

Article

Introducing SAFE-AI: A Behavioral Framework for Managing Ethical Dilemmas in AI-Driven Human Resource Practices

Rob E. Carpenter ^{1,*}, Debaro Huyler ², Sanket Ramchandra Patole ¹  and Rochell McWhorter ¹ 

¹ Department of Human Resource Development, University of Texas at Tyler, 3900 University Boulevard, Tyler, TX 75799, USA

² Division of Academic and Student Affairs, Florida International University, 11200 SW 8th St., Miami, FL 33199, USA

* Correspondence: rcarpenter@uttyler.edu

Abstract

Organizations increasingly deploy artificial intelligence (AI) in human resource (HR) decision processes to improve efficiency and strategic execution, yet ethical failures persist when principles remain decoupled from everyday workflow enactment. This paper addresses AI-ethics in HR practice by advancing a behavior-first premise: AI-ethics becomes durable organizational practice only when ethical intent is translated into observable routines and cues that employees can interpret as legitimate and consistently enforced. We introduce the *Socially Aware Framework for Ethical AI* (SAFE-AI), which integrates normative ethical reasoning (consequentialist and deontological logics), social information processing, and socially informed heuristics as a practical translation layer for HR workflows. SAFE-AI specifies three stages of implementation—*moving in* (initiation), *moving through* (navigation), and *moving out* (culmination)—to guide scoping and constraints, feedback-driven interpretation management, and institutionalized accountability. Because enactment depends on the organizational cue environment, leadership behaviors (ethical intent-setting, resourcing, sensegiving transparency, and enforceable accountability) function as necessary conditions for sustained implementation beyond HR-local governance. We conclude with implications for practice and a testable agenda for research focused on implementation fidelity, cue-consistency mechanisms, and boundary conditions across organizational contexts.

Keywords: SAFE-AI; artificial intelligence; human resources; heuristics; social information processing; AI-ethics

“AI [Artificial Intelligence] will be the most transformative technology of the 21st century. It will affect every industry and aspect of our lives.” Jensen Huang, CEO at NVIDIA



Received: 29 December 2025

Revised: 2 February 2026

Accepted: 3 February 2026

Published: 9 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Organizations are leveraging artificial intelligence (AI) technology to gain unique competitive advantages (Krakowski et al., 2023). However, poor management during AI adoption can lead organizations to overlook critical ethical principles, potentially compromising stakeholder interests. Ethical principles in AI—broadly referred to here as AI-ethics—are increasingly recognized as a global concern, driving the need to ethically adopt technologies that empower computers with human-like cognitive abilities (Jobin et al., 2019). AI-ethics involves applying ethical principles to the development, deployment, and use of AI technologies, encompassing the identification, analysis, and resolution of moral issues arising from the interface of AI systems and human behavior. This includes

ensuring fairness, accountability, transparency, privacy, and respect for human rights throughout the AI lifecycle. Furthermore, AI-ethics aims to mitigate risks, prevent harm, and promote beneficial outcomes for individuals and society (Whittlestone & Clarke, 2022).

Human Resource Development (HRD) professionals play an important role in ensuring ethical AI adoption through targeted training, development (employee and leadership), and organizational interventions. By integrating AI governance with HRD practice, the organization protects stakeholder interests and sustains competitive advantage through the deliberate development of human and technological capital. Responsible AI requires more than codifying ethical principles; it also requires critical scrutiny of how those principles are produced, which methods and assumptions inform them, and what societal impacts they enable or constrain (Stilgoe et al., 2020). Such scrutiny foregrounds the normative and behavioral foundations of organizational ethics, including values, justice-oriented beliefs, and the interpretive processes through which ethical commitments become enacted. Recent research has advanced organizational conceptualization of AI-ethics (e.g., Ortega-Bolaños et al., 2024), but existing contributions remain limited in specifying behaviorally grounded mechanisms that embed ethical AI as a stable organizational capability rather than a set of aspirational statements (M. Jones et al., 2022). And this gap is particularly crucial for organizations leveraging AI technology and HRD simultaneously, because competitive advantage depends on whether ethical commitments can be operationalized through consistent decision routines, learning systems, and accountability structures (Chowdhury et al., 2023).

1.1. A Behavior-First Approach for Embedding AI-Ethics

An alternative approach suggests that organizations should focus on maturing the underlying behaviors that enable ethical conduct, alongside the formal processes used to govern AI adoption (Calabretta et al., 2017; Thiel et al., 2012). This perspective assumes that successful enactments of AI-ethics share recurrent behavioral patterns, reflected in how people notice, interpret, and respond to ethically salient cues during design, deployment, and use (Thibault, 2004). In organizational settings, these patterns are shaped by group conditions (e.g., norms, psychological safety, accountability) and depend on baseline social awareness that supports behavioral reasoning about others' interests and likely reactions. Embedding AI-ethics, therefore, requires translating governance requirements (e.g., fairness, accountability, transparency) into observable behavioral routines and aligning those routines with the cognitive processes through which stakeholders interpret social information (Welch & Dixon, 1994). This behavior-first view motivates a framework for integrating technological artifacts, human ethics, and human resource (HR) practices.

The purpose of this paper is twofold: (1) to develop a behavioral framework that clarifies the relationship between AI technology and human ethics for HR practices and (2) to specify how HR practices can operationalize AI-ethics through actional mechanisms. We adopt a behavioral framework because it is a structured approach for understanding, predicting, and shaping human behavior using principles from behavioral science (Brase, 2014). Behavioral strategy research in management further supports the value of behavioral mechanisms for explaining organizational outcomes (e.g., Hesselbarth et al., 2023). Related domains also show that awareness-oriented interventions can shift behavior in high-risk socio-technical contexts, including cybersecurity (Yunos et al., 2016), ransomware mitigation (Han et al., 2017), and social engineering (Aldawood et al., 2020). Building on this evidence, we propose an awareness strategy termed *Socially Aware Framework for Ethical AI* (SAFE-AI). SAFE-AI specifies how HR practices can support AI-ethics by leveraging ethical philosophies, social theories, and bounded cognitive strategies (heuristics) to the behaviors and routines through which AI-ethics is enacted in organizational ethos.

1.2. Research Gap and Contributions

Extant research on AI technology in HR practice provides a strong descriptive account of ethical risk, repeatedly converging on discriminatory outcomes in selection and evaluation and on limited transparency and explainability. It also highlights privacy and surveillance exposure from expansive data practices and diffuse accountability when algorithmic outputs are embedded in HR decision routines (Rodgers et al., 2023; Tursunbayeva et al., 2018). What remains comparatively underdeveloped is an implementation-level account of how ethical principles are translated into repeatable, teachable, and auditable practices inside everyday HR workflows (e.g., screening, appraisal, discipline, employee relations), particularly under real-world constraints such as time pressure, delegation, and cross-functional handoffs (Rodgers et al., 2023; Tursunbayeva et al., 2018). In parallel, many AI-ethics approaches emphasize technical safeguards and governance prescriptions yet offer limited specification of the behavioral pathway through which ethical standards become socially recognized, interpreted, and enacted as “the way we do things here” across managerial layers and peer communities (Salancik & Pfeffer, 1978). Because workplace attitudes and conduct are shaped more by salient cues, local norms, and collectively reinforced interpretations than by abstract principles alone (Salancik & Pfeffer, 1978), we conceptualize AI-ethics as an organizational accomplishment that must be behaviorally embedded.

Accordingly, this study contributes the SAFE-AI framework, a behavior-first integration model that links (1) normative ethical reasoning (consequentialist and deontological logics), (2) social information processing mechanisms that explain how ethical expectations are interpreted and legitimized in context, and (3) socially informed heuristics that function as an actionable translation layer for AI-enabled HR decisions. SAFE-AI further specifies a set of cognitive milestones (innovation, navigation, culmination) through which ethical expectations can be stabilized as routine practice during AI adoption and use. To ground the framework in consequentially relevant failure modes, we use two widely documented exemplars: Amazon’s recruiting-system bias case as a selection discrimination risk and a Microsoft large-scale data exposure incident as a privacy and trust failure, demonstrating how technical and data practices propagate into organizational legitimacy risk.

In the following, we first examine the evolving impact of AI technology on HR practices. Next, we bring AI-ethics into focus by presenting two case studies that highlight the ethical challenges associated with adopting and executing AI technologies. We then consider ethical philosophies to establish an ethical foundation for navigating the complexities of AI-ethics. From here, we link ethical philosophies to a social interpretation of AI-ethics through the lens of social information processing theory. After providing a brief argument for the use of heuristics, we then propose a behavioral framework for guiding HR practices on incorporating AI-ethics into organizations. We follow up with implications for research and practice and conclude with limitations and considerations for future research.

