Digital Social Science

Innovative

Review

Ethical Leadership in the Age of Artificial Intelligence: A Systematic Literature Review

Mobarak Hossain*

Department of Business Administration, Bangladesh Islami University, Dhaka 1214, Bangladesh
* Correspondence: mobarakru1987@gmail.com

Abstract: Artificial Intelligence (AI) has emerged in organizational settings at a swift pace, fundamentally reshaping the comprehension of ethical leadership. The study, called a systematic literature review, analyzed 87 peer-reviewed studies that were published in the period from 2018 to 2025, to see how AI technologies transform the roles, resources and ethical landscape of organizational leaders. Algorithmic decision-making, automated surveillance and AI-enhanced governance are transforming the existing discourse on ethical leadership – whether that's transformational leadership, servant leadership, or authentic leadership – and we take a look at how they're doing it in this article based on the PRISMA 2020 guidelines. The five thematic clusters we identified are: (1) shifting the moral responsibility for AI from humans to AI systems, (2) algorithmic bias as an ethical leadership problem, (3) transparency and explainability in securing trust of AI system leadership, (4) data governance and data privacy as leadership needs, and (5) building new ethical leadership skills for AI. We suggest an integrative theory, the Responsible AI Leadership Model (RAILM), to combine themes and provide directions for empirical studies in the future.

Keywords: ethical leadership; artificial intelligence; algorithmic decision-making; responsible AI; organizational ethics; digital leadership

Citation: Mobarak Hossain. 2025. Ethical Leadership in the Age of Artificial Intelligence: A Systematic Literature Review. *Digital Social Sciences 2(2),* 24-31. https://doi.org/10.69971/dss.2.2.2025.45

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license https://creativecommons.org/licenses/by/4.0/.

1. Introduction

Ethical leadership and AI are among the most impactful areas in today's management literature. Organizational leaders are now encountering new ethical dilemmas that are not covered by current ethical frameworks, as AI systems increasingly make decisions on everything from hiring to strategic investment decisions (Floridi, Josh and Thomas et al. 2021; Mittelstadt and Patrick Russell 2022). Now the issue isn't whether AI will impact organizational life, but whether leaders have the ethical insight needed to lead this change in a responsible way. Earlier, when decisions were attributed to discrete human actors, ethical leadership was theorized as the manifestation of normatively appropriate behavior in one's personal behavior and interpersonal relationships (Brown, Linda and David 2005). The deployment of AI today can dramatically change this traceability. When a machine learning model refuses to finance a loan, suggests a workforce cut, or decides a treatment protocol for a patient, the responsibility shifts to the data scientists, algorithm design, procurement officers, and senior executives that may not have the technical expertise to question the systems that they deploy (Mittelstadt, Patrick, and Sandra et al. 2022; Rinta-Kahila Tapani, Ida, and Nicole et al. 2023). The academic answer to this challenge has been disjointed. Important contributions have been made by researchers in management science, computer ethics, organizational behavior and information systems, but integrative synthesis is lacking in development. To fill this gap, this systematic literature review will chart the conceptual landscape between the theory of ethical leadership and AI governance, mapping the extent of convergence and controversy, and an integrated approach to inform research and practice. This review has the following organization. Section 2 provides the theory and concept supports. The methodology is described in section 3 following PRISMA 2020 reporting guidelines and consists of a search strategy, inclusion criteria and analytic approach. The results are presented under five emerging themes in Section 4. In Section 5, the Responsible AI Leadership Model (RAILM) is presented. Implications for research, practice and policy are discussed in section 6. Section 7 concludes.

2. Theoretical Background

2.1 Classical Ethical Leadership Theory

Ethical leadership scholarship is grounded in normative philosophy with a focus on Kantian deontology, utilitarian consequentialism and virtue ethics concepts translated into behavioral and social learning models that are relevant to organizational settings (Treviño, Laura and Michael 2000). The major paradigm in the field, developed by Brown, Treviño, and Harrison (2005), conceptualized ethical leadership as a dyadic, socially learned process, that is, ethical leadership is learned by followers from their leaders and the leaders behave in an ethical way in order to hold followers accountable by providing them with rewards and punishments and communicating openly about ethics. This perspective holds leadership to be fundamentally relational—ethics becomes actual through visible human actions. This framework has been enhanced by further theoretical developments. Transformational leadership (Bass and Roland 2006; Avolio, Fred and Todd 2009) explained that inspirational motivation and idealized influence are important factors that foster the moral reasoning of followers. Servant leadership theory (Greenleaf 1977; Dierendonck 2011) emphasized the leader's fiduciary duty to a group of stakeholders other than shareholders. Authentic leadership models (Luthans and Buce 2003; Walumbwa, Bruce and William et al. 2008) emphasized the importance of self-awareness, relational transparency, and internalized moral perspective in maintaining ethical behavior in stressful situations. The ethical leader is always a human agent, with agency, empathy, and moral deliberation, with whom other humans typically engage in an organizational context that is relatively straightforward to interpret. All these assumptions are challenged by the introduction of artificial intelligence in leadership environments (Vakkuri Kai-Kristian and Marianna et al. 2020; Bankins and Formosa 2023).

