https://doi.org/10.1080/02642069.2021.1989177.
Floridi, L., Cowls, J., 2021. A unified framework of five principles for AI in society. In: Ethics, Governance, and Policies in Artificial Intelligence. Springer, pp. 5–17, 10.1007/978-3-030-81907-1_2.
Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., Vayena, E., 2021. An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations, in: Ethics, Governance, and Policies in Artificial Intelligence. Springer, pp. 19–39. Doi: 10.1007/978-3-030-81907-1_3.
Freiman, O., 2023. Making sense of the conceptual nonsense "trustworthy AI". AI Ethics. 3, 1351–1360. https://doi.org/10.1007/s43681-022-00241-w.
Fuchs, D.J., 2018. The dangers of humanlike bias in machine-learning algorithms. Mo. S&t's Peer to Peer. 2, 1. https://scholarsmine.mst.edu/peer2peer/vol2/iss1/1.
Future of Life Institute, 2017. Beneficial AI 2017, (Retrieved Nov 24, 2022). https://futureoflife.org/event/bai-2017. (Accessed Nov 16, 2022).
Gasser, U., Almeida, V.A.F., 2017. A layered model for AI governance. IEEE Internet Comput. 21, 58–62. https://doi.org/10.1109/MIC.2017.4180835.
Ghallab, M., 2019. Responsible AI: Requirements and challenges. AI Perspect. 1, 1–7.
Gillath, O., Ai, T., Branicky, M.S., Keshmiri, S., Davison, R.B., Spaulding, R., 2021. Attachment and trust in artificial intelligence. Comput. Hum. Behav. 115, 106607. https://doi.org/10.1016/j.chb.2020.106607.
Google, 2019. Perspectives on issues in AI governance. https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf.
Grimshaw, J.M., Shirran, L., Thomas, R., Mowatt, G., Fraser, C., Bero, L., Grilli, R., Harvey, E., Oxman, A., O'Brien, M.A., 2001. Changing provider behavior: An overview of systematic reviews of interventions. Med. Care. 39 (Supplement 2), II2–I45. https://doi.org/10.1097/00005650-200108002-00002.
Hagendorff, T., 2020. The ethics of AI ethics: An evaluation of guidelines. Minds Mach. 30, 99–120. https://doi.org/10.1007/s11023-020-09517-8.
Hamon, R., Junklewitz, H., Sanchez, I., 2020. Robustness and explainability of Artificial Intelligence: From technical to policy solutions (Publications Office of the European Union. Issue. https://data.europa.eu/doi/10.2760/57493.
Helbing, D., 2019. Machine intelligence: blessing or curse? It depends on us! towards digital enlightenment. Springer, pp. 25–39.
Hernández-Orallo, J., Martínez-Plumed, F., Avin, S., Whittlestone, J., 2020. AI Paradigms and AI Safety: Mapping Artefacts and Techniques to Safety Issues.
Hilb, M., 2020. Toward artificial governance? The role of artificial intelligence in shaping the future of corporate governance. J. Manag. Gov. 24, 851–870. https://doi.org/10.1007/s10997-020-09519-9.
Holmström, J., Hällgren, M., 2022. AI management beyond the hype: exploring the co-constitution of AI and organizational context. AI & Soc. 37, 1575–1585. https://doi.org/10.1007/s00146-021-01249-2.
IBM, 2019. IBM everyday ethics for AI. https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf.
Jakesch, M., Buçinca, Z., Amershi, S., Olteanu, A., 2022. How different groups prioritize ethical values for responsible AI, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 310–323. Doi: 10.1145/3531146.3533097.

Janssen, M., Brous, P., Estevez, E., Barbosa, L.S., Janowski, T., 2020. Data governance: Organizing data for trustworthy Artificial Intelligence. Gov. Inf. q. 37, 101493. https://doi.org/10.1016/j.giq.2020.101493.
Jelinek, T., Wallach, W., Kerimi, D., 2021. Policy brief: The creation of a G20 coordinating committee for the governance of artificial intelligence. AI Ethics. 1, 141–150. https://doi.org/10.1007/s43681-020-00019-y.
Ji-fan Ren, S., Fosso Wamba, S., Akter, S., Dubey, R., Childe, S.J., 2017. Modelling quality dynamics, business value and firm performance in a big data analytics environment. Int. J. Prod. Res. 55, 5011–5026. https://doi.org/10.1080/00207543.2016.1154209.