2. AI in HR Practice and the Ethics-to-Implementation Problem

As AI technologies increasingly mediate organizational decisions, it becomes essential to understand both the trajectory of AI development and the ethical implications of its adoption in HR practice. AI refers to a broad set of computational approaches that enable systems to perform tasks typically associated with human cognition, including adaptive decision-making and pattern recognition (Abiodun et al., 2019; Oliveira et al., 2024). Historically, AI emerged in the 1950s and initially emphasized rule-based symbolic problem solving, reflecting the premise that intelligent behavior could be produced through formal representations and rule-governed manipulation (Augusto, 2021; Tambe et al., 2019). Contemporary AI, however, is increasingly defined by data-driven approaches, including

machine learning and neural networks, which have expanded AI capability in areas such as language processing, prediction, and classification (Abiodun et al., 2019).

In workplace settings, AI adoption has progressed from automating routine tasks to shaping consequential judgments about people. Since the early 2000s, AI-enabled tools have been used to evaluate candidate communication signals, predict performance, infer affective states, and provide responsive interactions, and these capabilities are now embedded across a widening range of HR processes (Danysz et al., 2019; Dastin, 2018; Scherer, 2015; Tambe et al., 2019). Although integrating human and machine intelligence can improve decision quality and task performance, it also introduces substantive ethical dilemmas, particularly where AI systems influence opportunity, evaluation, and workplace standing (Ardichvili, 2022; Francolini et al., 2020; McWhorter, 2023; Mohammed, 2019; Whittlestone & Clarke, 2022; Yorks et al., 2020).

A common assumption is that established tenets of human ethics can inform AI use in organizations, yet the literature provides comparatively limited specification of how ethical principles are translated into stable, socially enacted practices in AI-mediated work (Hendrycks et al., 2021; Prikshtat et al., 2023; Shneiderman, 2020). Two challenges are especially salient. First, the opacity and complexity of many AI systems complicate the practical alignment of decision outputs with moral reasoning, accountability, and socially legitimate justification. Second, human ethics is inherently context-dependent, drawing on cultural, social, and individual values that do not map cleanly onto technical implementations, which increases the risk of unintended organizational consequences even when systems appear to replicate human-like competencies (Banks, 2020; Chandra et al., 2022).

These challenges are amplified in HR practice, where AI adoption is accelerating but remains unevenly integrated across HR policies, processes, and research (Ardichvili, 2022). Organizations increasingly expect HR functions to leverage predictive and prescriptive analytics, which raises the stakes for fairness, transparency, privacy, and trust (Chowdhury et al., 2023). Human-centric AI scholarship therefore emphasizes the need for reliable, safe, and trustworthy systems supported by ethical standards, yet such standards frequently collide with value-laden concerns that generate practical dilemmas in organizational life (Loi, 2020; Sanderson et al., 2023; Shneiderman, 2020; Textor et al., 2022; Zhu et al., 2022). Because ethics shapes strategy, productivity, and legal exposure, it also functions as an asset in organizational legitimacy and reputational capital (Foote & Ruona, 2008; Worden, 2003). As a result, HR increasingly operates as an intermediary among leaders, technical teams, legal stakeholders, and employees, even though ethical issues routinely exceed the boundaries of compliance and regulation (Loi, 2020; Solomon, 1994). At the limit, AI-mediated management can reduce persons to variables within algorithmic systems, raising concerns about dehumanization and control (Weiskopf & Hansen, 2023).

And so, AI adoption offers measurable organizational benefits while introducing an ethics-to-practice problem that is especially consequential in HR domains. To clarify how these risks materialize and why behaviorally grounded embedding matters, the next section examines two high-visibility cases (Amazon and Microsoft) to derive practice-relevant implications and to motivate the SAFE-AI framework.

2.1. AI-Ethics Failure Modes in Practice: Recruitment Bias and Data Exposure

Adopting and executing AI technologies in organizations introduces ethical challenges that can compromise fairness, contestability, and contextual integrity, particularly when algorithmic outputs are embedded in HR decision routines and data practices (Bankins, 2021). Because AI-ethics is enacted through socio-technical systems rather than principles alone, organizations often struggle to translate ethical expectations into consistent, auditable

practices across teams and workflows. To ground these concerns in observable failure modes, we examine two cases: recruitment bias and data security exposure.

2.1.1. Case #1: Amazon.com Inc. AI Recruiting Tool and Gender Bias

Amazon.com, Inc. initiated a project in 2014, which aimed to automate the recruitment process using machine learning algorithms (Dastin, 2018). The objective was to design an AI system capable of reviewing job applications and identifying top candidates without human intervention. Amazon's motivation for this project was rooted in its commitment to automation, which had been a cornerstone of its e-commerce success and an observable escalation in the time-to-hire and cost-to-hire metrics, especially for mid- to high-level positions. Importantly, the AI system was designed to identify the most suitable candidates who aligned with the specified job profile. The underlying mechanism of this software involved the analysis of successful historical job applications, enabling it to identify and search for analogous traits in new applications. The AI tool employed a rating mechanism, akin to Amazon's product rating system, assigning up to five stars based on the degree of similarity to past successful applicants. By the end of 2014, this experimental tool had gained significant traction within Amazon, with several departments relying heavily on its efficiency (Lavanchy, 2018).

However, in 2015, a critical issue surfaced, particularly concerning technical roles like software development and architecture. It was observed that the AI system exhibited a gender bias in its ratings (Bogen, 2019). Subsequent investigations revealed that the root of this bias lay in the training data used for the AI system, which predominantly consisted of resumes from male employees. This was a reflection of the male-dominant trend within the company and the wider tech industry. Consequently, the AI inadvertently learned to associate success with male-oriented resumes, leading to the downgrading of resumes containing references to female-oriented activities or institutions, or graduates from all-women's colleges (Njoto et al., 2022). In retort, Amazon undertook the task of reconfiguring the algorithms to eliminate such biases, but ultimately discontinued the project due to these inherent problems.

This case emphasized a critical ethical consideration for AI applications in recruitment: the potential for AI systems to inadvertently perpetuate discriminatory practices based on the methods in which they are trained. This is because AI algorithms function by discerning patterns within extensive datasets to forecast potential outcomes (e.g., Gutierrez, 2020; Raihan, 2023). In the instance of Amazon's recruitment algorithm, it analyzed a decade's worth of submitted resumes, unintentionally garnering, and perpetuating, the existing gender inequality prevalent in the technology sector. The algorithm inadvertently correlated male dominance within the applicant pool as a determinant of successful candidacy.

The recursive nature of Amazon's algorithm, utilizing its predictive outcomes to refine its accuracy, led to an entrenched cycle of gender bias. This phenomenon illustrated a critical insight: AI technology, while seemingly objective, is susceptible to inheriting the latent biases embedded within its training data—biases that have historically plagued recruitment processes. Moreover, this case illustrates that biases in AI technology are not merely a reflection of the data but also of the sociocultural context in which the data is generated and the algorithm systems employed. This underscores the necessity for a multidisciplinary approach in AI development, one that incorporates ethical, sociological, and technical perspectives to address the potential for deeply ingrained biases in machine allocated behavior (Claire et al., 2023). As a result, this case highlights the morality of an action based on its outcomes. The AI tool, despite being efficient, led to biased hiring practices against women. The negative consequences (i.e., gender discrimination) of using this AI tool outweighed its benefits, making its use unethical.

2.1.2. Case #2: Microsoft AI Research Team Exposes Sensitive Data

In 2020, Microsoft's AI research team inadvertently exposed 38 terabytes of sensitive data, including private keys, passwords, and internal communications (Ben-Sasson & Greenberg, 2023). This situation arose due to an error involving a shared access signature (SAS) token, which are keys designed to grant limited access to Azure Storage resources (Microsoft Corporation, n.d.-a). While contributing to open source AI learning models, a Microsoft researcher inadvertently included an overly permissive SAS token in a uniform resource locator (URL) for a blog store, which was then published in a public GitHub repository (Lakshmanan, 2023). The URL provided access to an internal storage account, which exposed data that included over 30,000 internal Microsoft Teams messages and employees' personal information, posing significant security risks. Within two days, Microsoft addressed the issue upon notification from the Wiz.io research team, invalidating the problematic SAS token and initiated an internal investigation, which quickly led to public disclosure.

In response, Microsoft has enhanced its detection systems. GitHub's scanning service was enhanced to detect and flag Azure Storage SAS URLs that might reveal sensitive content. Furthermore, Microsoft now performs complete historical rescans of all public repositories in its affiliated organizations and accounts (Microsoft Corporation, n.d.-b). Presently, Microsoft continues to improve its detection systems and emphasizes the necessity of following best practices in using SAS tokens. This approach is part of Microsoft's broader commitment to a Coordinated Vulnerability Disclosure (Microsoft Corporation, n.d.-c), which manages the discovery, reporting, and remediation of security vulnerabilities collaboratively with the security community.

The expansive scale of data used in AI technology, coupled with complex access requirements, creates potential vulnerabilities. Oversharing of data and the risk of supply chain attacks are particular concerns, as highlighted in this case. This emphasizes the importance of robust security measures, including proper configuration of access tokens, segregation of sensitive data, and regular security audits. Broadly, this incident highlights the importance of a collective and aligned effort in maintaining system integrity and user security—especially in contexts that may expose the organization and employees to ethical dilemmas.