2.2 AI Governance and Ethics Frameworks

In tandem with the theory of ethical leadership, there has been a significant body of research on AI ethics in the last ten years. A number of high-profile governance frameworks have converged on a shared set of values: fairness, accountability, transparency, privacy, safety, and societal benefit (Jobin, Marcelo and Effy 2019; Fjeld, Nele and Hannah et al. 2020). However, these frameworks have been criticized for their aspirational nature and lack of organization operationalization (Mittelstadt 2019; Raji, Elizabeth and Aaron et al. 2022). When leaders say they want AI to be ‘fair’ or ‘transparent’, they do not give the AI any kind of instructions on how to identify the algorithmic discrimination in a deployed hiring system, nor how to communicate to the employees that AI-generated performance is a factor in promotion. However, the translation process from an ethical principle to a leadership practice needs an organizational actor, a leader, who grasps the technical and ethical aspects of the deployment of AI (Dignum 2020; Jobin, Marcelo and Effy 2019).

2.3 Bridging Leadership and AI Ethics

A growing body of research is starting to connect these two bodies of literature. (Floridi, Josh and Thomas et al. 2021; Taddeo and Floridi 2018; Metcalf, Emily and Danah 2016) have argued for the need of 'ethically aligned design' for the development of AI, which include organizational accountability structures. Leadership theory, on the other hand, is emerging from the leadership side, with (Nishii and Patrick 2008) and (Waldman and Robert 2014) arguing that the theory should adapt to the concept of distributed agency in sociotechnical systems, and more recently, (Bankins and Paul 2023) have done so. This review is part of that bridging process in systematically charting the empirical and conceptual work in this intersection.

3. Methodology

3.1 Review Design

The methodology used in this study is systematic literature review (SLR) according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines (Page, Joanne and Patrick et al. 2021). Systematic reviews are suitable for when researchers want to identify and summarize an impenetrable and fragmented literature through systematic, replicable search and inclusion processes (Tranfield, David, Palminder 2003; Petticrew and Helen 2006). The subject matter, which dealt with many aspects from management science, information systems, computer ethics to organizational behavior, required a systematic approach to achieve a comprehensive and unbiased review.

3.2 Search Strategy

The six major academic databases searched were: Business Source Complete (EBSCO), Web of Science, Scopus, JSTOR, ABI/Inform and Google Scholar. The search terms were also developed iteratively by using the Boolean operators and the three conceptual clusters: leadership terms ("ethical leadership" OR "responsible leadership" OR "moral leadership" OR "transformational leadership" OR "servant leadership"), AI/technology terms ("artificial intelligence" OR "machine learning" OR "algorithmic decision-making" OR "automation" OR "AI governance"), and organizational terms ("ethics" OR "trust" OR "accountability" OR "governance" OR "organizational behavior"). Searches were carried out in January 2025 and were revised up to March 2025.

3.3 Inclusion and Exclusion Criteria

Selection of articles was based on clear inclusion and exclusion criteria, to ensure the relevance and quality of the review. The selected articles were published between January 2018 and March 2025, which were mostly peer-reviewed journal articles. Studies that were eligible were those that examined the link between leadership, management, or organizational ethics and the application of artificial intelligence (AI), automation, or algorithmic systems. Only articles written in English were considered. Empirical studies, theoretical papers, systematic review and narrative review of literature were used in the review to get a comprehensive view of the topic. Book chapters, conference papers, grey literature, editorials, opinion pieces and other non-peer reviewed publications were not considered for studies. Articles that explored the topic of AI ethics only from a technical or engineering point of view, without addressing organizational or leadership aspects were also omitted. Further, redundant records and retracted publications were excluded from the final data set.

3.4 PRISMA Flow and Study Selection

The first database search brought back 4612 records. Of these 3,219 records, 675 were screened by title and abstract of these, 675 were screened by title and abstract after deduplication. Inclusion/exclusion criteria were used by two independent reviewers,

who resolved differences by arbitration with the third reviewer (Cohen's kappa = 0.81; strong interrater agreement). After a full-text review, 321 articles were included in a final sample of 87 articles that fulfilled all the inclusion criteria.

Table 1. PRISMA 2020 Flow Diagram Summary.

Stage	Records / Articles	Action Taken
Database search (6 databases)	4,612	Initial retrieval
After deduplication	3,219	Duplicates removed
Title/abstract screening	3,219	Excluded: 2,898 (not relevant)
Full-text eligibility review	321	Excluded: 234 (criteria not met)
Final included studies	87	Synthesized in review

3.5 Analytic Approach

Studies included were analyzed using thematic synthesis as presented by (Thomas and Angela 2008). This process was a three-step one, which included line-by-line coding of findings from each article, developing descriptive themes, and generating analytical themes that are not based on the content of individual studies, but rather provide new and interpretive perspectives. Bibliometric analysis was also performed using the software VOSviewer to visualize the co-citation clusters and to determine the intellectual communities that exist in the corpus.

