5-11-2023

Accountability in Managing Artificial Intelligence: State of the Art and a way forward for Information Systems Research

Alexander Moltubakk Kempton
University of Oslo, alexansk@ifi.uio.no

Elena Parmiggiani
Norwegian University of Science and Technology, parmiggi@ntnu.no

Polyxeni Vassilakopoulou
University of Agder, polyxenv@uia.no

Recommended Citation

Kempton, Alexander Moltubakk; Parmiggiani, Elena; and Vassilakopoulou, Polyxeni, "Accountability in Managing Artificial Intelligence: State of the Art and a way forward for Information Systems Research" (2023). *ECIS 2023 Research Papers*. 361.
https://aisel.aisnet.org/ecis2023_rp/361

ACCOUNTABILITY IN MANAGING ARTIFICIAL INTELLIGENCE: STATE OF THE ART AND A WAY FORWARD FOR INFORMATION SYSTEMS RESEARCH

Research Paper

Alexander Moltubakk Kempton, Department of Informatics and HISP Centre, University of Oslo, Norway, alexansk@ifi.uio.no

Elena Parmiggiani, Department of Computer Science, Norwegian University of Science and Technology (NTNU), and Sintef Nord AS, Tromsø, Norway, parmiggi@ntnu.no

Polyxeni Vassilakopoulou, Information Systems Department, University of Agder, Norway, polyxenv@uia.no

Abstract

Establishing accountability for Artificial Intelligence (AI) systems is challenging due to the distribution of responsibilities among multiple actors involved in their development, deployment, and use. Nonetheless, AI accountability is crucial. As AI can affect all aspects of private and professional life, the actors involved in AI lifecycles need to take responsibility for their decisions and actions, be ready to respond to interrogations by those affected by AI and held liable when AI works in unacceptable ways. Despite the significance of AI accountability, the Information Systems research community has not engaged much with the topic and lacks a systematic understanding of existing approaches to it. This paper present the results of a comprehensive conceptual literature review that synthetizes current knowledge on AI accountability. The paper contributes to the IS literature by providing (i) conceptual clarification mapping different accountability conceptualizations; (ii) a comprehensive framework for AI accountability challenges and actionable responses at three different levels: system, process, data and; (iii) a framing of AI accountability as a a socio-technical and organizational problem that IS researchers are well-equipped to study highlighting the need to balance instrumental and humanistic outcomes.

Keywords: Artificial Intelligence, Accountability, AI management, Literature Review

1 Introduction

Artificial Intelligence (AI) technologies are increasingly used in critical application areas including processing medical images, controlling traffic, supporting complex legal decision-making processes, detecting tax fraud, automating credit underwriting and managing electricity grids. As AI keeps infusing contemporary life, the impact of AI on societies and individuals becomes significant. It is, therefore, critical to put in place arrangements to ensure accountability for AI-enabled systems and their outcomes (Australian Government, 2019). As Floridi and colleagues (2018) note, AI-enabled systems need to be handled as tools for enhancing human agency, without removing human responsibility.

Human responsibility for AI, however, is challenging to establish in practice. Multiple actors are involved with different roles in designing, developing and deploying AI, deciding when and how to use which algorithms, determining and manipulating data feeds for these algorithms, developing and validating models and overseeing algorithmic performance (including broader economic, societal, legal, and ethical impacts). These actors need to be able to justify their actions, respond to interrogations by those affected by AI deployment and to be liable when AI works in unacceptable ways. In other words, accountability arrangements need to be in place (Bovens, 2007, 2010; Bovens et al., 2014). Accountability can be viewed as a relationship between actors encompassing not only responsibility but also its enactment through visibility and liability (Boos & Grote, 2012). In the context of AI, accountability relates to the distribution and enactment of responsibility within extensive meshes of actors. Accountability, therefore, is one of the issues that makes managing AI unlike information technology management in the past (Berente et al., 2021).

The management of AI-enabled systems and the implications of different AI accountability arrangements is an exemplary sociotechnical concern that Information Systems (IS) researchers are well-positioned to study. IS research examines more than technologies or social phenomena, or even the two side by side; it investigates emergent sociotechnical phenomena (Lee, 2001). As a key interdisciplinary tradition, IS can contribute to the AI accountability discourse interfacing technical, organizational and ethical perspectives (Berente et al., 2021; Sarker et al., 2019). Nevertheless, there is a striking paucity of IS research on AI accountability while the volume of related research originating from other disciplines is growing. This makes it difficult for IS researchers to follow and relate their own work within the state of the art on the topic. This gap is odd, as accountability has previously been studied within IS but in different contexts including security management (Vance et al., 2015), information technology for development (Bernardi, 2017), Internet of Things applications (Boos & Grote, 2012) and new organizational forms based on social media and crowd-sourced content (Scott & Orlikowski, 2014). In this prior IS work, researchers have investigated how to establish or enhance accountability and the impact of emerging accountability regimes and accountability pressures at the level of individuals and organizations.

We performed a comprehensive literature review that can serve as a basis for orienting IS researchers interested on AI accountability. The review provides a synthesis of an extensive corpus of 131 papers systematizing the state of the art and identifying areas for future research (Ortiz de Guinea & Paré, 2017; Schryen et al., 2015). The literature synthesis was guided by the following questions:

- RQ1: How is AI accountability defined and conceptualized in the literature?
- RQ2: How can the literature on AI accountability inform the management of AI?

Our contribution is threefold. First, we contribute to conceptual clarification by mapping and analyzing different accountability conceptualizations. This is important because an issue with the extant body of research is that researchers tend to define accountability in different ways and therefore address different accountability issues, practices, and challenges. Second, we develop a comprehensive framework that maps and describes emerging AI accountability challenges and actionable AI management responses at the system, process and data levels. Finally, as a third contribution, we frame AI accountability as a socio-technical and organizational problem that IS researchers are well-equipped to study and identify areas for future research, thus providing a research agenda for the IS community. Specifically, we propose to approach AI accountability by jointly considering instrumental and humanistic outcomes (Sarker et al., 2019) and call for further research in the middle ground between AI business opportunities and ethical concerns.