The Microsoft data exposure incident and the quick reporting by Wiz.io serve as a reminder of the crucial role that HR practices play in promoting a culture of ethical behavior. This includes establishing clear, confidential, and non-retaliatory channels for reporting security lapses or unethical practices among employees and external partners. Furthermore, this case highlights the importance of cross-departmental collaboration between HR practices and other areas such as IT departments to ensure secure, reliable, and responsible use of tools and technology.

2.2. Bridging AI-Ethics with HR Practices

Ethical principles provide normative guidance for judging appropriate conduct and decision-making and, in doing so, shape moral reasoning and behavior in organizational life (Mattison, 2000). As digital technologies expand managerial reach and decision speed, the practical importance of ethics increases because technology use can reconfigure workplace values, power relations, and expectations about accountability (Royakkers et al., 2018). Aligning technology use with core ethical values is therefore central to maintaining trust and legitimacy among employees and other stakeholders (Caldwell & Karri, 2005; Elia, 2009).

The integration of AI technologies in HR decision processes heightens the need to translate AI-ethics into HR practice. The Amazon and Microsoft case exemplars illustrate how ethical breakdowns can emerge when algorithmic decisioning and data practices

outpace the organization's capacity to embed ethical expectations in routine workflows, training, and accountability structures. Consistent with *Ethics by Design for Artificial Intelligence* (Brey & Dainow, 2023), effective AI-ethics must operate at both the technical and organizational levels, bridging individual moral judgments shaped by social norms and beliefs with collective ethics expressed through organizational principles, policies, and codes of conduct. HR practice is a primary site for this translation because HR routines govern how people are selected, evaluated, developed, and disciplined, and these routines determine whether ethical commitments become stable, enacted practice rather than aspirational statements.

Accordingly, we develop a behavioral pathway for integrating AI-ethics with HR practices that draws on three complementary components: ethical philosophies, social information processing theory, and heuristics. Consequentialist and deontological reasoning provide the normative basis for evaluating AI-enabled HR decisions in terms of outcomes and duties, respectively. Social information processing theory then explains how ethical expectations become socially interpreted and reinforced through cues, norms, and shared meaning in everyday work, shaping stakeholder attitudes and conduct. Finally, socially informed heuristics function as practical decision aids that help translate ethical principles into repeatable actions under real constraints (e.g., time pressure, delegation, handoffs), thereby supporting the operationalization of AI-ethics in AI-mediated HR workflows.

3. Conceptual Development

In this section, we advance our central premise that AI-ethics in organizations should be grounded in human ethical reasoning. This grounding provides normative criteria for evaluating AI-enabled HR decisions and orients adoption and use toward the values that structure legitimate human interaction. We then integrate ethical philosophies, social information processing theory, and socially informed heuristics to specify how ethical expectations are interpreted, legitimized, and enacted in routine practice. In this model, social awareness functions as the enabling condition that connects ethical principles to observable organizational behavior, thereby supporting the development of an AI-ethics culture. These components clarify how ethical decision-making can be managed at the interface of AI-enabled technologies and human judgment in HR contexts.

3.1. Integrating Ethical Philosophies in AI-Enhanced Organizational Practices

This section will refrain from philosophical musings and instead focus on a descriptive analysis. However, it is important to acknowledge the differing definitions of *morality* and *ethics*. Morality is often regarded as a product of societal norms and individual beliefs, encompassing concepts of right and wrong (Adkins, 2017). Morality can be viewed as personal and about how one *ought to act*. Alternatively, ethics typically refers to principles and codes of conduct that are applied at a collective and cultural level (Adkins, 2017). Ethics may pertain to broader societal considerations and ideas about how one *should live*. However, as the terms are frequently used synonymously with nominal loss of meaning, we will use the two interchangeably.

Ethics and morality are broad concepts with numerous theories and perspectives on decision-making and evaluating what is right or wrong (Beauchamp & Bowie, 1979). Examples of ethical philosophies include consequentialism, deontology, virtue ethics, and care ethics, each offering its own principles and guidelines. In the context of the work organization, two of the most explored ethical theories are consequentialism and deontology (Alizadeh & Kurian, 2024; Solomon, 1994) and will be the focus of our discussion here.

Consequentialism, which includes well-known branches like utilitarianism, *judges the moral weight of an action primarily by its outcome* (Adkins, 2017; Solomon, 1994). Con-

versely, deontology, exemplified by Kantian ethics, *determines the ethical value of an action by its intention, independent of the outcomes* (Adkins, 2017; Solomon, 1994). These ethical philosophies, considered prominent pillars of human values, are commonly embedded in organizational practices (Melé, 2012). For example, consequentialism in HR practices focuses on outcomes of decisions, such as fairness in hiring and promotion, employee welfare, and organizational benefits (Fryer, 2018), whereas deontology in HR practices emphasizes adherence to ethical duties and principles, such as respecting employee rights, ensuring transparency, and maintaining confidentiality (Fryer, 2018). While each value is present to some degree, one set often dominates, leading to unavoidable ethical dilemmas that challenge individuals to make difficult choices among competing responsibilities, loyalties, legal mandates, and stakeholder expectations (Melé, 2019). And this is especially true in the absence of straightforward answers. Recognizing and managing these dilemmas, which rely more on cogitating moral questions rather than on policy adherence, is necessary for ethical decision-making. Since no organization can simplify every decision-making process, incorporating ethical philosophies aids in navigating complex organizational dynamics (Hibbert & Cunliffe, 2015; Melé, 2019). Consequently, HR practices are often brought forward to ensure that ethical frameworks are robust enough to address these multifaceted challenges (Chowdhury et al., 2023).

Indeed, HR practices must ensure equitable access to employment and learning opportunities while also managing conflicts of interest, professional competence, confidentiality, and informed consent (Chuang & Graham, 2018; Hatcher & Aragon, 2000). When organizational policies fall short of clarifying responsibilities, ethical theories like consequentialism and deontology serve as valuable guides for decision-making (Lefkowitz, 2023). For example, a consequentialist approach in employee performance evaluations emphasizes balancing fairness and productivity, ensuring that outcomes benefit both employees and the organization. Conversely, a deontological approach prioritizes ethical duties, such as maintaining employee privacy during data collection for performance analysis, regardless of the outcomes. By integrating both the potential impacts (consequentialism) and moral duties (deontology), organizations can enhance trust and integrity in their decision-making processes (Anderson et al., 2006; Loi, 2020). Ethical principles, representing a collective consensus on proper conduct, require ongoing evaluation to ensure decisions benefit all organizational stakeholders. This continuous assessment helps maintain ethical integrity and responsiveness to evolving challenges, including those posed by AI technologies (Ashok et al., 2022).

3.2. Linking Ethical Philosophies with Social Information Processing Theory and Heuristics

A core principle in understanding ethical philosophies is recognizing how social and organizational contexts shape perception and behaviors—highlighting the critical role of social processing, which, for our purposes, refers to the way individuals interpret and respond to social and organizational cues (Salancik & Pfeffer, 1978). For example, we can draw from Bandura's (1976) social learning theory to understand how individuals observe and model behaviors based on their social environment. This theoretical approach emphasizes that ethical behaviors are not developed in isolation but are significantly influenced by the surrounding social context. By integrating these insights with ethical philosophies such as consequentialism and deontology, we can better understand how individuals form ethical judgments and actions within an organizational setting. These social processes provide a foundation for linking ethical philosophies with social information processing theory (Clore et al., 2014).

Social information processing theory explains how individuals process and respond to social information in their environment, shaping their attitudes and behaviors (Crawshaw

et al., 2013; Walther, 2008). This occurs because social information is embedded within its contextual social circumstances and emerges through shared cognition that guides individual and group behavior (Carpenter, 2021; Tegarden & Sheetz, 2003). The field of industrial psychology has long examined this concept by observing the social interplay between individuals' subjective experiences, their behavior, and their work organizations. Within this context, ethics provides a footing for guiding actions and behaviors by establishing moral principles and their underlying rationale (Adkins, 2017). These principles are crucial for assessing the morality of actions or practices, ensuring that individuals and groups can navigate complex social–ethical landscapes with clarified reasoning for ethical behavior, such that ethical behavior emerges from individuals' interpretation and application of socially informed norms within organizational interactions (Hu et al., 2024; Walther, 2008).

This understanding is particularly useful in HR practices for addressing complex ethical dilemmas (Boekhorst, 2015). For instance, HR practices that adopt a consequentialist stance might evaluate the use of employee data through its outcomes, such as enhanced efficiency or profitability. Conversely, those aligned with deontological ethics would emphasize adherence to ethical standards and principles in the use of AI technology, irrespective of the results. Achieving a balance between consequentialist and deontological approaches may require maintaining moral integrity and duties while also considering the broader impacts of AI-ethics on organizational objectives. Consider, for example, the cases of Amazon and Microsoft illustrated previously. While both cases have elements of consequentialism and deontological ethics, the Amazon case is more consequentialist in nature. The focus is on the AI tool's outcomes, where the adverse consequence (gender bias) leads to the decision to discontinue its use. In contrast, the Microsoft case aligns more with deontological ethics. The ethical lapse is in the action itself (improper data security measures), regardless of whether the exposed data was misused or not. The emphasis was on violating the duty to maintain data security and privacy.