4. Findings

4.1 Review Design Overview of the Literature

The 34 journals in which the 87 studies were published ranged from 1994 to 2002. The highest representation was from the Journal of Business Ethics (n = 14), Leadership Quarterly (n = 11), Journal of Management Information Systems (n = 9), Journal of Business Ethics (n = 8), and Organization Science (n = 7). It was observed that the number of publications was increasing sharply with only 8 qualifying articles in 2018–2019 and 31 in 2023–2024 indicating the field's swift emergence. In terms of methodologically, 44 percent were conceptual or theoretical, 31 percent were quantitative (survey or secondary data analysis), 16 percent were qualitative (interviews or case studies), and 9 percent were mixed methods. Thematic synthesis produced five key clusters which are explored in turn below: (1) moral agency and distributed responsibility, (2) algorithmic bias as an ethical issue for leadership, (3) transparency, explainability, and trust in leadership, (4) data governance and privacy as leadership concerns, and (5) AI ethical leadership capabilities.

4.2 Theme 1: Moral Agency and Distributed Responsibility

One of the main themes that frequently appear in the literature is the issue of moral agency displacement in AI-enriched organizations. The classical ethical leadership theory takes as its premise a clear conception of a single identifiable human agent who decides and is accountable for decisions to stakeholders. The assumption is being challenged by AI systems in at least three ways. The first is the introduction of a 'responsibility gap' as (Mittelstadt, Patrick and Sandra et al. 2022) put it, in which an adverse impact occurs but it is not easily traceable to any single individual human decision maker. (Rinta-Kahila, Tapani, Ida, and Nicole et al. 2023) reported this phenomenon in an insurance company case study using a machine learning model for claim adjudication. Senior leaders said they were unaware of the technical aspects of the system, data scientists said they were just doing their job, and the procurement team said they had trusted the vendors to tell them the truth. No one took responsibility for the lack of ethics. Second, a number of authors have contended that AI does not simply take the place of human moral agency, but that it generates new instances of 'distributed agency,' where humans and algorithms share responsibility in a manner that our current legal and ethical frameworks are inadequate to address (Floridi et al. 2021; Bankins and Formosa 2023). One of the most detailed theoretical accounts in the corpus is that of Bankins and Formosa (2023) who distinguish between 'moral patiency' (the capacity to be wronged), ‘moral responsibility' (the capacity to be held accountable) and ‘moral agency' (the capacity to act for moral reasons). They believe that existing AI systems can negatively impact moral patients, but that they are not themselves moral agents, so human leaders need to take a bigger responsibility for the systems they use. Third, the literature is fraught with conflicting conceptions of how far one can go in designing ethics into an AI rather than representing them. The Responsible Autonomy of AI, The ethics of autonomous AI systems has always been a factor of the norms of human designers and the organizations within which they are applied (Dignum, 2020). This 'value-laden' perspective on AI has shifted the moral burden onto organizational leaders to not only be reactive but also be proactive in setting the moral boundaries of AI systems (Taddeo and Floridi 2018; Vakkuri et al. 2020).

4.2 Theme 2: Algorithmic Bias as an Ethical Leadership Challenge

Data politics and the ethics of AI are intertwined. These machine learning models are trained based on historical data reflecting previous discrimination and, when used within an organizational context, can introduce and reinforce discrimination at scale (O’Neil 2016; Obermeyer et al. 2019). This dynamic is well known in the literature in several areas of the organization. In the context of hiring and talent management, Dastin (2018) and other academic studies (Pena, Iñaki, and Pablo et al. 2020; Köchling and Wehner 2020) have reported consistent discrimination by women and ethnic minorities against AI-powered recruitment systems that were trained using historical hiring data that may also have been historically discriminatory. Binns (2018) argues that this can be a unique ethical challenge for algorithmic leaders: these outputs are often perceived as objective and widely applicable, which makes it difficult to uncover and question bias. In performance management, Ajunwa (2020) and Duggan et al. (2020) report the development of ‘algorithmic management’ (the monitoring, evaluation and disciplining of employees with the help of algorithms based on machine learning), which they describe as a type of ‘surveillance capitalism’ that affects the dignity and autonomy of workers. Duggan

et al. (2020) suggest that from an ethical leadership standpoint, leaders who use algorithmic performance measures without considering their ethical implications are thereby delegating the responsibility of caring for employees to the algorithms. In each of these settings, the literature highlights a common theme: ethical leadership failure is not simply technical, involving using a biased algorithm, but organizational—how to set up governance processes that can identify and address algorithmic bias. (Raji Elizabeth and Aaron et al. 2022) conducted a survey of internal audit practices at the organizations that deploy AI and discovered that less than 30 percent of them had formal procedures in place to detect bias in AI, and even fewer had board-level oversight on the use of AI for ethics. The literature proposes that proactive institutional design – the intentional development of institutional structures, processes, and cultures that intentionally address algorithmic bias by making it an organizational priority – is key to responsible AI leadership (Canca 2022; Metcalf, Keller and boyd 2016).