The remainder of the paper is organized as follows. First, we elaborate on AI-specific accountability challenges. Then, we present our method for selecting and analyzing the articles for this review. We continue by offering a synthesis of our findings. We conclude the paper by discussing the implications of our findings and providing directions for further research.

2 The Challenge of Accountability in Managing AI

AI refers to technological artefacts performing the cognitive functions typically associated with humans, including perceiving, reasoning, and learning (Rai et al., 2019). The term does not denote a specific technology, it is an umbrella term for computational advancements that references human intelligence in addressing ever more complex decision-making problems (Berente et al., 2021). The recent rise of interest on AI is linked to successes in data-driven modelling and especially machine learning (Whittaker, 2021).

Until some years ago, AI was based on logic and knowledge-based approaches. In these traditional approaches, algorithm experts had full control over AI models. In contemporary machine learning, however, only the processes by which models learn from data can be controlled (Kane et al., 2021). In other words, instead of developing models, now we develop processes to train models using data. The models created in this way are as good as the data used for training (Benbya et al., 2021). AI applications often rely on "found data", i.e. data collected for some other purpose (Brown, 2021). Such datasets may not include all relevant aspects or representative instances of phenomena under study and may lead to inaccurate, or biased model outputs. Ensuring accountability for AI is especially challenging when multiple actors are involved in defining the processes for training models, defining what data will be used and sourcing the data.

Ensuring accountability for AI is also challenged by model inscrutability. AI applications based on machine learning tend to lack transparency as it is often impossible to explain how inputs lead to outputs. The inner workings of models are difficult to understand, and models may provide inferences that cannot be directly explained, i.e. they are "blackboxed" (Asatiani et al., 2021; Burrell, 2016; Rai, 2020). For instance, "deep learning" models cannot be represented in standard forms, such as closed equations, decision trees, or graphs (Strandburg, 2019). Inscrutability makes it hard to assess whether models will be reliable in unusual new situations. This is important because AI applications based on machine learning are adaptive to the social and socio-technical structures they are embedded in (Kempton, 2022). This inherent inscrutability has drawn the attention of AI experts leading to the development of explainable AI (XAI); a research domain aiming to produce approximate interpretations of inscrutable machine-learning models and for developing machine learning techniques that produce explainable models. However, XAI insights are not always easily communicable to non-experts. Ensuring accountability for inscrutable AI is a major challenge.

3 Method

We performed a conceptual literature review to provide a synthesis of prior research and identify areas for future research (Ortiz de Guinea & Paré, 2017; Schryen et al., 2015). We chose this approach because conceptual reviews are well-suited for focusing on a single concept and examining how the concept and its core attributes have been defined in the literature (Ortiz de Guinea & Paré, 2017). This type of literature review is suitable when the goal is to develop understanding with exhaustive coverage of the literature and high systematicity (idem). The approach we followed is based on the three-step process proposed by Kitchenham (2004). Specifically, the three-step process includes: a) planning the review, where a detailed protocol containing specific search terms and inclusion/exclusion criteria is developed, b) conducting the review, where the identification, selection, quality appraisal, examination and synthesis of prior published research is performed and c) reporting the review, where the write-up is prepared. We used these steps as our methodological framework.

Identification. To identify articles to be reviewed, we first performed a targeted literature search using specific terms. Then we extended the selection of articles by performing backward and forward searches starting from the corpus of literature identified via the search terms. We utilized Scopus as our search engine for the first step of the search process. Scopus was chosen for being one of the most comprehensive databases of scientific literature and for its advanced search capabilities (Gusenbauer, & Haddaway, 2020). In addition, it employs rigorous quality control measures to ensure the quality and accuracy of the indexed literature, which helps to minimize the risk of low-quality articles. We specifically searched for: (AI OR "artificial intelligence") AND (accountability OR accountable) in the abstract, title or authored-defined keywords. We decided not to restrict the year of publication as we are interested in including any relevant paper from the early eras of AI research until the conclusion of this literature review (October 2022). In total, 1090 papers were identified in this first phase.

Selection and Quality Appraisal. Articles were imported to a shared spreadsheet and manually screened by all authors collaboratively. We first considered the titles, then the abstracts, and finally the full texts. When abstracts were not available, we read the full texts. Following common practices used in good quality literature reviews (Vrontis & Christofi, 2019) we established specific exclusion criteria to reduce selection bias, guarantee the quality of the papers selected, and increase the review validity. We excluded papers that did not meet one of the following criteria:

- Articles should clearly address managing AI
- Articles should not only include a presentation of technical designs, unless they also include a discussion of accountability in managing AI
- Articles should be written in English and be peer-reviewed research papers

Therefore, documents that are not research papers (e.g., interviews, research proposals), are not focusing on AI systems (but only causally mention AI) or not engaging with AI management (but only casually mention accountability) were excluded.

To ensure the inclusion of as many relevant papers as possible, we performed a backward and forward search to identify potentially interesting articles that were cited by the articles resulting from the search described above. These articles were screened based on the exclusion criteria.

After following this process, we ended up with a corpus of 131 research papers that span across different research fields (Table 1).

Research Field	Number of publications
Computer Science and Software Engineering	39
Law	25
Health and Medical Sciences	16
Social Science	15
Information Systems	11
Human Computer Interaction	8
eGovernment and Public Administration	7
Organization, Management, Information Science	6
Education	4
Total	**131**

Table 1. Corpus of papers analyzed in this review

Examination and Synthesis. We followed a thorough coding process. We defined an initial coding protocol based on the research questions. We initially coded the same subset of the papers′ corpus independently in parallel to strengthen intercoder reliability. We then discussed our coding. This helped us to align in a common coding practice. All remaining papers were distributed among the authors. To ensure validity, the emerging results were discussed among all authors in regular meetings. The codes were derived through an inductive-deductive approach inspired by a hermeneutic tradition (Boell & Cecez-Kecmanovic, 2014) in which our understanding gradually emerged through iterative refining of the codes.