Jobin, A., Ienca, M., Vayena, E., 2019. The global landscape of AI ethics guidelines. Nat. Mach. Intell. 1, 389–399. https://doi.org/10.1038/s42256-019-0088-2.
Jöhnk, J., Weißert, M., Wyrtki, K., 2021. Ready or not, AI comes—An interview study of organizational AI readiness factors. Bus. Inf. Syst. Eng. 63, 5–20. https://doi.org/10.1007/s12599-020-00676-7.
Johnson, D.G., Verdicchio, M., 2017. AI anxiety. J. Assoc. Inf. Sci. Technol. 68, 2267–2270. https://doi.org/10.1002/asi.23867.
Kitchenham, B., 2004. Procedures for performing systematic reviews. https://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf.
Kodiyan, A.A., 2019. An overview of ethical issues in using AI systems in hiring with a case study of Amazon's AI based hiring tool. Researchgate Preprint 1–19.
Korinek, A., 2020. Integrating ethical values and economic value to steer progress in artificial intelligence (Vol. w26130). The Oxford Handbook of Ethics of AI. Doi: 10.1093/oxfordhb/9780190067397.013.30.
Kuziemski, M., Misuraca, G., 2020. AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. Telecommun. Policy. 44, 101976. https://doi.org/10.1016/j.telpol.2020.101976.
Larsson, S., Anneroth, M., Felländer, A., Felländer-Tsai, L., Heintz, F., Ångström, R.C., 2019. Sustainable AI: An inventory of the state of knowledge of ethical, social, and legal challenges related to artificial intelligence. AI Sustainability Center. 68, 101926. https://doi.org/10.1016/j.techsoc.2022.101926.
Leavy, S., 2018. Gender bias in artificial intelligence: the need for diversity and gender theory in machine learning 1st International Workshop on Gender Equality in Software Engineering (GE). IEEE Publications/ACM, Gothenburg, Sweden, 10.1145/3195570.3195580.
Lee, C.-C., Comes, T., Finn, M., Mostafavi, A., 2022. Road map towards responsible AI in crisis resilience management. Cornell University, arXiv Preprint ArXiv: 2207.09648.
Lubberink, R., Blok, V., van Ophem, J., Omta, O., 2019. Responsible innovation by social entrepreneurs: an exploratory study of values integration in innovations. J. Respons. Innov. 6, 179–210. https://doi.org/10.1080/23299460.2019.1572374.
Mannes, A., 2020. Governance, risk, and artificial intelligence. AI Mag. 41, 61–69. https://doi.org/10.1609/aimag.v41i1.5200.
Mäntymäki, M., Minkkinen, M., Birkstedt, T., Viljanen, M., 2022. Defining organizational AI governance. AI Ethics. 2, 603–609. https://doi.org/10.1007/s43681-022-00143-x.
Margetts, H., 2022. Rethinking AI for good governance. Dædalus. 151, 360–371. https://doi.org/10.1162/daed_a_01922.
Martin, K., Waldman, A., 2023. Are algorithmic decisions legitimate? The effect of process and outcomes on perceptions of legitimacy of AI decisions. J. Bus. Ethics. 183, 653–670. https://doi.org/10.1007/s10551-021-05032-7.
Matthews, J., 2020. Patterns and antipatterns, principles and pitfalls: Accountability and transparency in artificial intelligence. AI Mag. 41, 82–89. https://doi.org/10.1609/aimag.v41i1.5204.
Meske, C., Bunde, E., Schneider, J., Gersch, M., 2022. Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities. Inf. Syst. Manag. 39, 53–63. https://doi.org/10.1080/10580530.2020.1849465.
Mezgár, I., 2021. From ethics to standards; an overview of AI ethics in CPPS. IFAC PapersOnLine. 54, 723–728. https://doi.org/10.1016/j.ifacol.2021.08.084.
Mikalef, P., Conboy, K., Lundström, J.E., Popovič, A., 2022. Thinking responsibly about responsible AI and 'the dark side' of AI. Eur. J. Inf. Syst. 31, 257–268. https://doi.org/10.1080/0960085X.2022.2026621.