4.3 Theme 3: Transparency, Explainability, and Leadership Trust

Ethical leadership is about trust. The basic model proposed by Brown, Treviño, and Harrison (2005) defines trustworthiness as the foundation of ethical leadership influence and is defined by the extent to which followers are willing to engage in ethical behavior and are willing to report ethical concerns in the organization (Mayer, Davis, and Schoorman 1995; Walumbwa et al. 2008). The 'black box' problem is a unique risk to trust that arises with the deployment of AI. The processes of many high performing AI systems, especially deep learning, are not interpretable even by the designer of the system (Rudin 2019). A job applicant may not be able to offer the kind of transparent explanation that Treviño, Hartman, and Brown (2000) suggest is at the core of ethical leadership if he or she has been screened out of the job, and a medical diagnosis may result from an opaque algorithmic process in which the leader who relies on the result cannot provide such an explanation. Failure to provide a reason for a decision is not just a communication problem, but an ethics issue because it is the violation of the rights of affected parties to understand and question consequential decisions that involve them (Wachter, Mittelstadt, and Russell 2017). There are multiple dimensions of AI transparency that link directly to leadership, as discussed in the literature. The public disclosure of information about the functioning of an AI system is called algorithmic transparency, the disclosure of information about how the AI system is used in decision making processes is called process transparency, and the disclosure of information about the decisions that the AI system produces is called outcome transparency (Larsson and Heintz 2020). Although ethical leadership is expected to be accountable for all three dimensions, the literature indicates that outcome transparency might be especially sensitive in situations where complete algorithmic transparency is not technically feasible or commercially sensitive, which might result in less trust among stakeholders (Doshi-Velez and Kim 2017; Mittelstadt, Patrick and Sandra 2022). New research has started to provide AI decision-making leaders with tools for communicating AI decisions in human-understandable terms, including ‘explainable AI’ (XAI).Research in the field of ‘explainable AI’ (XAI) is beginning to provide AI decision-making leaders with technical tools that can help them communicate AI decisions in human-readable terms, such as LIME, SHAP and counterfactual explanations (Samek, Wiegand, and Müller 2017; Molnar 2020). But (Rudin 2019) warns that post-hoc explanations of black-box models are always going to be an imperfect approximation and recommends the use of inherently interpretable models wherever possible. This may mean taking a more proactive role in shaping the technical design decisions taken upstream, rather than simply dealing with the communication challenges downstream, for ethical leaders (Bankins and Formosa 2023).

4.4 Theme 4: Data Governance and Privacy as Leadership Imperatives

The AI systems leaders implement are, by nature, systems for data collection, analysis and use. Ethical leadership in the AI age is thus part of data governance, which encompasses the policies, processes and institutional structures that govern the collection, storage, sharing and use of data within the organization (Janssen, Brous and Estevez 2020). Data governance, as identified in the literature reviewed here, is a neglected aspect of ethical leadership practice. The term privacy stands tall when it comes to data ethics issues in the corpus. A theoretical framework for some of these studies is the influential idea of ‘surveillance capitalism' (Zuboff 2019) which concerns the commodification of human behavioral data for profit that implicates employees, customers and citizens without them having given meaningful consent. A set of questions arises for ethical leaders: What information does the organization need to gather? From whom? For what purposes? When is it appropriate to give data to third parties or government officials? Data governance has been transformed from a compliance role to a strategic leadership role as a result of the General Data Protection Regulation (GDPR) and similar privacy policies in countries such as the United States, Canada, China, and Brazil (Voigt and von dem Bussche 2017; Hoofnagle, van der Sloot and Borgesius 2019). There are many studies in our corpus that show a 'privacy leadership gap': organisations often have more technical staff with greater knowledge and engagement with data ethics than do their senior leaders, leading to a culture in which it is considered as a topic for the IT department rather than a leadership issue (Martin and Shilton 2016; Spiekermann-Holt 2022). In addition to regulatory requirements, there is a more positive approach to data stewardship as part of the concept of ethical leadership. The literature also includes a more enthusiastic view of data stewardship as a leadership responsibility that is ethical. Janssen et al. (2020) suggest a ‘data governance maturity model', which separates two categories: reactive (compliance driven) and proactive (values driven) approaches to data ethics. Leaders following the latter approach integrate privacy concerns into product design, procurement and organizational culture, and don't view data ethics as a constraint to be managed. This proactive orientation is in line with Treviño et al's (2000) definition of the ‘moral person' aspect of ethical leadership, which refers to internalization of values that drive conduct without the presence of external oversight.

4.5 Theme 5: AI Ethical Leadership Competencies

The fifth group of literature deals with the special skills required for the ethical governance of AI. The field of this work is interdisciplinary, combining adult education, leadership development and human-computer interaction research, and is applied. According to Canca, there are four key competencies that are essential in an ethical leader of AI: (1) AI literacy (understanding, at a conceptual level, how AI systems work and what they can and cannot do); (2) ethical reasoning (identifying and analyzing the ethical implications of AI deployment decisions); (3) stakeholder engagement (soliciting and incorporating diverse perspectives on AI governance); and (4) institutional design (establishing organizational structures, policies, and cultures to realize ethical AI principles). The framework is supported in many respects by the research of (Mittelstadt, Patrick and Sandara 2022) and (Floridi, Josh and Thomas 2021), which further emphasize the importance of systems thinking (the capacity to foresee second and third-order