For synthesizing the papers, we were interested in exploring definitions and conceptualizations of AI accountability in the identified literature (see RQ1). In this process, we gradually engaged with theoretical imports (Boell & Cecez-Kecmanovic, 2014). Bovens and colleagues' conceptualization of accountability (Bovens, 2007, 2010; Bovens et al., 2014) emerged as an important conceptual apparatus to unpack how accountability unfolds in different domains. As a result, we decided to classify the identified articles along the three dimensions of accountability proposed by Bovens and colleagues (2014): obligation, interrogation, and sanctioning. Furthermore, we worked to derive insights from the literature on AI accountability that inform the management of AI. To do this, we surfaced the different levels of analysis in the literature reviewed. The resulting categories are (i) system accountability, (ii) process accountability, and (iii) data accountability. The findings of our examination and synthesis are presented in the section that follows.

4 Findings

4.1 AI Accountability Definitions

Our first research question is: *How is AI accountability defined and conceptualized in the literature*? An interesting finding is that a significant part of the literature reviewed lacks conceptual clarity. Almost one out of two (specifically, 56 out of 131) reviewed papers use the term accountability without defining what the term means. This is a significant issue as the accountability concept is quite malleable and its liberal use can lead to confusion.

Among the papers that do define accountability, several focus on the obligation of those involved in AI lifecycles to account for their actions. In this perspective, accountability is conceptualized as a particular form of responsibility (for instance: Hayes et al., 2020; Milosevic, 2019; Verdiesen et al., 2021). A different theme of accountability conceptualizations focuses on the interrogation ability of AI stakeholders (including users, affected parties and regulators). In this second perspective, accountability is conceptualized in close relation to transparency and auditability. For instance, Pedersen and Johansen state: "accountability is a term that comprises important aspects such as the provenance of, access to, transparency of, and auditability of, algorithms and data" (2020, p. 520 p. 520). We also identified a third theme: conceptualizing accountability by focusing on the post-hoc sanctioning of blamable agents (when things go wrong). In this third perspective, accountability is conceptualized in close relation to liability. For instance, Ibrahim and colleagues argue for developing systems that can provide evidence for the causes of undesired events and explain that: "accountability in this context refers to developing a system's (forensic) capabilities in holding misbehaving parties responsible for violations. In the case of a drone crash, it is imperative to find and address the root cause to prevent future mishaps; in aircraft accidents, accountability is part of the judicial process to assign liability and responsibility." (Ibrahim et al., 2020 p. 2978).

Overall, accountability is defined in part of the literature in a narrow way focusing on one or two of the three perspectives identified but we also found a significant number of papers (specifically, 19 papers) that define the concept adopting a quite comprehensive definition. In these papers, accountability is viewed not only as a matter of actors assuming responsibility for their actions, but also covers mechanisms for transparency and liability. The comprehensive view on AI accountability corresponds with Bovens and colleagues (2014) definition and includes a) the obligation of those involved in AI development and deployment to answer for and justify actions, b) the ability of stakeholders to interrogate about AI and c) the sanctioning ability when AI systems work in unacceptable ways. Table 2 provides an overview of the different perspectives on accountability found in the literature reviewed.

Dimension	**Number of publications** *(some publications include more than one dimensions)*
Obligation of those involved in AI development and deployment to answer for and justify actions	53
Interrogation ability of stakeholders about AI	41
Sanctioning ability when AI systems work in unacceptable ways	37

Table 2. Perspectives on Accountability in the Literature Reviewed

4.2 AI accountability in the management of AI

Our second research question is: *How can the literature on AI accountability inform the management of AI?* In answering this question, we sought to understand different approaches for going from principles to practice. As illustrated in figure 1, we found that the majority of papers deal with accountability on the level of values and principles. How these values and principles can be translated into AI management practices is not always clear. However, we also identified papers that do operationalise the abstract concepts and can inform practice.

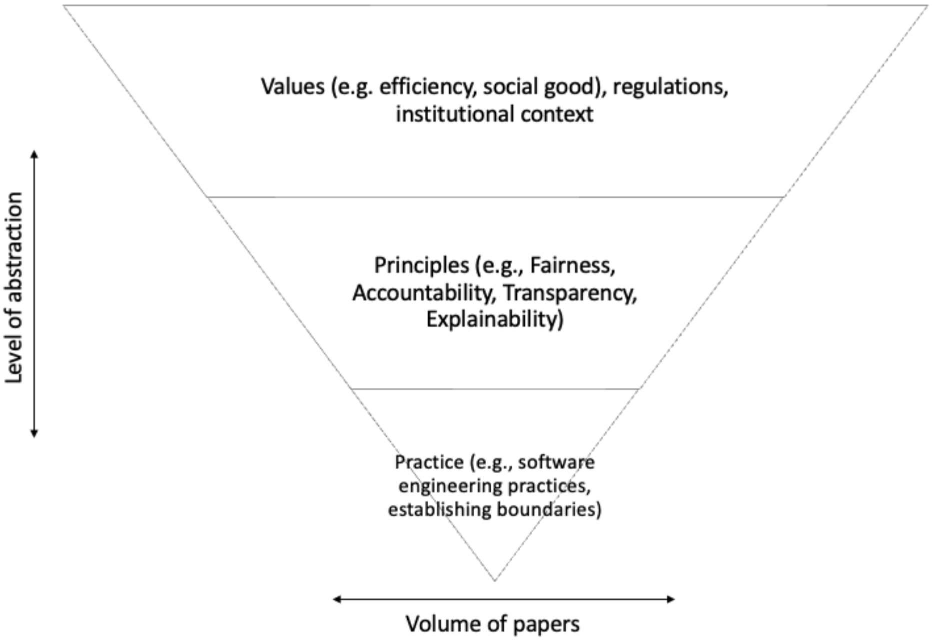
Values (e.g. efficiency, social good), regulations, institutional context
Principles (e.g., Fairness, Accountability, Transparency, Explainability)
Practice (e.g., software engineering practices, establishing boundaries)
Level of abstraction
Volume of papers

Figure 1. Different levels of abstraction in the literature reviewed

Reviewing the papers that operationalise the high-level values and principles, we derived three categories of approaches to AI accountability that form our proposed framework on AI Accountability at three different levels of analysis: system, process, and data. It is interesting to observe that the approaches proposed in the literature and captured by the three categories are quite varied. They range from approaches to raise societal awareness to principles for organizing software development practices. Table 3 provides an overview of the findings per level pointing to arrangements that need to be in place to make AI Accountability possible.