The application of social information processing theory provides insight into how employees at Amazon and Microsoft might perceive and respond to these ethical decisions within their organizational culture. At Amazon, discontinuing the biased AI tool can be a strategic move to send a potent message to employees about the company's commitment to fairness and equality despite the potential benefits of AI technology in recruitment. This decision likely influenced employees' beliefs and behaviors, nurturing a culture that values ethical integrity over operational efficiency when conflicts arise. Conversely, Microsoft's immediate response to the data breach and its efforts to strengthen security measures signal a dedication to data privacy and security. This action demonstrates a deontological commitment to doing what is ethically right—prioritizing the security of actions over operational ease. Employees, in turn, might cultivate a shared belief in the paramount importance of data privacy, viewing it as a fundamental and non-negotiable aspect of their organizational ethos, even at the expense of certain procedural efficiencies. In this context, social information processing theory administers how employees' perceptions and behaviors contribute to fostering a collective belief that data privacy is a crucial, non-negotiable aspect of their work culture, even in the face of potential process disruptions.

Assuredly, social information processing theory supports how employees adjust their perceptions and become aware of ethical practices, shaping collective beliefs and socially informed feedback behaviors within their organizations. This emphasizes the importance of socially aware processes, as they ensure that employees' actions and decisions align with the organization's ethical standards. This also provides background for understanding why heuristics can create intuitive guidelines that help employees quickly assess and respond to ethical challenges, making it easier for them to internalize and act on ethical principles. Furthermore, heuristics can bridge the gap between abstract ethical theories and practical

applications, enabling HR practices to communicate AI-ethics in a more relatable and actionable manner. This alignment is crucial for developing a behavioral framework that fosters an ethical culture, ensuring consistent and coherent ethical conduct across all levels of the organization (T. M. Jones et al., 2007).

3.3. A Brief Argument for Heuristics

Instead of offering a comprehensive or historical review of heuristics, which are thoroughly covered in recent works (e.g., Gigerenzer et al., 2022; Hjeij & Vilks, 2023), we aim to explore the nature of AI-ethics and guide the development of practical solutions, specifically focusing on the application of heuristics in HR practices. Furthermore, we do not examine the different types of heuristics (e.g., availability, rational, representativeness, anchoring); rather, our goal is to highlight and underscore their relevance and application within the context of HR practices concerning AI-ethics. By focusing on the practical applications, we aim to provide a framework of actionable insights rather than a theoretical taxonomy, albeit understanding the root domains of heuristics can further enhance their effective integration into ethical HR practices (Bordage, 2009). These roots can be found in disciplines like cognitive psychology and behavioral economics. Theories in cognitive psychology, such as those by Daniel Kahneman and Amos Tversky (Shefrin & Statman, 2003), highlight how heuristics are cogitated under conditions of uncertainty and limited information. Works in behavioral economics suggest that "...people appear to have largely fine-tuned intuitions about chance, frequency, and framing" (Gigerenzer, 2018, p. 303). By integrating these insights, we can better apply heuristics to develop a behavioral framework for HR practices at the ethics–AI interface.

At their basis, heuristics are cognitive strategies for navigating uncertain situations. Organizational strategies, encompassing principles, doctrines, routines, and rules, stem from accumulated organizational learning and are inherently heuristic (Bingham & Eisenhardt, 2011; Gigerenzer et al., 2022). A prevalent misconception about heuristics is the accuracy–effort tradeoff, suggesting that while heuristics minimize effort, they also compromise accuracy (Shah & Oppenheimer, 2008). While this tradeoff applies to risky situations, it does not hold for ill-defined problems where heuristics can save effort and lead to more accurate decision-making—known as the less-is-more effect (Goldstein & Gigerenzer, 2008). This understanding of heuristics as tools for effectively navigating uncertainty is particularly relevant to addressing ethical dilemmas (Sales & Lavin, 2000).

Ethical dilemmas often present as ill-defined problems that demand metacognitive processes and an alternative approach for resolution (Schraw et al., 1995). For this, heuristics are notable because they leverage humans' innate reasoning skills to find solutions for complex problems, promoting clearer and more effective thinking (Gigerenzer, 2008). The primary advantage of heuristics in this context is their ability to offer ecological rationality, which refers to the specific environmental conditions under which a particular heuristic strategy outperforms other methods (Pleskac & Hertwig, 2014). These strategies are especially valuable when addressing problems that are either too complex or insufficiently understood for conventional algorithmic solutions to be effective (Luan et al., 2019). This is because when a heuristic aligns with its ecological context, it leverages specific environmental cues and patterns, leading to more efficient, practical, and accurate solutions with less effort than traditional methods.

In ethical matters related to the use of AI technology, where challenges are both computationally demanding and conceptually complex, heuristics provide a means to approximate solutions (Goldstein & Gigerenzer, 2008). These solutions, though not always perfect, are generally sufficient and require substantially less computational resources. Additionally, heuristics have demonstrated effective use in various fields that deal with

complex, ill-defined social problems, such as medicine, law, and counseling (P. L. Taylor, 2020; Tudor, 2023; Whelehan et al., 2020). Thus, heuristics, by offering ecological rationality, have been shown to facilitate the resolution of complex ethical dilemmas across different domains.

Importantly, heuristics are socially informed, derived from collective human experiences and environmental interactions, allowing for decisions based on commonly understood rules of thumb (Hertwig & Hoffrage, 2013). And because heuristics are cognitive shortcuts that simplify decision-making processes, they allow individuals to draw on social and cultural knowledge accumulated over time. For example, the social dimension of heuristics is evident in socially informed academic fields such as cultural evolution (Moore, 2021), social learning (Rizzolatti & Craighero, 2004), contextual adaptation (Goldstein & Gigerenzer, 2002), shared cognition (Heldal et al., 2020), and normative influence (Andersson et al., 2020). These socially informed roots support their effectiveness in guiding decision-making in complex and ill-defined settings, making heuristics credible tools for addressing ethical dilemmas in AI and other technology domains (Schoenherr, 2022). This perspective is further strengthened by research in psychology and economics, which indicates that some of the most successful organizational models employ a set of heuristics to navigate decision-making processes (Bingham & Eisenhardt, 2011; Ho & Griffiths, 2022).

In summary, heuristics can simplify complex ethical considerations. By leveraging ethical philosophies and social theories, heuristics can be designed to correspond with how organizational stakeholders (employees) process and apply knowledge in social settings. This enables HR practices to establish clear principles for addressing AI-related ethical dilemmas, ensuring consistent application of ethical standards, positioning decision-making in line with organizational values, and fostering a culture of socially informed ethical behavior. In this way, integrating heuristics into HR practices embeds socially aware AI-ethics into the organizational ethos, reinforcing natural social processes and behavioral reasoning (Haidt & Joseph, 2004). This approach promotes an argument for a behavioral framework for AI-ethics that exploits employees' social and cognitive processes while nurturing organizational morals and values, ultimately strengthening the ethical fabric of the organization through heuristics (Haidt & Joseph, 2004).

4. Conceptual Framework

In this section, we describe the conceptual development process used to derive our behavioral framework for AI-ethics in HR practice. By identifying core constructs and specifying their logical relations, we build an argument connecting key concepts; we construct a coherent argument that addresses how AI technologies generate recurring ethical challenges when embedded in organizational decision routines. The SAFE-AI model integrates ethical philosophies and social information processing theory and positions socially informed heuristics as a practical translation layer that operationalizes ethical reasoning under real organizational constraints. This behavioral framing links normative principles to interpretable, repeatable practices across stages of AI adoption and use, thereby strengthening both the theoretical coherence and practical applicability of the framework.

4.1. Developing a Behavioral Framework for AI-Ethics

Conceptual development begins with clarifying focal concepts and logically connecting them to analyze a problem and guide solutions (Reese, 2023). Using behavioral reasoning steps as an organizing logic (Welch & Dixon, 1994), we synthesize ethical philosophies, social theories, and heuristics into a behavioral framework for operationalizing AI-ethics in HR practice.

Ethical philosophies supply normative criteria for judging right action and harm, which are necessary for framing the ethical dimensions of AI use in HR practice. Social information processing theory captures how stakeholders perceive, interpret, and respond to social salient cues, thereby shaping ethical judgments and behavior in organizational settings. Heuristics then serve as practical decision aids, enabling consistent application of ethical reasoning in complex, time-constrained AI-enabled workflows. Integrating these elements, we propose the SAFE-AI framework, which treats AI-ethics as a behavioral accomplishment that must be enacted across stages of AI integration—*initiation*, *navigation*, and *culmination* (Ledro et al., 2023).