Mikalef, P., Fjørtoft, S.O., Torvatn, H.Y., 2019. Developing an artificial intelligence capability: A theoretical framework for business value. International Conference on Business Information Systems.
Mikalef, P., Gupta, M., 2021. Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. Inf. Manag. 58, 103434. https://doi.org/10.1016/j.im.2021.103434.
Mills, S., Baltassis, E., Santinelli, M., Carlisi, C., Duranton, S., Gallego, A., 2020. Six steps to bridge the responsible AI gap, in:.
Min, J., Kim, J., Yang, K., 2023. CSR attributions and the moderating effect of perceived CSR fit on consumer trust, identification, and loyalty. J. Retailing Con. Serv. 72, 103274. https://doi.org/10.1016/j.jretconser.2023.103274.
Minkkinen, M., Niukkanen, A., Mäntymäki, M., 2024. What about investors? ESG analyses as tools for ethics-based AI auditing. AI Soc. 39, 329–343. https://doi.org/10.1007/s00146-022-01415-0.
Mökander, J., Floridi, L., 2021. Ethics-based auditing to develop trustworthy AI. Minds Mach. 31, 323–327. https://doi.org/10.1007/s11023-021-09557-8.
Mora-Cantallops, M., Sánchez-Alonso, S., García-Barriocanal, E., Sicilia, M.-A., 2021. Traceability for trustworthy ai: A review of models and tools. Big Data Cogn. Comput. 5, 20. https://doi.org/10.3390/bdcc5020020.
Nath, V., Levinson, S.E., 2014. Autonomous military robotics. Springer. https://doi.org/10.1007/978-3-319-05606-7_5.
Nishant, R., 2023. The formal rationality of artificial intelligence-based algorithms and the problem of bias. D, & Ravishankar, M. J. Inf. Technol, Schneckenberg, 02683962231176842.
Nzobonimpa, S., Savard, J.F., 2023. Ready but irresponsible? Analysis of the government artificial intelligence readiness index. Policy Internet. 15, 397–414. https://doi.org/10.1002/poi3.351.
Okoli, C., 2015. A guide to conducting a standalone systematic literature review. Commun. Assoc. Inf. Syst. 37, 43. https://doi.org/10.17705/1CAIS.03743.
Pagallo, U., Aurucci, P., Casanovas, P., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Schafer, B., Valcke, P., 2019. AI4People-On good AI governance: 14 priority actions, a SMART model of governance, and a regulatory toolbox. https://ssrn.com/abstract=3486508.
Pan, S.L., Nishant, R., 2023. Artificial intelligence for digital sustainability: An insight into domain-specific research and future directions. Int. J. Inf. Manag. 72, 102668. https://doi.org/10.1016/j.ijinfomgt.2023.102668.
Papagiannidis, E., Enholm, I.M., Dremel, C., Mikalef, P., Krogstie, J., 2023. Toward AI governance: Identifying best practices and potential barriers and outcomes. Inf. Syst. Front. 25, 123–141. https://doi.org/10.1007/s10796-022-10251-y.
Paré, G., Trudel, M.-C., Jaana, M., Kitsiou, S., 2015. Synthesizing information systems knowledge: A typology of literature reviews. Inf. Manag. 52, 183–199. https://doi.org/10.1016/j.im.2014.08.008.
Porritt, K., Gomersall, J., Lockwood, C., 2014. JBI's systematic reviews: Study selection and critical appraisal. Am. J. Nurs. 114, 47–52. https://doi.org/10.1097/01.NAJ.0000450430.97383.64.
Radu, R., 2021. Steering the governance of artificial intelligence: National strategies in perspective. Policy Soc. 40, 178–193. https://doi.org/10.1080/14494035.2021.1929728.
Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., Barnes, P., 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona Spain, pp. 33–44, 10.1145/3351095.3372873.
Rakova, B., Yang, J., Cramer, H., Chowdhury, R., 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. Proc. ACM Hum. Comput. Interact. 5 (CSCW1), 1–23.
Rana, N.P., Chatterjee, S., Dwivedi, Y.K., Akter, S., 2022. Understanding dark side of artificial intelligence (AI) integrated business analytics: Assessing firm's operational inefficiency and competitiveness. Eur. J. Inf. Syst. 31, 364–387. https://doi.org/10.1080/0960085X.2021.1955628.