Level	Description	Main findings	Key references
System	Describes how accountability can be achieved by configuring sociotechnical systems of software and human actors (including human-in-the-loop and human-outside-the-loop)	Accountability by design (exploring design characteristics for AI accountability) Accountability rooted in societal awareness about the significance of different human/AI configurations. Accountability partially relinquished within bounded environments whose inputs and outputs are controlled.	Adams & Hagras, 2020; Addis & Kutar, 2019; Arrieta et al., 2020; Asatiani et al., 2021; Bogina et al., 2022; Chiao, 2019; Gualdi & Cordella, 2021; Janssen et al., 2022; Kim et al., 2020; Knowles & Richards, 2021; Liu et al., 2019; Naiseh et al., 2020; Nussbaumer et al., 2023; Rjoob et al., 2020; Sjöström et al., 2022; Tambe et al., 2019; Vassilakopoulou, 2020
Process	Describes how accountability can be achieved by structuring AI lifecycle processes (including development, evaluation, and monitoring)	Accountability can be achieved through software engineering practices and by modifying the development lifecycle. Accountability can be achieved by introducing accountability gates in AI development.	Baird & Maruping, 2021; Cobbe et al., 2021; Hutchinson et al., 2021; Kroll, 2018; Raji et al., 2020; Vakkuri et al., 2019

Data	Describes how accountability can be achieved by foregrounding the infrastructure and work that shape the data	Accountability can be achieved by unearthing and tracing mundane data work involved in AI lifecycles.	Hutchinson et al., 2021; Orr & Davis, 2020; Tarafdar et al., 2023

Table 3. Making AI accountability possible at the system, process and data level

4.2.1 System

The first identified category of research articles engages with AI accountability at the *system* level encompassing studies that propose achieving accountability by configuring sociotechnical arrangements of software and human actors. The design of such sociotechnical systems includes arrangements like human-in-the-loop and human-outside-the-loop. We identified three subgroups of approaches within this category.

The first and most frequently found approach is *accountability by design.* With this term, we describe approaches suggesting that accountability can be achieved (completely or partly) through the appropriate design of AI-infused sociotechnical systems. For example, Sjöström and colleagues take a design science research approach to build and evaluate software that embeds mechanisms for the governance of privacy and accountability in healthcare (Sjöström et al., 2022). Nussubaumer and colleagues propose an ethics-by-design approach to develop and implement a decision support system for emergency management (Nussbaumer et al., 2023). Vassilakopoulou (2020) proposes the reuse of classical sociotechnical design principles for accountability by design through regulation and operational coordination. Explainability is often identified as a key prerequisite for accountability by design, because it can help humans to better control AI and enable auditing AI for regulatory compliance (Arrieta et al., 2020; Rjoob et al., 2020). A large proportion of papers in this group, conceptualize accountability as an effect of AI explainability. While most such papers discuss this topic technically with regards to algorithmic design (e.g. Adams & Hagras, 2020; Kim et al., 2020), there is also research with a focus on the interaction between systems and users, discussing when and how explanations should be given to enable accountability (Arrieta et al., 2020; Naiseh et al., 2020). Taking the sociotechnical argument further, Gualdi and Cordella, suggest opening algorithms, and exposing them to public scrutiny to direct light on the accountability of the assemblage constituted by technological, institutional and legal dimension and not independently on each dimension (Gualdi & Cordella, 2021). Researchers also point out that although explainability is important for accountability it is not sufficient by itself. For instance, Janssen and colleagues (2022) performed empirical research for accountability in the context of AI in public services and found that ensuring the ability to understand needs to be complemented with training of decision makesrs and careful algorithm choices in the first place.

The second approach identified in the literature suggests rooting accountability in *societal awareness* about the significance of different humans/AI configurations. Several researchers point to the need for taking action to develop this awareness. For example, Addis and Kutar, observe that there is often a lack of understanding of the need for AI accountability, and what this implies, both among managers and in general outside computer science specialists (Addis & Kutar, 2019). One top-down avenue to deal with this problem is to publish informative guidelines to support the management of AI in organizations or governmental agencies and to develop and distribute educational material (Bogina et al., 2022; Chiao, 2019; Liu et al., 2019). Along a different avenue, some scholars target the relationship between accountability of AI and public discourse, by proposing standards for talking about AI accountability in public fora such as mass media, in addition to standards for documenting and reporting decisions. Knowles and Richards (2021) present a theoretical framework that accounts for the distinct institutional nature of public trust in AI and the new role documentation could play in fostering accountability. An interesting suggestion for strengthening awareness was put forward by Tambe and colleagues (2019) in the context of AI for human resource management. They proposed the creation of "AI Councils" with stakeholders' representatives that should debate the assumptions and the data that are to be fed into AI models.

The third approach relates to *setting general boundary conditions* around AI-enabled systems, that is, defining and controlling the borders for AI operations. Delimiting AI within well-defined ranges of operation can be a away of ensuring accountability even when AI remains intentionally inscrutable due to IP rights or privacy concerns (Burrell, 2016). A promising concept to formalize this approach is given by Asatiani and colleagues (2021) who propose establishing clear boundaries within which AI is to interact with its surroundings, choosing and curating training data carefullu, managing both input and output data appropriately while being able to compromise some explainability in favor of accuracy.

4.2.2 Process

The second category of approaches for ensuring AI accountability in practice relates to *process accountability*. Overall, this category captures approaches that focus on the lifecycles of AI-infused systems (including not only design, but also development, evaluation, and monitoring) and propose to achieve accountability by structuring them purposefully. For example, Baird and Maruping (2021) present AI accountability as an element in the delegation of work to AI algorithms that involves the alignment of new dependencies across human and algorithmic actors throughout AI lifecycles. Several papers in this category posit that accountability can be achieved through *software engineering practices and changes in the development lifecycle*. Kroll (2018) argues that in the context of developing AI, software engineering needs to reflect human values, and he calls for research on how practices like requirements engineering can change towards this goal. Furthemore, Vakkuri and colleagues (2019) performed empirical reseach on ethical AI in the industry and came up with the conclusion that responsibility of developers and development is under-discussed motivating more research on development processes.