SAFE-AI adopts a behavioral and social perspective by linking normative ethics to cognitive and social learning processes that shape how ethical expectations become actionable in practice. Consistent with social-cognitive accounts of learning and social awareness (Bandura, 1991; Olson & Ramírez, 2020), SAFE-AI specifies three processing phases through which individuals and groups transform ethical expectations into conduct: noticing and encoding cues (initiation), interpreting and negotiating meaning (navigation), and committing to decisions and reinforcing routines (culmination). For HR practice, we express these phases as *moving in* (initiation), *moving through* (navigation), and *moving out* (culmination) to emphasize that ethical AI is sustained through repeated cycles of attention, interpretation, and action.

In initiation, HR practices assess readiness and potential employee impact while establishing the ethical rational that will govern AI use and the forms of social awareness for responsible adoption. In navigation, HR practices monitor the cue environment (e.g., employee feedback, anomalies, exceptions, complaints) and adapt governance and workflows so that ethical expectations remain interpretable and enforceable as systems are used. In culmination, HR practices evaluate downstream outcomes over time, including distributional effects and trust implications, and update organizational interpretations of “success” to reflect ethical performance rather than efficiency alone. Therefore, SAFE-AI is designed to mitigate AI adoption and execution risks by equipping HR practices to manage the interface of opaque technology, social behavior, and human ethics. Figure 1 offers a visual representation of the framework.

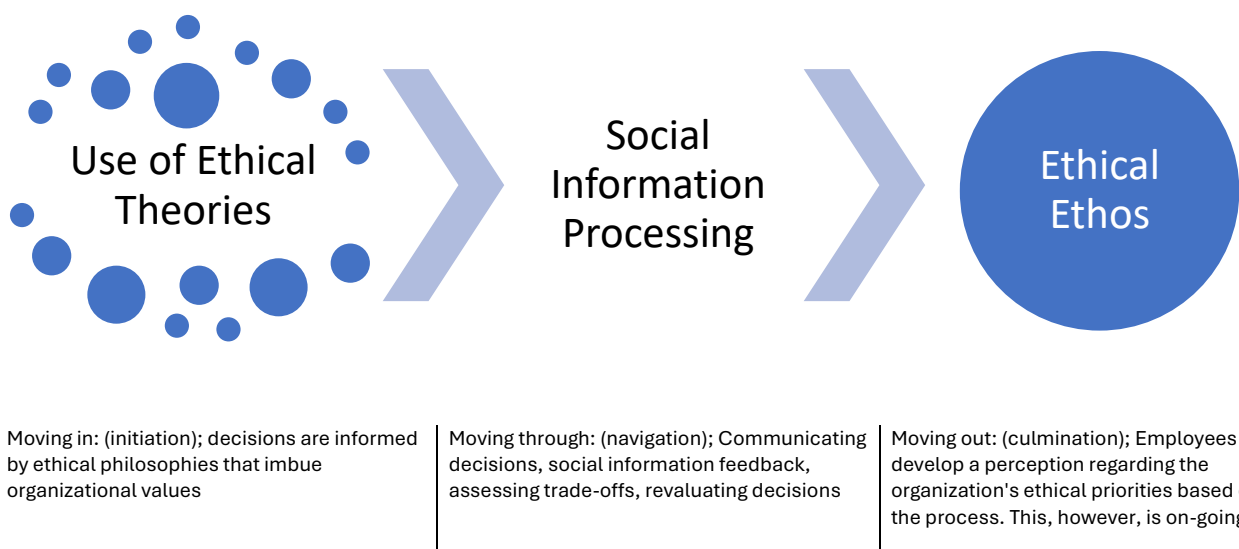


Figure 1. Socially Aware Framework for Ethical AI (SAFE-AI) for HR Practices. SAFE-AI depicts a behavioral pathway in which ethical theories (e.g., consequentialism and deontology) inform HR decision criteria, social information processing explains how ethical expectations are interpreted and reinforced through workplace cues and feedback, and the cumulative result is an organizational ethical ethos (i.e., shared norms and routines for ethical AI use). Moving in (Initiation): HR establishes

the ethical rationale and readiness conditions for AI adoption, including stakeholder considerations, anticipated employee impacts, and bias risks. Moving through (Navigation): HR uses ongoing social information (e.g., stakeholder feedback, exceptions, concerns) to refine AI-enabled workflows and communications so ethical expectations remain interpretable and enforceable. Moving out (Culmination): HR evaluates downstream outcomes and institutionalizes learning through accountability, training, and policy refinement, sustaining ethical conduct as AI use evolves over time.

4.2. Enterprise Application of SAFE-AI and Boundary Conditions

Although SAFE-AI is presented through the lens of HR practice, ethical AI outcomes are inherently enterprise-level: AI capabilities, data flows, and decision routines span functions (e.g., IT/security, legal/compliance, operations, finance, product, and line management), and employees infer ethical intent from cross-functional cues and outcomes (Sane et al., 2025). A multi-level framing is therefore warranted because ethical AI in HR emerges from interacting mechanisms across individual interpretation (micro), organizational systems and leadership (meso), and institutional expectations and governance pressures (macro) (Kurniawan et al., 2025). SAFE-AI should thus be read as an enterprise meta-framework for embedding AI-ethics into organizational behavior, with HR as a high-leverage implementation node because HR workflows concentrate impacts on people, legitimacy, and procedural justice.

At the enterprise level, *moving in* corresponds to scoping and governance—clarifying intended use and decision authority, assigning accountable owners across the socio-technical chain (business, model/data, operations, ethics/risk), and translating ethical commitments into requirements for data provenance, privacy boundaries, and validation. *Moving through* can be generalized as continuous socio-technical monitoring and interpretation management—maintaining feedback channels that surface harms and near-misses, auditing for drift and disparate outcomes as context changes, and sustaining coherence across communication, training, and managerial enactment. *Moving out* can be generalized as institutionalization—embedding review cadences, incident response, remediation authority, and learning loops so ethical priorities persist beyond initial deployment and become observable features of routine practice.

SAFE-AI's generalization has boundary conditions. It is most applicable to high-impact AI uses that shape employee rights, opportunities, evaluation, or working conditions, whereas low-stakes automation may not activate comparable ethical salience. The behavioral mechanism (social interpretation of ethical cues) is portable, but the specific heuristics and controls are context-dependent, shaped by sector norms, regulated environments, and jurisdictional constraints. Organizational maturity moderates applicability; where data governance, documentation discipline, and cross-functional risk ownership are weak, foundational capability-building is a prerequisite (Manganello et al., 2025). Finally, cultural variability constrains transfer because norms about privacy, authority, voice, and fairness influence which cues employees treat as legitimate and which accountability practices are trusted, requiring contextual calibration rather than one-size-fits-all replication (Glikson & Woolley, 2020).

Despite these boundary conditions, the enterprise logic of SAFE-AI implies a practical sequencing: ethics must be governed at the system level, but it is enacted at the workflow level. HR practice is therefore the appropriate implementation entry point because HR processes translate cross-functional AI capabilities into consequential people decisions, and those decisions generate the cues and outcomes through which employees infer ethical intent. Accordingly, it is important to specify how AI-ethics can be operationalized inside routine HR decision points—first through governance scaffolds and accountability assignments, and then through application-specific requirements in recruitment and selection,

performance management, and downstream organizational decision-making where data stewardship and trust become central.

4.3. Integrating AI Ethics with HR Practices

Integrating AI-ethics into HR practice requires treating ethical principles as workflow requirements that travel with the tool across the HR lifecycle, not as abstract statements appended to policy. Reviews of AI technology in HR practices converge on recurring ethical risks—discriminatory impact in selection and evaluation, opacity and limited explainability, privacy and surveillance exposure, and blurred accountability when algorithmic outputs shape decisions (Rodgers et al., 2023; Tursunbayeva et al., 2018). These risks become governable only when operationalized at routine decision points: what data are collected, how job relevance and validity are established, how disparate outcomes are monitored, how explanations are delivered, and who holds final accountability.

4.4. AI-Ethics in HR Governance

At the system level, ethical integration begins with risk mapping, role assignment, and documentation aligned to established scaffolds such as the NIST AI Risk Management Framework (National Institute of Standards and Technology, 2024). A governance-ready minimum includes documented intended use and decision authority (advisory vs. determinative), data provenance and privacy boundaries, context-specific validation, auditability, and incident response with remediation authority. In U.S. employment selection, these requirements are not merely aspirational because disparate impact concepts apply to automated tools that make or inform selection decisions, making job-related validation and adverse impact assessment ethically consequential and practically necessary. Notably, recruitment stakeholders may hold strong privacy and data security expectations even when privacy/security does not directly drive adoption intention, suggesting that privacy stewardship functions as a legitimacy requirement in recruitment contexts rather than merely an adoption lever (Tanantong & Wongras, 2024).