effects of the implementation of AI) and psychological safety (the creation of environments that allow employees to voice their concerns regarding the ethics of AI). A second key challenge identified relates to the technical expertise needed for an effective and responsible governance of AI and the generalist nature of senior leadership positions. Simple mental models of the AI, as a neutral tool or a superintelligent agent, are common among senior executives (Dignum 2020; Rinta-Kahila, Tapani, Ida, and Nicole et al. 2023), both of which hinder good governance of ethical issues. This will entail more than just the technical upgrading of individual leaders; it will require organizational structures that connect technical knowledge with effective decision-making power, including Chief Ethics Officer positions, AI Ethics Boards, and cross-functional ethics review committees (Canca 2022; Raji, Elizabeth and Aaron et al. 2022). Additionally, the gender and intersectionality aspects of ethical competency in AI leadership are starting to be acknowledged. (Criado-Perez 2019) and (D'Ignazio and Klein 2020) describe the tendency for AI teams to be composed mainly of men and white people, which helps create a system of blind spots around the differential effects of AI on different populations. There are a number of studies in our corpus that suggest that a diverse leadership team, both in terms of gender and ethnicity, is better equipped to recognize and reduce inadvertent harms of AI to underrepresented populations, although the evidence for that remains limited: (Criado-Perez 2019; West, Meredith and Kate 2019).

5. The Responsible AI Leadership Model (RAILM)

5.1 Framework Overview

To derive from our thematic synthesis, we put forth an integrative theoretical framework called the Responsible AI Leadership Model (RAILM), which re-thinks the concept of ethical leadership in an AI-infused organizational context. RAILM is a framework which expands the classical tradition of ethical leadership in three dimensions: an expanded moral agency, a systemic approach to accountability and proactive institutional design.

5.2 Core Propositions

The organizing principles of RAILM are five propositions that are interrelated:

Proposition 1: When designing, deploying, and facing outcomes of AI systems, ethical leaders in AI-supported organizations are accountable for having a heightened moral agency, even in situations where they do not have a deeper understanding of the systems themselves. (Bankins and Formosa 2023; Rinta-Kahila, Tapani, Ida, and Nicole et al. 2023). This proposition questions the 'moral luck' defense which has been advanced by leaders in cases where documented harm has occurred, that is, because the algorithmic system was left to its own devices, the leader is not accountable. RAILM believes that leaders who opt for the deployment of AI systems are ethically responsible for the probable outcomes of their actions.

Proposition 2: Bias detection, fairness auditing, and recourse mechanisms for stakeholders need to be embedded in organizational practice to guarantee ethical AI leadership (Raji Elizabeth and Aaron et al. 2022; Canca 2022). Algorithmic harm at the organizational level is not solely the fault of individual leader actions; ethical leadership should be reflected in designing organizations to accommodate distributed ethical evaluation of AI systems.

Proposition 3: Leadership trust in AI-infused contexts relies on a sense of transparency of the results and a meaningful explanation of the results to stakeholders who may be impacted (Wachter, Mittelstadt, and Russell 2017; Mittelstadt, Patrick and Sandra 2022). If leaders fail to or refuse to elucidate consequential decisions made by AI, they are compromising the relational aspect of ethical leadership influence.

Proposition 4: Data stewardship, which is deliberate responsible governance of data collection, use and sharing based on values, is a fundamental duty of ethical leadership in the AI age (Janssen, Brous and Estevez 2020; Spiekermann-Holt 2022). The way that data governance is viewed by the leaders as merely a technical or compliance role is a neglect of an important part of an ethical role.

Proposition 5: The development of AI-specific ethical leadership skills, such as AI literacy, ethical reasoning, stakeholder engagement, systems thinking and institutional design capability, is a leadership development priority and an organizational responsibility (Canca 2022; Floridi, Josh and Thomas et al. 2021). If these competencies are not developed by leaders, then structure paves the way for moral failure in the organization.

5.3 RAILM and Established Leadership Theories

RAILM does not replace existing theories of ethical leadership but builds upon them to apply them to the context of AI. It is a focus on increased moral responsibility, which is rooted in the fiduciary perspective on the welfare of the stakeholders that is inherent in servant leadership (van Dierendonck 2011). It is related to the structural aspect of transformational leadership (Bass and Riggio 2006), which is about institutional design. It builds on the relational transparency aspect of authentic leadership (Walumbwa et al. 2008) by emphasizing transparency and explainability. It also shares an underlying premise with the moral person conception of ethical leadership (Treviño, Hartman, and Brown 2000) in that it emphasizes the internalization of values in a proactive manner.

6. Discussion

6.1 Implications for Research

The review highlights a number of key research avenues for future studies. First, the literature is far more conceptual and theoretical than empirical, and a lot of empirical research, quantitative and qualitative, is needed to test the propositions that have been outlined here. Longitudinal research that tracks the development of organizations' AI governance strategies in response to ethical missteps, regulatory requirements, and stakeholder participation would be especially valuable to the field. Second, although there are a number of studies that discuss the ethics of AI in North American or Western European situations, there is a significant lack of research on AI ethical leadership in the Global South where AI is being rapidly utilized and regulatory frameworks are often still in their infancy (Jobin, Ienca, and Vayena 2019). The propositions of RAILM could have significant meaning in different cultural contexts, with implications for their translation. Third, the gender and intersectional aspects of ethical leadership in AI require far more continued study. The majority of the research reviewed in this paper approaches the study of organizational leaders

as a general category, and future studies could investigate how gender, race, class, and other social roles impact leaders' responses to issues of AI ethics (Criado-Perez 2019; D'Ignazio and Klein 2020). Fourthly, there is a need for the development of measurements. RAILM's competency framework calls for validated tools that can measure ethical leadership in relation to and within AI in workplaces. Scale development research, based on the tradition in measuring ethical leadership that has been developed over the years (Brown, Treviño and Harrison 2005), would be a valuable contribution.