A subgroup of papers concretize this view by showing how accountability can be achieved by modifying AI development lifecycles. AI development lifecycles consist of sets of steps that are iteratively performed during both design and use of the systems. Cobbe and colleagues (2021) specify these steps as commissioning (which includes ideation and possible procurement processes), model development, decision making (pertaining to use and operations) and investigation (including auditing). The authors argue that every step needs to be designed with accountability in mind. For example, during the model development stage, information about the selection of training data needs to be recorded to enable audit into potential bias. During the decision-making stage, there needs to be logging mechanisms in place that records details of inputs and outputs, which makes it possible to investigate what has occurred during use and detect failures. Together, changes in the lifecycle make it possible to perform an internal or external audit. This approach, therefore, follows recent calls from both research and policy that responsible and trustworthy AI depends on auditing mechanisms (Mökander & Floridi, 2021).

A variant of the same approach is to *inject accountability gates in AI lifecycles.* Following a stage-gate approach, accountability can be achieved by making defined activities obligatory before one can proceed from one stage to another. Hutchinson and colleagues (2021) provide a version of this approach, where they stress the importance of accompanying each stage of the development lifecycle with specific documentation practices. These practices enable the previous stage of the lifecycle to be audited while the next stage proceeds in an accountable manner. After passing each of the gates, the development lifecycle will have produced a set of auditable documents (Raji et al., 2020). This documentation should also account for stakeholder engagement processes behind the requirement specifications.

4.2.3 Data

The third and last category of approaches for ensuring AI Accountability in practice is foregrounding *data work*. This category encompasses a limited but growing set of studies that take data as their unit of analysis and propose to achieve accountability by foregrounding the infrastructure and work that shape data. This category too stresses the need for achieving accountability by design, but it proposes to do so by understanding how *decisions are taken as part of collective data work practices*. This category is partly related to the previous, in that it takes a processual perspective as a starting point. For example, a notable contribution is provided by Hutchinson and colleagues (2021) who develop a framework to report all decisions that are taken about the data involved in the AI-enabled system stemming from a lifecycle-perspective on system development inspired by software engineering. They draw on an understanding of data as infrastructures, that is, sociotechnical arrangements that shape what and how we can know about the world (Bowker & Star, 2000). Although reminiscent of waterfall software engineering approaches, the framework by Hutchinson and colleagues is primarily a s*tep-by-step bookkeeping or tracing approach* to document what data are needed and why, who uses them and how, who stores them and them, how training sets are defined and tested, and the data maintained.

Understanding the actual data work performed throughout the AI systems' lifecycle is crucial in a data accountability perspective. Inspired by a materialized action approach and actor-network theory, Orr and Davis (2020) study AI practitioners in their day-to-day work to develop AI systems. The authors find that practitioners have a central role in distributing ethical responsibilities across a range of different actors as they take apparently mundane decisions about the data and the algorithms when designing systems. An important implication of this finding is that accountability and ethics cannot be defined ex ante but emerge as part of collective data work practices. Tarafdar and colleagues (2023) take this perspective in investigating human-algorithm daily interactions in algorithmic work in the case of Uber drivers and uncover ambiguities in the roles taken and assigned by the Uber algorithm. In general, these studies demonstrate that data accountability requires in-depth analysis of actual practices.

5 Discussion

With this literature review, we extend the literature on AI management by framing AI accountability as a socio-technical and organizational problem that IS researchers are well-equipped to study. As AI technologies are becoming increasingly widespread in organizations, IS scholars have been working to identify and investigate emerging AI management challenges. They have engaged with questions like how tasks can be delegated between people and technologies (Baird & Maruping, 2021), how one can handle algorithmic bias and discrimination (Dolata et al., 2022; Kordzadeh & Ghasemaghaei, 2022), and how to approach tensions between groups of workers and invisible workarounds (Pachidi et al., 2021). However, while accountability is recognized as an important organizational issue (Karunakaran et al., 2022) and the literature on AI has identified accountability as a societal concern, there is a lack of comprehensive research in the IS field into the challenges AI bring in terms of accountability and the socio-technical responses to these challenges (Dolata et al., 2022). Nevertheless, there is a growing body of research related to AI accountability in disciplines outside IS. The time is right for taking stock of the literature making it possible for IS researchers to relate their own work and contribute. IS research is well-positioned to contribute to the AI accountability discourse bringing together technical, organizational and ethical perspectives (Berente et al., 2021; Sarker et al., 2019).

An important argument in the extant AI accountability literature is that artefacts such as AI models and AI-enabled systems cannot be held accountable in themselves; accountability is reserved for humans and organizations (Raji et al., 2020; Singh et al., 2019). Hence, ensuring accountability entails making possible to determine which humans and organizations are involved in AI development and deployment. As Ågerfalk reminds us, "an algorithm is an algorithm. Until we have reached technological singularity, humans develop algorithms (Makridakis, 2017). Machine learning as a form of automated action means that systems may modify their behaviour over time. The boundaries of such modifications are still managed by humans within technological, organisational and institutional frames. It is not a question of monsters but of agency concerning explainable and accountable AI" (Ågerfalk, 2020 p.5) This paper's framing of AI accountability as an issue of IS management affords researchers to investigate the conditions under which AI systems are designed and developed. It also couples these conditions to the use of AI, as the characteristics of AI make it challenging to demarcate the boundaries between the agency of human users making decisions and the predictions and recommendations made by technology, not least when system use also provides training data that feed and renew models. AI accountability thus provides a lens to study when, where, and by whom decisions are taken.