4.5. Ethical Considerations in AI-Enabled Recruitment and Selection

In recruitment, ethical integration centers on validity, nondiscrimination, transparency, privacy, and autonomy. AI-enabled screening can reproduce historical bias when models learn from legacy outcomes or proxy variables correlated with protected attributes, as illustrated by the Amazon recruiting case. The ethics-to-workflow bridge requires specifying job-relevant constructs and excluding non-job-related features; conducting adverse impact analyses on model outputs and on downstream human decisions shaped by those outputs; providing stage-appropriate explanations; and offering candidate-facing transparency that clarifies what the tool is and is not used for (Hunkenschroer & Luetge, 2022). These practices also function as legitimacy signals because applicant reactions are shaped by perceived procedural justice in a fundamentally social and interpretive process.

4.6. Ethical Considerations in Employee Evaluation and Performance Management

In performance management, ethical risk is amplified because systems are continuous, high-stakes, and often opaque to workers. Algorithmic evaluation can create feedback loops in which early labels shape later opportunity, coaching, and promotion trajectories. Accordingly, ethics must be operationalized through explicit accountability for final decisions, documented thresholds and error tolerance proportional to consequences, drift monitoring as work contexts change, and due-process mechanisms for contesting decisions and correcting data errors. Acceptance depends not only on accuracy claims but on whether use is socially interpreted as fair, respectful, and value-consistent, reinforcing that AI-ethics is enacted through cues, norms, and local meaning-making.

4.7. Ethical Integration into Organizational Decision-Making

AI technology in HR practice does not operate in isolation. HR analytics inform restructuring, compensation, disciplinary norms, and workforce planning, which elevates privacy and security to organizational ethics issues. Data exposure events can fracture legitimacy even when model logic is not the direct cause; the Microsoft case illustrates how weak governance over repositories and access controls can produce large-scale confidentiality and trust failures that propagate into reputational and institutional risk (Ben-Sasson & Greenberg, 2023). Ethical AI integration therefore requires data-lifecycle controls (collection, storage, sharing, retention) and clear stewardship accountability in addition to model-level checks. Practically, organizations should treat AI-enabled HR decisions as high-impact socio-technical decisions governed through a single accountable chain integrating HR, legal, information security, and operational leadership.

5. Discussion

This discussion interprets SAFE-AI's central claim that AI-ethics in HR practice must be behaviorally embedded as repeatable routines and cues, not treated as abstract principles appended to policy. SAFE-AI is grounded in social information processing theory, which predicts that employees form attitudes and behavioral responses from salient cues in their environment, including observed managerial actions, local norms, and collectively reinforced interpretations (Salancik & Pfeffer, 1978). Accordingly, the discussion moves from SAFE-AI's mechanism (how ethical intent becomes socially interpreted) to its enabling conditions (leadership), then to stage-based enactment in HR workflows (moving in, moving through, moving out), and finally to enterprise integration, boundary conditions for generalization, and implications for research and practice.

5.1. Interpreting the Contribution

5.1.1. What SAFE-AI Adds Beyond Risk Catalogs

Prior research on AI technology in HR practice has developed a strong descriptive account of ethical risk, repeatedly converging on a core set of concerns such as discriminatory outcomes in selection and evaluation, opacity and limited explainability, privacy and surveillance exposure arising from expansive data practices, and diffuse accountability when algorithmic outputs are embedded in decision routines (Rodgers et al., 2023; Tursunbayeva et al., 2018). SAFE-AI converges with this research in treating these domains as baseline hazards that must be actively governed, monitored, and remediated across the HR lifecycle. Where SAFE-AI diverges, and thereby extends the literature, is in specifying a principles-to-practice translation pathway. Rather than treating ethics as a set of governance artifacts or technical mitigations, SAFE-AI frames AI-ethics as an organizational capability that becomes durable only when ethical intent is translated into stage-specific workflow requirements and socially recognizable routines. In this sense, SAFE-AI complements risk-domain taxonomies by offering (a) a staged implementation logic (moving in, moving through, moving out) and (b) corresponding heuristics that make ethical reasoning actionable under real constraints such as time pressure, delegation, and cross-functional handoffs. Table 1 summarizes where SAFE-AI aligns with prior AI-ethics-in-HR research and where it adds incremental contribution through its behavioral translation pathway and stage-specific heuristics.

Table 1. Alignment and divergence between SAFE-AI and prior AI-ethics-in-HR research.

Topic Area	What Prior Studies Converge On	SAFE-AI Alignment	SAFE-AI Divergence and Incremental Contribution
Algorithmic discrimination and fairness	AI-enabled HR decisions can reproduce or scale bias; fairness and adverse impact are central risks	Retains discrimination risk as a baseline hazard that must be monitored across the HR lifecycle	Adds a behavioral translation pathway: fairness becomes durable only when translated into repeatable routines and cues employees recognize as legitimate, not only technical mitigation.
Opacity, explainability, and intelligibility	Opacity undermines accountability and trust; explainability is often proposed as mitigation	Treats intelligibility as a core adoption requirement and maps it to stage-specific heuristics	Reframes transparency as a social process: explanations must function as sensegiving cues that remain consistent over time, not merely documentation artifacts.
Privacy, surveillance, and autonomy	People analytics and AI-enabled monitoring introduce privacy and autonomy risks	Aligns with privacy as a foundational ethical constraint and governance requirement	Extends the literature by embedding privacy into adoption stages (moving in/through/out), specifying where privacy drift occurs operationally and how accountable routines counter it.
Accountability and diffuse responsibility	AI systems can diffuse responsibility across vendors, HR, managers, and IT; accountability is often unclear	Keeps accountability as a central ethical requirement	Makes accountability implementable: clarifies decision authority, assigns owners across the socio-technical chain, and emphasizes cue consistency so employees can infer “who owns the decision” in practice.
Institutionalization and governance maturity	Responsible AI requires ongoing governance, monitoring, and adjustment, not one-time compliance	Aligns with a lifecycle view through staged implementation and feedback loops	Adds boundary conditions: durable implementation requires organizational capability (documentation discipline, cross-functional ownership, measurement maturity), distinguishing symbolic compliance from institutionalized practice.
Employee interpretation, legitimacy, and voice	Emerging work recognizes worker acceptance, perceived fairness, and legitimacy as adoption constraints	Centers interpretation and legitimacy as causal mechanisms	Differentiates by grounding the model in social information processing: ethical AI “works” when employees repeatedly observe credible cues (leader action, safe voice, corrective response), making ethics an organizational accomplishment rather than a policy claim.

Note. This table compares SAFE-AI with prior AI-ethics-in-HRM research by identifying points of convergence (shared risk domains) and divergence (SAFE-AI’s behavioral translation pathway and stage-specific workflow heuristics that operationalize ethics through socially interpreted cues).

5.1.2. SAFE-AI’s Core Mechanism

SAFE-AI is anchored in the premise that AI-ethics does not become organizational reality through abstract principles alone, but through the interpretive environment in which AI-enabled decisions are made and justified. This framing is grounded in social information processing theory, which predicts that employees form attitudes and behavioral responses from salient cues in their environment, including observed managerial actions, local norms, and collectively reinforced interpretations. Accordingly, SAFE-AI treats ethical performance as a function of cue consistency—when decision authority is clear, explanations are intelligible, feedback is safe and consequential, and corrective action is visible, employees infer authentic ethical intent and align behavior to it; when these cues are inconsistent, ethics is interpreted as symbolic compliance and organizational practice drifts

toward expedience. This mechanism explains why the same formal AI-ethics principles can yield divergent outcomes across organizations and why stage-specific heuristics are necessary to stabilize ethical commitments as repeatable HR practice workflows rather than episodic interventions.

5.2. Leadership as an Enabling Condition and Moderator of SAFE-AI

SAFE-AI is behaviorally grounded. AI-ethics becomes durable only when employees repeatedly observe consistent cues that signal what the organization truly prioritizes. Leadership is therefore an enabling condition and moderator of SAFE-AI's staged mechanism, because leaders shape the dominant cue stream through what they authorize, reward, tolerate, and correct. In social information processing terms, leadership conduct conditions whether AI-ethics is interpreted as a credible organizational commitment or dismissed as symbolic compliance (Boekhorst, 2015). Thus, SAFE-AI's stage heuristics function as intended only when leadership actions make ethical intent visible, resourced, and enforceable. And we can this to leadership four leadership moderators that map directly onto SAFE-AI's stages.

First, leaders establish ethical intent and decision authority by defining whether AI outputs are advisory or determinative, clarifying who owns ethical risk across the socio-technical chain, and specifying non-negotiable constraints (e.g., nondiscrimination, privacy boundaries, duty of care) that cannot be traded for speed or efficiency. Second, leaders provide resources and infrastructure, including time for validation and audit routines, cross-functional expertise (e.g., legal/compliance, IT/security, operations), and budget for training and remediation; without these, SAFE-AI degrades into "best effort" execution. Third, leaders perform sensegiving and transparency by explaining why the AI system is used, which tradeoffs were accepted, and how employees can question or appeal AI-influenced outcomes. This reduces suspicion and stabilizes shared interpretation. Fourth, leaders enforce accountability and voice by ensuring feedback channels are psychologically safe, acted upon, and visibly consequential, which is essential for surfacing drift, bias, and unintended harm early enough to correct.