6.2 Implications for Practice

There are several implications for organizational leadership and governance professionals that can be drawn from RAILM. Most fundamentally, it transforms AI governance from a technical compliance role to an ethical leadership role. If leaders are able to adopt this mindset, they will make decisions about the use of AI with the same moral rigor that they show when making other decisions that have a high moral component. In practice, this entails: doing proactive ethical impact assessments before introducing AI systems in sensitive environments; establishing cross-functional AI ethics review processes with real power to stop or alter deployments; implementing AI literacy development programs for senior executives and board members; creating safe channels for employees to voice concerns about AI ethics without incurring professional risks; and publishing clear reports of AI governance practices to external stakeholders (Canca 2022; Floridi et al. 2021). To boards of directors, RAILM recommends that AI governance, like financial and legal compliance, be a permanent agenda item and have dedicated oversight mechanisms. Leading organizations are introducing special positions like Chief AI Ethics Officer and Responsible AI team, providing a framework for others to adapt (Raji Elizabeth and Aaron et al. 2022).

6.3 Implications for Policy

This review has policy implications for those tasked with creating frameworks for the governance of AI by the legislature and regulators. Current AI regulations, such as the EU AI Act suggested by the European Commission (2021) and the US National AI Initiative Act, are mainly technical and organizational in nature, covering the methods and processes of AI development. The leadership aspect is only peripherally addressed. It recommends that laws and regulations explicitly mandate executive accountability in AI governance, including: disclosure of executive decisions around AI deployment; requirement for board-level oversight of high-risk AI applications; liability provisions for executives who fail to discharge their duty of care properly in relation to AI governance. This would result in institutionalized incentives towards building ethical leadership competencies and taking proactive governance action in the field of AI.

6.4 Limitations

There are some limitations to this review which should be noted. Although we conducted thorough searches using the PRISMA guidelines, the search might have missed some relevant publications in non-English languages or some publications in journals not indexed in the databases we searched. Although it is correct that there are only scholarly articles in peer-reviewed journals, practitioner literature and policy documents could have relevant insights into the practice of ethical leadership in AI. Further, the interdisciplinary nature of the topic involved synthesizing different programs of research with ontological commitments and methodological approaches. It is necessary to simplify and generalize complex and contested debates in the theoretical framework that we propose, RAILM. Future researchers are invited to test its propositions empirically and to elaborate its theory.

7. Conclusion

This systematic literature review has identified 87 peer-reviewed publications and analyzed the current research landscape around the issue of ethical leadership and artificial intelligence. We identified five overarching thematic clusters that we believe are key and have included in the Responsible AI Leadership Model: moral agency and distributed responsibility, algorithmic bias, transparency and trust, data governance and privacy, and AI ethical leadership competencies. At the core of RAILM is the hope that ethical leadership in the AI era demands much more than was ever expected of leaders before. As AI systems become impactful organizational actors, human leaders still have the same ethical duties as before, and they are only becoming greater. When leaders deploy AI systems, they enter into a morally significant responsibility for the values of those systems and the harm they inflict. This broader understanding of ethical leadership is not just a theoretical construct, but it also reflects the growing demands of regulators, employees, customers, and civil society organizations for organizational leadership to ensure the ethical use of AI systems. Those who are best suited to lead this charge will be those who possess not only a missionary spirit of ethical concern, but also have the institution design skills to turn their moral concern into organizational structures and processes in which they systematically guard against harms arising from AI. With each new advancement in AI capabilities, the importance of ethical AI leadership only intensifies, moving towards greater autonomy, widespread deployment, and integration into organizational life. This review is a welcome addition to a discussion that is coming to a head in the twenty-first century as a defining issue of management scholarship and practice.

References

Acquisti, Alessandro, Laura Brandimarte, and George Loewenstein. 2015. Privacy and Human Behavior in the Age of Information. *Science* 347: 509–514. https://doi.org/10.1126/science.aaa1465

Ajunwa, Ifeoma. 2020. The Paradox of Automation as Anti-Bias Intervention. *Cardozo Law Review* 41: 1671–1742. https://scholarship.law.unc.edu/cgi/viewcontent.cgi?article=1490&context=faculty_publications&ref=internet.exchangepoint.tech

Avolio, Bruce J., Fred O. Walumbwa, and Todd J. Weber. 2009. Leadership: Current Theories, Research, and Future Directions. *Annual Review of Psychology* 60: 421–449. https://doi.org/10.1146/annurev.psych.60.110707.163621

Bankins, Sarah, and Paul Formosa. 2023. The Ethical Implications of Artificial Intelligence (AI) for Meaningful Work. *Journal of Business Ethics* 185: 725–740. https://doi.org/10.1007/s10551-023-05339-5

Bass, Bernard M., and Ronald E. Riggio. 2006. Transformational Leadership. *2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates*. *Psychology Press*: 1-296. https://www.taylorfrancis.com/books/mono/10.4324/9781410617095/transformational-leadership-bernard-bass-ronald-riggio