We primarily contribute to the IS literature on managing AI by providing (i) conceptual clarification by mapping different accountability conceptualizations; (ii) a framework encompassing actionable responses to address AI accountability in practice on the level of systems, processes, and data. Conceptually, accountability points to the relationship between actors´ responsibilities and their enactments through visibility and liability (Boos & Grote, 2012). As shown in the literature review, however, scholars have tended to adopt loose or limited definitions of what accountability entails. We, therefore, recommend that IS researchers follow a comprehensive view on AI accountability that includes a) the obligation of those involved in AI development and deployment to answer for and justify actions, b) the ability of stakeholders to interrogate about AI and c) the sanctioning ability when AI systems work in unacceptable ways.

Our second contribution to IS is synthesizing existing approaches to operationalize accountable AI into a framework (Table 3). This framework classifies the approaches to the levels of system, process, and data, as responses to accountability challenges posed by AI. As such, we contribute with what Gregor refers to as Type 1 theory, as a description and classification that provides an overview of current knowledge and aims to be helpful in further analysis (Gregor, 2006). Most research to date – across disciplines –focused on accountability on an abstract level, in terms of values and principles. Accountability is an organizational and socio-technical issue, meaning that it is a concern IS researchers are especially well-equipped to approach. We believe our proposed framework can be a valuable starting point for further research. We emphasize data accountability - as one level in the framework - as an important venue for IS research. Data has become a primary interest in our field, and as Pentinnen and Aaltonen recently proposed, IS is in a position to evolve into a "material science" of the digital economy (Aaltonen & Penttinen, 2021). Researchers have argued and shown that data is never a neutral resource or a simple representation of reality but the results of practical socio-technical endavours encompassing chains of activities carried out by diverse actors (Jones, 2019; Parmiggiani et al., 2022). Data always results from some nexus of practices and technology, and data is the main reason why an AI model ends up in the way it does. It follows logically that data accountability is essential for AI accountability.

These two contributions form the basis of our third contribution; foregrounding accountability as a way to balance goals of efficiency and humanistic values in managing AI. This is further elaborated in the subsection that follows.

5.1 Accountability to balance the aims for efficiency and humanistic values

The AI discourse in IS tends to drift in two directions: on the one hand we find an instrumental perspective, often accompanied by rhetorics highlighting the potential of increased efficiency, effectiveness and speed for organizations (Borges et al., 2021; Brynjolfsson & Mcafee, 2017). On the other hand, some scholars take a humanistic perspective, proposing ethical guidelines, stressing the negative potential consequences or dark sides of AI for humanity (Floridi et al., 2018; Mikalef et al., 2022; Vassilakopoulou et al., 2022) and the risk for "a dystopian future state where ubiquitous data collection feeds ML systems that users do not understand, that lack user feedback, and that result in behavioral control of humans using internet-based platforms" (Kane et al., 2021, p. 372 p. 372).

In terms of the nascent conversation on managing AI, these two directions stretch along a sociotechnical 'axis of cohesion' that drives IS research (Benbya et al., 2021; Sarker et al., 2019). Both views have a lot to offer. For example, an instrumental view to managing AI has the potential to provide organizations with roadmaps for developing capabilities to get actionable insights and for orchestrating new types of AI-infused resources. The humanistic perspective has the strength of reminding us that human beings should be centerstage, thus preventing technology-driven oppressive agendas. It is also useful on the normative level informing policies on how to regulate AI.

The way we deal with the relation between instrumental and humanistic perspectives is important, because it has consequences for the type of future we envision for our societies and for the way forward in IS research (Sarker et al. 2019). We extend the IS literature by proposing to *approach AI accountability as a means to jointly consider instrumental and humanistic outcomes.* These two types of outcomes together can form a virtuous cycle wherein both are synergistically connected (idem). We do not find it useful to consider the relation between instrumental and humanistic perspectives as tension or dualism. On the contrary, our analysis of the literature illustrates that it can be considered as a duality: these two perspectives while conceptually distinct can be mutually enabling and a constituent of one another (Farjoun, 2010). Our aim is not to advocate for AI accountability as a regime of strict “AI policing”. As Kane and colleagues warn, strict calls for accountability might have the paradoxical effect of turning to authoritarianism (Kane et al., 2021). Even more importantly, we are not putting forward AI accountability as ethics-washing (Bietti, 2020) to facilitate business opportunities associated with AI. A fertile middle ground can exist for AI accountability that leads to awareness and mitigation of AI’s multifaceted risks for the realization of the considerable potential benefits of AI.

Our review illustrates that it is possible to enact accountability in practice at the system, process and data level (Table 3). It can be useful to define a buffer space for balancing performance benefits and the risks associated with AI (Asatiani et al., 2021) while still disrupting the path that ignores users’ humanity (Kane et al., 2021). Future applications of accountable AI can develop in the middle ground between instrumental and humanistic approaches, balancing efficiency and risks. We envision future AI accountability research in IS to *proactively explore the boundary conditions of this middle ground between business opportunities and ethical concerns of AI*. Several papers in this literature review provide evidence of promising efforts in this direction.

For the future, it will be crucial to cover system, process and data levels as a whole. As our review demonstrates, the data level – data accountability – is particularly underdeveloped in AI accountability research. This is an important reason for IS researchers to pay more attention to how data are chosen, prepared, cleaned and reused (Parmiggiani et al. 2022). We single out this, as an important venue for IS research. From this perspective, we envision that future research in IS could draw on data studies and *further investigate how AI accountability can encompass strategies to trace and manage data decisions*.

6 Conclusion

Part of the widespread discourse on AI is centered on the promise of providing solutions to problems by utilizing limited resources as AI is becoming increasingly easy to develop and deploy. However, there is no silver bullet: work and resources are needed for establishing AI accountability mechanisms that will allow societies to continue innovating while mitigating the risks. The UN has already asked for putting AI in halt till we put appropriate safeguards in place (United Nations, 2021). The IS field has much to offer in approaching the issue. This paper provides a sound basis for IS research on AI accountability by synthesizing the literature, clarifying different conceptualizations and providing directions for research through the development of an analytical framework. By emphasizing the balance between humanistic and instrumental values (Sarker et al. 2019) and by utilizing the field’s methods and concepts for studying complex sociotechnical managerial issues, IS researchers can contribute insights for establishing AI accountability at the system, process and data levels.