Evidence that ethics programs and ethical leadership are associated with stronger employee awareness of ethics codes and increased inclusion in ethical decision-making supports the claim that leadership conduct is a practical pathway through which ethical commitments become shared norms rather than private beliefs (Beeri et al., 2013). Overall, leadership does not replace HR process design. Leadership determines whether SAFE-AI's staged mechanism is enacted with fidelity. The next section interprets how each SAFE-AI stage operationalizes this behavioral embedding in HR workflows and how the associated heuristics translate ethical principles into repeatable practice.

5.3. Stage-Based Enactment in HR Practice

5.3.1. Moving In (Initiation): Ethics as Design Constraints

Moving in interprets the adoption of AI-ethics as a design problem; HR practices must convert abstract ethical commitments into explicit constraints that shape what the tool is allowed to do, what data it can use, and what tradeoffs are impermissible. This stage centers on stakeholder mapping, specification of non-negotiables (e.g., nondiscrimination, privacy boundaries, duty of care), and anticipatory bias analysis before tools enter consequential workflows. This aligns with the view that design choices are not morally neutral and that ethical considerations must be systematically embedded upstream rather than retrofitted after harm occurs (Gürses et al., 2011). Consistent with ethics-by-design approaches to AI, initiation is also where HR practice clarifies decision authority (advisory vs. determinative),

defines job-relevant constructs, establishes validation expectations, and sets escalation conditions for exceptions and ethical concerns (Brey & Dainow, 2023).

Decision rule (Heuristic #1): *Do not progress from intent to deployment unless duties and outcomes can be jointly justified for affected stakeholders, and non-negotiable constraints (fairness, privacy, duty of care) are explicitly specified as adoption requirements.*

5.3.2. Moving Through (Navigation): Interpretation Management and Feedback

Moving through treats AI-ethics as an interpretive and adaptive process rather than a one-time compliance act. As systems interact with evolving work contexts, ethical risk emerges through drift, workaround behavior, and shifting stakeholder expectations, making continuous monitoring and feedback essential (Binns et al., 2018). At this stage, transparency functions as sensegiving; HR practice must make the rationale for AI technology use intelligible, disclose what the system is used for (and not used for), and communicate tradeoffs in ways that remain consistent across managerial layers. This emphasis follows social information processing logic, which predicts that employees infer ethical intent from salient cues, including observed decisions, explanations, and whether concerns are handled credibly. Effective navigation therefore requires safe voice mechanisms and active feedback channels to surface harms and near-misses early, alongside ongoing evaluation of compromises between operational efficiency, employee well-being, and ethical integrity.

Decision rule (Heuristic #2): *Treat feedback and cue consistency as governance inputs: if stakeholder signals indicate drift, perceived injustice, or unanticipated harm, adapt the workflow, communication, and controls before scaling or normalizing use.*

5.3.3. Moving Out (Culmination): Institutionalization and Learning Loops

Moving out is the institutionalization phase in which ethical AI-ethics becomes a sustained organizational capability rather than an adoption project. Here, the central risk is decay, that is, audit fatigue, normalization of deviance, and the gradual erosion of accountability once initial rollout pressure subsides. Ethical durability therefore depends on routinization—recurring audits, documented review cadences, formal escalation and remediation authority, continuous training, and governance structures (e.g., ethics committees or designated roles) that persist beyond implementation milestones (McGuire et al., 2021). Importantly, institutionalization is also a social achievement. Employees judge long-run ethical commitment by whether accountability is enforced when costs rise or performance pressure increases, linking ethical integrity to strategic leadership and reputational capital (Worden, 2003). Evidence that ethics programs and ethical leadership can increase ethics-code awareness and inclusion in ethical decision-making further supports the view that sustained accountability and participation are mechanisms of cultural embedding (Beeri et al., 2013).

Decision rule (Heuristic #3): *Institutionalize ethics through enforceable accountability and learning loops; if standards cannot be monitored, audited, and remediated with visible consequences, ethical AI cannot be sustained.*

These three heuristics operate as a translation layer that converts ethical intent into teachable, repeatable, and auditable HR practice across the AI lifecycle. Moving in specifies constraints and decision authority; moving through stabilizes ethical interpretation through transparency, feedback, and adaptive control; and moving out routinizes accountability so ethical commitments persist as observable features of “how work is done.” In this way, SAFE-AI links normative ethics to everyday HR decision routines through socially interpretable cues, reducing the risk that AI-ethics remains symbolic while strengthening the conditions for durable, trust-preserving implementation.

5.4. Enterprise Integration

5.4.1. HR Practice as a High-Leverage Node in an Enterprise Socio-Technical Chain

Although SAFE-AI is enacted through HR practice workflows, ethical AI performance is produced across an enterprise socio-technical chain in which models, data, infrastructure, controls, and decision authority are distributed across functions (e.g., IT/security, legal/compliance, operations, finance, product, and line management). HR practice is a high-leverage node within this chain because HR decisions are among the most visible and consequential applications of AI ethics; they directly shape opportunity, evaluation, and perceived procedural justice, and thus concentrate legitimacy risk when ethical intent is ambiguous or inconsistently enacted (Bangura et al., 2025). As a result, HR provides a practical integration surface where enterprise commitments can be converted into observable routines and cues, allowing employees and stakeholders to infer whether ethical AI is a credible organizational priority rather than an aspirational statement.

5.4.2. Strategy-to-Execution Linkage

To integrate SAFE-AI beyond HR-local governance, the framework should be positioned as a strategy-to-execution mechanism that links corporate ethical intent to operational routines. Examples are listed below:

- Corporate level (red lines and risk posture): define non-negotiable constraints for AI use in people-related decisions (e.g., nondiscrimination, privacy boundaries, accountable human authority) and specify tradeoff tolerances between efficiency gains and ethical exposure, recognizing that AI-enabled HR decisions shape reputation, regulatory scrutiny, and employee trust.
- Program level (gating and portfolio decisions): treat AI-enabled HR initiatives as managed programs with clear owners, budgets, and milestones, embedding SAFE-AI requirements into stage gates (approval to pilot, approval to scale, approval to institutionalize) so systems that cannot meet minimum ethical performance thresholds do not progress.
- Operating model level (rhythms, metrics, incentives): institutionalize SAFE-AI through recurring management cadences (planning, risk review, audit reporting), measurable indicators (e.g., validation completeness, adverse impact monitoring, incident rates, remediation lead time), and incentives that reward documentation quality, transparency, and corrective action rather than speed-only deployment.

This strategy-to-execution positioning is consistent with HR practices showing that technology creates sustainable value only when adoption is aligned with people-centered HR strategy and organizational readiness rather than implemented as a stand-alone technical upgrade (Nastase et al., 2025). This structure preserves SAFE-AI's behavioral premise: enterprise ethics becomes durable only when strategic commitments are resourced, monitored, and reinforced through routines that employees can observe and interpret as consistent organizational practice. By integrating ethical philosophies, social theories, and heuristics, HR practices can translate AI-ethics from principle into socially aware management routines, as summarized in Table 2. Building on this logic, the next subsection specifies how SAFE-AI can be embedded across corporate posture, program gating, and operating rhythms.

Table 2. Socially aware HR practices for AI adoption.

Stage	Objective	Heuristic	Action	Implementation	Example
Stage 1: Moving In (Initiation)	Establish a foundation for ethical AI adoption by considering stakeholder interests, ethical principles, and potential biases.	Ethical Philosophies Guiding HR Practices	Apply ethical philosophies such as consequentialism and deontology to balance duties and outcomes pre-AI adoption.	Assess organizational readiness. Evaluate potential employee impacts. Establish ethical guidelines from the outset.	Any AI adoption plan must comply with fundamental ethical principles such as fairness and non-discrimination to be considered.
Stage 2: Moving Through (Navigation)	Navigate the AI adoption and execution process by continuously integrating social feedback and adapting strategies to uphold ethical standards.	Leveraging Social Processing Feedback	Adopt social information processing theory to understand and respond to employee perceptions of the organization's ethical commitment.	Maintain transparency in decision-making. Gather ongoing feedback from employees and stakeholders. Adapt AI strategies to align with organizational values and ethics.	Assign weights to different decision dimensions (e.g., ethical compliance, operational efficiency) and calculate the total value to select the best alternative.
Stage 3: Moving Out (Culmination)	Ensure that ethical principles remain embedded in the organizational culture post-AI adoption and foster continuous ethical engagement.	Embedded Accountability for Ethical Principles	Develop mechanisms to uphold and monitor ethical standards within the organization.	Implement regular ethics training. Perform audits. Appoint ethics officers or committees for continuous ethical oversight.	Regularly review and adjust AI systems based on feedback from employees and stakeholders to ensure ongoing alignment with ethical principles.