Ransbotham, S., Gerbert, P., Reeves, M., Kiron, D., Spira, M., 2018. Artificial intelligence in business gets real. MIT Sloan Manag. Rev. 60280.
Reddy, S., Allan, S., Coghlan, S., Cooper, P., 2020. A governance model for the application of AI in health care. J. Am. Med. Inform. Assoc. 27, 491–497. https://doi.org/10.1093/jamia/ocz192.
Rowe, F., 2014. What literature review is not: Diversity, boundaries and recommendations. Eur. J. Inf. Syst. 23, 241–255. https://doi.org/10.1057/ejis.2014.7.
Ryan, M., 2020. In AI we trust: Ethics, artificial intelligence, and reliability. Sci. Eng. Ethics. 26, 2749–2767. https://doi.org/10.1007/s11948-020-00228-y.

Salunke, S., Weerawardena, J., McColl-Kennedy, J.R., 2011. Towards a model of dynamic capabilities in innovation-based competitive strategy: Insights from project-oriented service firms. Ind. Mark. Manag. 40, 1251–1263. https://doi.org/10.1016/j.indmarman.2011.10.009.
Sandelowski, M., Barroso, J., 2006. Handbook for synthesizing qualitative research. Springer Publishing Company.
Schiff, D., Rakova, B., Ayesh, A., Fanti, A., Lennon, M., 2021. Explaining the principles to practices gap in AI. IEEE Technol. Soc. Mag. 40, 81–94. https://doi.org/10.1109/MTS.2021.3056286.
Schlögl, S., Postulka, C., Bernsteiner, R., Ploder, C., 2019. Artificial intelligence tool penetration in business: Adoption, challenges and fears International Conference on Knowledge Management in Organizations. Zamora, Spain, pp. 259–270.
Schneider, J., Abraham, R., Meske, C., Vom Brocke, J., 2023. Artificial intelligence governance for businesses. Inf. Syst. Manag. 40, 229–249. https://doi.org/10.1080/10580530.2022.2085825.
Seppälä, A., Birkstedt, T., Mäntymäki, M., 2021. From ethical AI principles to governed AI. ICIS.
Shneiderman, B., 2020a. Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Trans. Interact. Intell. Syst. (tiis) 10, 1–31. https://doi.org/10.1145/3419764.
Shneiderman, B., 2020b. Human-centered artificial intelligence: Three fresh ideas. AIS Trans. Human-Computer Interact. 12, 109–124. https://doi.org/10.17705/1thci.00131.
Siala, H., Wang, Y., 2022. SHIFTing artificial intelligence to be responsible in healthcare: a systematic review. Soc. Sci. Med. 296, 114782. https://doi.org/10.1016/j.socscimed.2022.114782.
Singapore Government, 2020. Model artificial intelligence governance framework. Singapore Digital (SG:D), Issue, second ed. https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf.
Smuha, N.A., 2021. Beyond a human rights-based approach to AI governance: Promise, pitfalls, plea. Philos. Technol. 34, 91–104. https://doi.org/10.1007/s13347-020-00403-w.
Taeihagh, A., 2021. Governance of artificial intelligence. Policy Soc. Taylor & Francis. 40, 137–157. https://doi.org/10.1080/14494035.2021.1928377.
Tallon, P.P., Ramirez, R.V., Short, J.E., 2013. The information artifact in IT governance: Toward a theory of information governance. J. Manag. Inf. Syst. 30, 141–178. https://doi.org/10.2753/MIS0742-1222300306.
Telia, 2019. Guiding principles on trusted AI ethics. https://www.teliacompany.com/globalassets/telia-company/documents/about-telia-company/publicpolicy/2018/guiding-principles-on-trusted-ai-ethics.pdf.
Templier, M., Paré, G., 2015. A framework for guiding and evaluating literature reviews. Commun. Assoc. Inf. Syst. 37, 6. https://doi.org/10.17705/1CAIS.03706.
Theodorou, A., Dignum, V., 2020. Towards ethical and socio-legal governance in AI. Nat. Mach. Intell. 2, 10–12. https://doi.org/10.1038/s42256-019-0136-y.
Thiebes, S., Lins, S., Sunyaev, A., 2021. Trustworthy artificial intelligence. Electron. Markets. 31, 447–464. https://doi.org/10.1007/s12525-020-00441-4.