Bélanger, France, and Robert E. Crossler. 2011. Privacy in the Digital Age: A Review of Information Privacy Research in Information Systems. *MIS Quarterly* 35: 1017–1042. https://doi.org/10.2307/41409971

Binns, Reuben. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. *Proceedings of Machine Learning Research* 81:149–159. https://proceedings.mlr.press/v81/binns18a.html

Brown, Michael E., Linda K. Treviño, and David A. Harrison. 2005. Ethical Leadership: A Social Learning Perspective for Construct Development and Testing. *Organizational Behavior and Human Decision Processes* 97: 117–134. https://doi.org/10.1016/j.obhdp.2005.03.002

Canca, Cansu. 2022. Operationalizing AI Ethics Principles. Communications of the ACM 63 (12): 18–21. https://doi.org/10.1145/3430368

Criado-Perez, Caroline. 2019. Invisible Women: Data Bias in a World Designed for Men. New York: Abrams Press. https://www.goodreads.com/en/book/show/41104077-invisible-women

Dastin, Jeffrey. 2022. Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women. *Auerbach*: 1-4. https://www.taylorfrancis.com/chapters/edit/10.1201/9781003278290-44/amazon-scraps-secret-ai-recruiting-tool-showed-bias-women-jeffrey-dastin

Dierendonck, Dirk van. 2011. Servant Leadership: A Review and Synthesis. *Journal of Management* 37: 1228–1261. https://doi.org/10.1177/0149206310380462

D'Ignazio, Catherine, and Lauren F. Klein. 2020. Data Feminism. *Cambridge, MIT Press*. https://data-feminism.mitpress.mit.edu/

Dignum, Virginia. 2019. Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Cham: Springer. https://doi.org/10.1007/978-3-030-30371-6

Doshi-Velez, Finale, and Been Kim. 2017. Towards a Rigorous Science of Interpretable Machine Learning. *arXiv* arXiv:1702.08608. https://arxiv.org/abs/1702.08608

Duggan, James, Ultan Sherman, Ronan Carbery, and Anthony McDonnell. 2020. Algorithmic Management and App-Work in the Gig Economy: A Research Agenda for Employment Relations and HRM. *Human Resource Management Journal* 30: 114–132. https://doi.org/10.1111/1748-8583.12258

European Commission. 2019. Ethics Guidelines for Trustworthy AI. *European Commission.* https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Commission. 2021. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). COM(2021) 206 final.

Fjeld, Jessica, Nele Achten, and Hannah Hilligoss et al. 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. *SSRN*: 1-39. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3518482

Floridi, Luciano, Josh Cowls and Thomas C. King et al. 2021. How to Design AI for Social Good: Seven Essential Factors. *Science and Engineering Ethics* 26: 1771–1796. https://doi.org/10.1007/s11948-020-00213-5

Greenleaf, Robert K. 1977. Servant Leadership: A Journey into the Nature of Legitimate Power and Greatness. *Paulist*: 1-370. https://www.amazon.com/Servant-Leadership-Legitimate-Greatness-Anniversary/dp/0809105543

Hoofnagle, Chris Jay, Bart van der Sloot, and Frederik Zuiderveen Borgesius. 2019. The European Union General Data Protection Regulation: What It Is and What It Means. *Information and Communications Technology Law* 28: 65–98. https://doi.org/10.1080/13600834.2019.1573501

Janssen, Marijn, Paul Brous, and Elsa Estevez et al. 2020. Data Governance: Organizing Data for Trustworthy Artificial Intelligence. *Government Information Quarterly* 37: 101493. https://doi.org/10.1016/j.giq.2020.101493

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence* 1: 389–399. https://doi.org/10.1038/s42256-019-0088-2

Köchling, Alina, and Marius Claus Wehner. 2020. Discriminated by an Algorithm: A Systematic Review of Discrimination and Fairness by Algorithmic Decision-Making in the Context of HR Recruitment and HR Development. *Business Research* 13: 795–848. https://doi.org/10.1007/s40685-020-00134-w

Larsson, Stefan, and Fredrik Heintz. 2020. Transparency in artificial intelligence. *Internet Policy Review*. https://doi.org/10.14763/2020.2.1469

Luthans, Fred, and Bruce J. Avolio. 2003. Authentic Leadership: A Positive Developmental Approach. *Edited by Kim S. Cameron, Jane E. Dutton, and Robert E. Quinn*: 241–261. https://cerf.radiologie.fr/sites/cerf.radiologie.fr/files/Enseignement/DES/Modules-Base/Luthans%20%26%20Avolio%2C%202003.pdf

Martin, Kirsten, and Katie Shilton. 2016. Why Experience Guides Ethics: The Practice of Ethics for Engineers in ICT. *Business and Society* 55: 521–549. https://doi.org/10.1177/0007650313517166

Mayer, Roger C., James H. Davis, and F. David Schoorman. 1995. An Integrative Model of Organizational Trust. *Academy of Management Review* 20: 709–734. https://doi.org/10.2307/258792

Metcalf, Jacob, Emily F. Keller, and Danah Boyd. 2016. Perspectives on Big Data, Ethics, and Society. *Council for Big Data, Ethics, and Society*: 1-23. https://bdes.datasociety.net/council-output/perspectives-on-big-data-ethics-and-society/