References

Aaltonen, A. and Penttinen, E. (2021). What makes data possible? A sociotechnical view on structured data innovations. *Proceedings of the 54th Hawaii International Conference on System Sciences, (HICSS)*, 5922–5931.

Adams, J. and Hagras, H. (2020). A type-2 fuzzy logic approach to explainable AI for regulatory compliance, fair customer outcomes and market stability in the global financial sector. *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE).*

Addis, C. and Kutar, M. (2019). AI Management: an exploratory survey of the influence of GDPR and FAT principles. *IEEE conference on SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation.*

Ågerfalk, P. J. (2020). Artificial intelligence as digital agency. *European Journal of Information Systems*, *29*(1), 1-8.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D. and Benjamins, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82-115.

Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T. and Salovaara, A. (2021). Sociotechnical envelopment of artificial intelligence: An approach to organizational deployment of inscrutable artificial intelligence systems. *Journal of the Association for Information Systems*, *22*(2), 325-352.

Australian Government, D. of Industry, Science and Resources. (2019). *Building Australia's artificial intelligence capability*. https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework/ai-ethics-principles

Baird, A. and Maruping, L. M. (2021). The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts. *MIS Quarterly*, *45*(1), 315-341.

Benbya, H., Pachidi, S. and Jarvenpaa, S. (2021). Special issue editorial: Artificial intelligence in organizations: Implications for information systems research. *Journal of the Association for Information Systems*, *22*(2), 281-303.

Berente, N., Gu, B., Recker, J. and Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, *45*(3), 1433-1450.

Bernardi, R. (2017). Health information systems and accountability in Kenya: A structuration theory perspective. *Journal of the Association for Information Systems*, *18*(12), 931 – 958

Bietti, E. (2020). From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy *ACM Conference on fairness, accountability, and transparency (FAccT 2020)*.

Boell, S. K. and Cecez-Kecmanovic, D. (2014). A hermeneutic approach for conducting literature reviews and literature searches. *Communications of the Association for information Systems*, *34*(1), 257-286.

Bogina, V., Hartman, A., Kuflik, T. and Shulner-Tal, A. (2022). Educating software and AI stakeholders about algorithmic fairness, accountability, transparency and ethics. *International Journal of Artificial Intelligence in Education*, *32*(3), 808-833.

Boos, D. and Grote, G. (2012). Designing Controllable Accountabilities of Future Internet of Things Applications. *Scandinavian Journal of Information Systems*, *24*(1), 3-28.

Borges, A. F., Laurindo, F. J., Spínola, M. M., Gonçalves, R. F. and Mattos, C. A. (2021). The strategic use of artificial intelligence in the digital era: Systematic literature review and future research directions. *International Journal of Information Management*, *57*, 102225.

Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework 1. *European law journal*, *13*(4), 447-468.

Bovens, M. (2010). Two Concepts of Accountability: Accountability as a Virtue and as a Mechanism. *West European Politics*, *33*(5), 946-967.

Bovens, M., Schillemans, T. and Goodin, R. E. (2014). Public accountability. *The Oxford handbook of public accountability*, *1*(1), 1-22.

Bowker, G. C. and Star, S. L. (2000). *Sorting things out: Classification and its consequences*. MIT press.

Brown, S. (2021). Machine learning, explained. *MIT Management Sloan School. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learningexplained.*

Brynjolfsson, E. and Mcafee, A. (2017). Artificial intelligence, for real. *Harvard Business Review*, *July*, 1-31.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512.

Chiao, V. (2019). Fairness, accountability and transparency: notes on algorithmic decision-making in criminal justice. *International Journal of Law in Context*, *15*(2), 126-139.

Cobbe, J., Lee, M. S. A. and Singh, J. (2021). Reviewable automated decision-making: A framework for accountable algorithmic systems. *ACM Conference on fairness, accountability, and transparency (FAccT 2021)*.

Dolata, M., Feuerriegel, S. and Schwabe, G. (2022). A sociotechnical view of algorithmic fairness. *Information Systems Journal*, *32*(4), 754-818.

Farjoun, M. (2010). Beyond dualism: Stability and change as a duality. *Academy of Management Review*, *35*(2), 202-225.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U. and Rossi, F. (2018). AI4People—An ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689-707.

Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly*, *30*(3), 611-642.

Gualdi, F., & Cordella, A. (2021). Artificial intelligence and decision-making: The question of accountability. Proceedings of the 54th Hawaii International Conference on System Sciences. *(HICSS)*, 2297-2306.

Gusenbauer, M. and Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods*, *11*(2), 181-217.

Hayes, P., Van De Poel, I. and Steen, M. (2020). Algorithms and values in justice and security. *AI & Society*, *35*(3), 533-555.

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P. and Mitchell, M. (2021). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. *ACM Conference on fairness, accountability, and transparency (FAccT 2021).*

Ibrahim, A., Klesel, T., Zibaei, E., Kacianka, S. and Pretschner, A. (2020). Actual causality canvas: a general framework for explanation-based socio-technical constructs. *Twenty-fourth European Conference on Artificial Intelligence (ECAI 2020).*

Janssen, M., Hartog, M., Matheus, R., Yi Ding, A. and Kuk, G. (2022). Will algorithms blind people? The effect of explainable AI and decision-makers' experience on AI-supported decision-making in government. *Social Science Computer Review*, *40*(2), 478-493.

Jones, M. (2019). What we talk about when we talk about (big) data. *The Journal of Strategic Information Systems*, *28*(1), 3-16.

Kane, G. C., Young, A. G., Majchrzak, A. and Ransbotham, S. (2021). Avoiding an oppressive future of machine learning: A design theory for emancipatory assistants. *MIS Quarterly*, *45*(1), 371-396.

Karunakaran, A., Orlikowski, W. J. and Scott, S. V. (2022). Crowd-based accountability: Examining how social media commentary reconfigures organizational accountability. *Organization Science*, *33*(1), 170-193.

Kempton, A. M. (2022). The digital is different: Emergence and relationality in critical realist research. *Information and Organization*, *32*(2), 100408.