Tolmeijer, S., Christen, M., Kandul, S., Kneer, M., Bernstein, A., 2022. Capable but amoral? Comparing AI and human expert collaboration in ethical decision making. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, pp. 1–17.
Too, E.G., Weaver, P., 2014. The management of project management: A conceptual framework for project governance. Int. J. Proj. Manag. 32, 1382–1394. https://doi.org/10.1016/j.ijproman.2013.07.006.
Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C.G., Van Moorsel, A., 2020. The relationship between trust in AI and trustworthy machine learning technologies, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 272–283. Doi: 10.1145/3351095.3372834.
Trocin, C., Mikalef, P., Papamitsiou, Z., Conboy, K., 2021. Responsible AI for digital health: A synthesis and a research agenda. Inf. Syst. Front. 1–19.
Van Grembergen, W., De Haes, S., Guldentops, E., 2004. Structures, processes and relational mechanisms for IT governance, in: Strategies for Information Technology Governance. IGI Global, pp. 1–36.
Varona, D., Suárez, J.L., 2022. Discrimination, bias, fairness, and trustworthy AI. Appl. Sci. 12, 5826. https://doi.org/10.3390/app12125826.
Venkataramanan, V., Packman, A.I., Peters, D.R., Lopez, D., McCuskey, D.J., McDonald, R.I., Miller, W.M., Young, S.L., 2019. A systematic review of the human health and social well-being outcomes of green infrastructure for stormwater and flood management. J. Environ. Manag. 246, 868–880. https://doi.org/10.1016/j.jenvman.2019.05.028.
Vollmer, S., Mateen, B.A., Bohner, G., Király, F.J., Ghani, R., Jonsson, P., Cumbers, S., Jonas, A., McAllister, K.S.L., Myles, P., Granger, D., Birse, M., Branson, R., Moons, K.G.M., Collins, G.S., Ioannidis, J.P.A., Holmes, C., Hemingway, H., 2020. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ 368 (l6927). https://doi.org/10.1136/bmj.l6927.
Wang, Y., Chaudhuri, K., 2018. Data poisoning attacks against online learning. arXiv Preprint ArXiv:1808.08994.
Winfield, A.F.T., Jirotka, M., 2018. Ethical governance is essential to building trust in robotics and artificial intelligence systems. Philos. Trans. A Math. Phys. Eng. Sci. 376. https://doi.org/10.1098/rsta.2018.0085.
Wirtz, B.W., Weyerer, J.C., Sturm, B.J., 2020. The dark sides of artificial intelligence: An integrated AI governance framework for public administration. Int. J. Public Admin. 43, 818–829. https://doi.org/10.1080/01900692.2020.1749851.
Wu, W., Huang, T., Gong, K., 2020. Ethical principles and governance technology development of AI in China. Engineering 6, 302–309. https://doi.org/10.1016/j.eng.2019.12.015.
Yerlikaya, S., Erzurumlu, Y.Ö., 2021. Artificial intelligence in public sector: A framework to address opportunities and challenges. The Fourth Industrial Revolution: Implementation of Artificial Intelligence for Growing Business Success 201–216. https://doi.org/10.1007/978-3-030-62796-6_11.
Yeung, K., Howes, A., Pogrebna, G., 2020. AI governance by human rights–centered design, deliberation, and oversight. The Oxford Handbook of Ethics of AI 77–106.
Zhang, D., Pee, L.G., Cui, L., 2021. Artificial intelligence in e-commerce fulfillment: A case study of resource orchestration at AliBaba's Smart Warehouse. Int. J. Inf. Manag. 57, 102304. https://doi.org/10.1016/j.ijinfomgt.2020.102304.
Zicari, R.V., Brodersen, J., Brusseau, J., Düdder, B., Eichhorn, T., Ivanov, T., Kararigas, G., Kringen, P., McCullough, M., Möslein, F., Mushtaq, N., Roig, G., Sturtz, N., Tolle, K., Tithi, J.J., van Halem, I., Westerlund, M., 2021. Z-Inspection ®: A process to assess trustworthy AI. IEEE Trans. Technol. Soc. 2, 83–97. https://doi.org/10.1109/TTS.2021.3066209.
Zou, J., Schiebinger, L., 2018. AI can be sexist and racist—It's time to make it fair. Nature 559, 324–326. https://doi.org/10.1038/d41586-018-05707-8.