Mittelstadt, Brent, Patrick Russell, and Sandra Wachter. 2022. Explaining Explanations in AI. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*: 279–288. https://doi.org/10.1145/3287560.3287574

Mittelstadt, Brent. 2019. Principles Alone Cannot Guarantee Ethical AI. *Nature Machine Intelligence* 1: 501–507. https://doi.org/10.1038/s42256-019-0114-4

Molnar, Christoph. 2020. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. *Raleigh*: 1-251. https://christophm.github.io/interpretable-ml-book/

Nishii, Lisa H., and Patrick M. Wright. 2008. Variability Within Organizations: Implications for Strategic Human Resources Management. *The People Make The Place, edited by Daniel B. Smith. Psychology Press*: 225–248. https://www.taylorfrancis.com/chapters/edit/10.4324/9780203809549-14/variability-within-organizations-implications-strategic-human-resources-management-lisa-nishii-patrick-wright

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science* 366: 447–453. https://doi.org/10.1126/science.aax2342

OECD. 2019. Recommendation of the Council on Artificial Intelligence. *OECD LEGAL Instruments*: 1-272. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

O'Neil, Cathy. 2016. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. *Crown*. https://dl.acm.org/doi/10.5555/3002861

Page, Matthew J., Joanne E. McKenzie and Patrick M. Bossuyt et al. 2021. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* 372: https://doi.org/10.1136/bmj.n71

Pena, Axel, Iñaki Pérez-Arnal, and Pablo García-Mochales et al. 2020. Bias in Multimodal AI: Testbed for Face Image-Based Gender Classification. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 30–31.

Petticrew, Mark, and Helen Roberts. 2006. Systematic Reviews in the Social Sciences: A Practical Guide. *Malden*. https://onlinelibrary.wiley.com/doi/book/10.1002/9780470754887

Raji, Inioluwa Deborah, Elizabeth Kumar, and Aaron Horowitz et al. 2022. The Fallacy of AI Functionality. *ACM Conference on Fairness, Accountability, and Transparency*: 959–972. https://doi.org/10.1145/3531146.3533158

Rinta-Kahila, Tapani, Ida Someh, and Nicole Gillespie et al. 2023. Algorithmic Decision-Making and the Problem of Accountability. *MIS Quarterly* 46: 2357–2400. https://doi.org/10.25300/MISQ/2022/16535

Rudin, Cynthia. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* 1: 206–215. https://doi.org/10.1038/s42256-019-0048-x

Samek, Wojciech, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable Artificial Intelligence: Interpreting Explaining and Visualizing Deep Learning. *ITU Journal* 1: 39–48. https://link.springer.com/book/10.1007/978-3-030-28954-6

Spiekermann-Holt, Sarah. 2022. Ethical IT Innovation: A Value-Based System Design Approach. *Routledge*. https://www.routledge.com/Ethical-IT-Innovation-A-Value-Based-System-Design-Approach/Spiekermann/p/book/9781482226355

Taddeo, Mariarosaria, and Luciano Floridi. 2018. The Debate on the Moral Responsibilities of Online Service Providers. *Science and Engineering Ethics* 22: 1575–1603. https://doi.org/10.1007/s11948-016-9754-x

Thomas, James, and Angela Harden. 2008. Methods for the Thematic Synthesis of Qualitative Research in Systematic Reviews. *BMC Medical Research Methodology* 8: 45. https://doi.org/10.1186/1471-2288-8-45

Tranfield, David, David Denyer, and Palminder Smart. 2003. Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review. *British Journal of Management* 14: 207–222. https://doi.org/10.1111/1467-8551.00375

Treviño, Linda Klebe, Laura Pincus Hartman, and Michael Brown. 2000. Moral Person and Moral Manager: How Executives Develop a Reputation for Ethical Leadership. *California Management Review* 42: 128–142. https://doi.org/10.2307/41166057

Vakkuri, Ville, Kai-Kristian Kemell, and Marianna Jantune et al. 2020. ECCOLA: A Method for Implementing Ethically Aligned AI Systems. *Journal of Systems and Software* 182: 1-16. https://www.sciencedirect.com/science/article/pii/S0164121221001643

Voigt, Paul, and Axel von dem Bussche et al. 2017. The EU General Data Protection Regulation (GDPR): A Practical Guide. *Cham.* https://doi.org/10.1007/978-3-319-57959-7

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law and Technology* 31: 841–887. https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf

Waldman, David A., and Robert M. Balven. 2014. Responsible Leadership: Theoretical Issues and Research Agenda. *Academy of Management Perspectives* 28: 224–234. https://doi.org/10.5465/amp.2014.0016

Walumbwa, Fred O., Bruce J. Avolio, and William L. Gardner et al. 2008. Authentic Leadership: Development and Validation of a Theory-Based Measure. *Journal of Management* 34: 89–126. https://doi.org/10.1177/0149206307308913

West, Sarah Myers, Meredith Whittaker, and Kate Crawford. 2019. Discriminating Systems: Gender, Race, and Power in AI. *AI Now Institute*. https://philpapers.org/rec/INSDSG

Zuboff, Shoshana. 2019. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. *Havard Business School*. https://www.hbs.edu/faculty/Pages/item.aspx?num=56791