Kim, B., Park, J. and Suh, J. (2020). Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information. *Decision Support Systems*, *134*, 113302.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele University Technical Report, UK*, *TR/SE-0401*(2004), 1-26. https://doi.org/https://doi.org/10.1.1.122.3308

Knowles, B. and Richards, J. T. (2021). The sanction of authority: Promoting public trust in AI. *ACM Conference on fairness, accountability, and transparency (FAccT 2021).*

Kordzadeh, N. and Ghasemaghaei, M. (2022). Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388-409.

Kroll, J. A. (2018). The fallacy of inscrutability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), 20180084.

Lee, A. S. (2001). Editor's comments: research in information systems: what we haven't learned. *MIS Quarterly*, *25*(4), v-xv.

Liu, H.-W., Lin, C.-F. and Chen, Y.-J. (2019). Beyond State v Loomis: artificial intelligence, government algorithmization and accountability. *International Journal of Law and Information Technology*, *27*(2), 122-141.

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

Milosevic, Z. (2019). Ethics in Digital Health: a deontic accountability framework. *IEEE 23rd International Enterprise Distributed Object Computing Conference (EDOC 2019).*

Mökander, J. and Floridi, L. (2021). Ethics-based auditing to develop trustworthy AI. *Minds and Machines*, *31*(2), 323-327.

Naiseh, M., Jiang, N., Ma, J. and Ali, R. (2020). Explainable recommendations in intelligent systems: delivery methods, modalities and risks. *International Conference on Research Challenges in Information Science*.

Nussbaumer, A., Pope, A. and Neville, K. (2023). A framework for applying ethics-by-design to decision support systems for emergency management. *Information Systems Journal*, *33*(1), 34-55.

Orr, W. and Davis, J. L. (2020). Attributions of ethical responsibility by Artificial Intelligence practitioners. *Information, Communication & Society*, *23*(5), 719-735.

Ortiz de Guinea, A. and Paré, G. (2017). What literature review type should I conduct? *The Routledge Companion to Management Information Systems* (pp. 73-82). Routledge.

Pachidi, S., Berends, H., Faraj, S. and Huysman, M. (2021). Make way for the algorithms: Symbolic actions and change in a regime of knowing. *Organization Science*, *32*(1), 18-41.

Parmiggiani, E., Østerlie, T. and Almklov, P. G. (2022). In the Backrooms of Data Science. *Journal of the Association for Information Systems*, *23*(1), 139-164.

Pedersen, T. and Johansen, C. (2020). Behavioural artificial intelligence: an agenda for systematic empirical studies of artificial inference. *AI & Society*, *35*(3), 519-532.

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137-141.

Rai, A., Constantinides, P. and Sarker, S. (2019). Editor's comments: next-generation digital platforms: toward human–AI hybrids. *MIS Quarterly*, *43*(1), iii-x.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *ACM Conference on fairness, accountability, and transparency (FAccT 2020)*.

Rjoob, K., Bond, R., Finlay, D., McGilligan, V., Leslie, S. J., Rababah, A., Iftikhar, A., Guldenring, D., Knoery, C. and McShane, A. (2020). Towards Explainable Artificial Intelligence and Explanation User Interfaces to Open the 'Black Box'of Automated ECG Interpretation. *Advanced Visual Interfaces. Supporting Artificial Intelligence and Big Data Applications* (pp. 96-108). Springer.

Sarker, S., Chatterjee, S., Xiao, X. and Elbanna, A. (2019). The sociotechnical axis of cohesion for the IS discipline: Its historical legacy and its continued relevance. *MIS Quarterly*, *43*(3), 695-720.

Schryen, G., Wagner, G. and Benlian, A. (2015). Theory of knowledge for literature reviews: an epistemological model, taxonomy and empirical analysis of IS literature. 36th International Conference on Information Systems (ICIS 2015).

Scott, S. V. and Orlikowski, W. J. (2014). Entanglements in practice. *MIS Quarterly*, *38*(3), 873-894.

Singh, J., Cobbe, J., & Norval, C. (2019). Decision provenance: Harnessing data flow for accountable systems. *IEEE Access*, *7*, 6562-6574.

Sjöström, J., Ågerfalk, P. and Hevner, A. R. (2022). The Design of a System for Online Psychosocial Care: Balancing Privacy and Accountability in Sensitive Online Healthcare Environments. *Journal of the Association for Information Systems*, *23*(1), 237-263.

Strandburg, K. J. (2019). Rulemaking and inscrutable automated decision tools. *Columbia Law Review*, *119*(7), 1851-1886.

Tambe, P., Cappelli, P. and Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*, *61*(4), 15-42.

Tarafdar, M., Page, X. and Marabelli, M. (2023). Algorithms as co-workers: Human algorithm role interactions in algorithmic work. *Information Systems Journal*, *33*(2), 232-267.

United Nations. (2021). *Urgent Action Needed over Artificial Intelligence Risks to Human Rights*. Retrieved 25 June from https://news.un.org/en/story/2021/09/1099972

Vakkuri, V., Kemell, K.-K. and Abrahamsson, P. (2019). AI Ethics in Industry: A Research Framework. *CEUR Workshop Proceedings*.

Vance, A., Lowry, P. B. and Eggett, D. (2015). Increasing Accountability Through User-Interface Design Artifacts. *MIS Quarterly*, *39*(2), 345-366.

Vassilakopoulou, P. (2020). Sociotechnical Approach for Accountability by Design in AI Systems. *28th European Conference on Information Systems (ECIS 2020)*.

Vassilakopoulou, P., Parmiggiani, E., Shollo, A. and Grisot, M. (2022). Responsible AI: Concepts, critical perspectives and an Information Systems research agenda. *Scandinavian Journal of Information Systems*, *34*(2), 89-104.

Verdiesen, I., Santoni de Sio, F. and Dignum, V. (2021). Accountability and control over autonomous weapon systems: A framework for comprehensive human oversight. *Minds and Machines*, *31*(1), 137-163.

Vrontis, D. and Christofi, M. (2019). R&D internationalization and innovation: A systematic review, integrative framework and future research directions. *Journal of Business Research*. *128*, 812-823.

Whittaker, M. (2021). The steep cost of capture. *Interactions*, *28*(6), 50-55.