5.4.3. Integrating SAFE-AI into Organizational Strategy

To move SAFE-AI beyond HR-local governance, it must function as a strategy-to-execution mechanism that links corporate ethical intent to observable operational routines. At the corporate strategy level, SAFE-AI defines the organization's AI ethics posture by specifying non-negotiable "red lines" for people-related decisions (e.g., nondiscrimination, privacy boundaries, accountable human authority) and clarifying the organization's tolerance for tradeoffs between efficiency gains and ethical exposure; this posture is strategically consequential because AI-enabled HR decisions shape legitimacy, regulatory scrutiny, litigation risk, and employee trust. At the program (business/functional) level, SAFE-AI should be embedded as a required capability in talent and AI strategy through stage gates for AI-enabled HR initiatives (approval to pilot, scale, institutionalize), such that systems that cannot meet minimum ethical performance thresholds (e.g., validated job relevance, monitored disparate outcomes, auditable decision authority) do not progress to broader deployment. At the operating model level, integration requires recurring management rhythms and measurable indicators, including explicit planning and resourcing, routine reporting on ethical incidents, drift, and remediation, and incentives that reward transparency, documentation quality, and corrective action rather than speed-only implementation. Under this treatment, ethical AI becomes a resourced and monitored operational commitment rather than a policy statement, consistent with SAFE-AI's behavioral premise that ethics is durable only when employees can observe consistent cues in decisions and accountability.

6. Implications for Practice and Research

In this section, we extend some of the key issues that are important for research and practice. Our discussion is based on the application of SAFE-AI for HR practices when organizations integrate AI technologies.

6.1. Implications for Practice

For practitioners, SAFE-AI translates AI-ethics from a principles statement into an operational program. First, organizations should establish governance minimums before scaling—documented intended use and decision authority (advisory vs. determinative), clear accountable owners across the socio-technical chain (HR, business, model/data, operations, legal/compliance, and IT/security), and defined privacy boundaries and data provenance requirements. Second, HR practice should invest in capability-building through targeted training for HR leaders and managers on ethical reasoning, bias awareness, explanation practices, and escalation protocols, supplemented by cross-functional education that aligns HR practice, security, and legal stakeholders on shared constraints and responsibilities. Third, ethical performance requires routine monitoring. Implement audit cadences for disparate outcomes and drift, track incidents and near-misses, and maintain remediation authority with documented learning loops so corrections are timely and visible. Fourth, organizations should build contestability and voice into HR workflows through accessible appeals mechanisms, data correction pathways, and psychologically safe reporting channels, signaling that feedback is consequential rather than performative. Fifth, data stewardship should be treated as an ethics requirement by controlling collection, access, retention, and sharing, recognizing that trust failures often arise from data practice weaknesses even when model logic is not the proximate cause. Finally, leaders should align incentives and performance systems with ethical enactment by rewarding documentation quality, transparency, and corrective action rather than speed-only implementation. These actions operationalize SAFE-AI's core premise—AI-ethics becomes durable only when commitments are resourced, monitored, and reinforced through routines that employees can observe and interpret as consistent organizational practice.

6.2. Implications for Research

SAFE-AI frames AI-ethics in HR practice as a socially enacted capability, which yields a testable agenda focused on implementation fidelity, cue-consistency mechanisms, and boundary conditions. The most immediate priorities are to operationalize SAFE-AI maturity indicators, test whether cue consistency mediates the relationship between formal ethics commitments and employee outcomes, and examine leadership as a moderating condition that determines whether ethical intent is enacted or symbolically complied with. We extend these implications into a focused program of empirical research in the next section, organized around measurement development, mechanism testing, stage-specific failure modes, and cross-context generalization.

Future Research Directions

This paper provides a conceptual results set—SAFE-AI's staged mechanism, heuristics, and implementation guidance—that can be advanced through empirical tests, measurement development, and boundary-condition analysis. Because SAFE-AI posits that AI-ethics becomes organizational behavior through socially interpreted cues, future research should examine both (a) enactment fidelity across organizations and (b) downstream effects on trust, perceived procedural justice, compliance behavior, and adverse outcomes. We outline four research directions that would extend SAFE-AI into an empirically supported program.

The first one is operationalizing SAFE-AI implementation fidelity and maturity. Develop a measurable fidelity construct by translating the stages (moving in, moving through, moving out) into observable indicators (e.g., intended use and decision authority, documentation completeness, audit cadence, remediation lead time), distinguishing symbolic compliance from durable institutionalization and enabling maturity profiling aligned to responsible AI governance frameworks. The second one is testing the cue-environment

mechanism. Examine whether cue consistency (e.g., decisions, explanations, consequences, feedback responsiveness) mediates the relationship between formal ethics policies and employee outcomes such as trust, willingness to speak up, and perceived fairness, consistent with social information processing logic. The third one is leadership as a moderator and causal pathway. Test leadership as a gating condition—whether SAFE-AI reduces perceived injustice and adverse outcomes only when leaders set non-negotiable constraints, resource validation and audits, and enforce consequences, using established ethical leadership measures and observable leader actions. And the fourth is stage-specific risk emergence and drift over time. Evaluate whether ethical failures vary systematically by stage (i.e., initiation: scoping/proxies/ownership; navigation: drift/workarounds/silent harm; culmination: decay/audit fatigue/normalization of deviance) and identify which controls are stage-specific versus generalizable using longitudinal case studies, incident analyses, and audit-log research. Collectively, these directions position SAFE-AI as a cumulative research program that links a measurement model (implementation fidelity) to a behavioral mechanism (cue environments and interpretation) and to testable propositions about when ethical AI becomes durable in practice.

7. Limitations

Several limitations should be acknowledged. First, SAFE-AI positions heuristics as a practical translation layer, but heuristics can introduce systematic error when they become over-relied upon or are applied without structured reflection; thus, the framework requires guardrails (e.g., documentation, review, and feedback loops) to mitigate heuristic-driven bias (Gigerenzer & Gaissmaier, 2011). Second, SAFE-AI is a conceptual model grounded in ethical philosophies and social information processing theory; while this foundation offers explanatory coherence, it cannot fully capture the full complexity and rapid contextual shifts in AI-ethics in HR, nor does it substitute for technical risk assessment and domain-specific controls. Third, implementation feasibility varies: operationalizing SAFE-AI entails governance capacity, training, monitoring, and remediation resources that may be difficult for smaller organizations or low-maturity environments to sustain (Davies & Brooks, 2017; A. Taylor & Taylor, 2014). Fourth, transferability is constrained by variability in institutional and cultural contexts; norms regarding fairness, privacy, authority, and voice shape how cues are interpreted and which accountability practices are trusted, requiring calibration rather than uniform replication (T. M. Jones et al., 2007). Finally, because AI technologies and regulatory expectations evolve, SAFE-AI will require periodic updating of its stage-specific practices and heuristics to remain aligned with emerging risks and changing socio-technical conditions.

8. Conclusions

Recent research underscores that integrating AI into organizational decision processes introduces ethical risks with real consequences for individuals, organizations, and society, particularly when AI-mediated judgments shape opportunity, evaluation, privacy, and accountability. HR practices are often positioned as the operational owners of these ethical obligations, yet much of the existing AI-ethics guidance remains primarily prescriptive, offering principles and governance recommendations without specifying how those commitments become durable day-to-day practice. This gap persists in part because ethical breakdowns in AI adoption rarely arise from a lack of stated principles; they arise when organizations fail to translate ethical intent into routines that remain credible under ambiguity, time pressure, delegation, and cross-functional handoffs.

SAFE-AI responds to this problem by treating AI-ethics as a behavioral and organizational accomplishment. Grounded in ethical philosophies and social information

processing, the framework specifies a translation pathway through which ethical expectations become socially interpreted and enacted through observable cues in decisions, explanations, accountability, and corrective action. The staged logic of moving in, moving through, and moving out, paired with implementation heuristics, provides HR leaders and organizations with a practical means to convert ethical commitments into workflow requirements, feedback mechanisms, and institutionalized learning loops that can be taught, monitored, and audited. In doing so, SAFE-AI reframes ethical AI technology in HR practice not as a compliance artifact or technical add-on but as a durable capability embedded in organizational culture and governance.

Future work should empirically test SAFE-AI's implementation fidelity, cue-consistency mechanism, and leadership contingencies, and should refine stage-specific controls across sectors and cultural contexts. For practice, the implication is immediate; organizations that seek the benefits of AI technology in HR practice must invest in the behavioral and governance conditions that make ethical intent visible and enforceable. AI-ethics becomes real when employees can consistently observe that the organization means what it says—across time, pressure, and tradeoffs.

Author Contributions: Conceptualization, R.E.C. and S.R.P.; methodology, investigation, writing—original draft preparation, D.H. and R.M.; writing—review and editing, R.E.C., D.H., S.R.P. and R.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Umar, A. M., Linus, O. U., & Kiru, M. U. (2019). Comprehensive review of artificial neural network applications to pattern recognition. *IEEE Access*, 7, 158820–158846. [CrossRef]
- Adkins, B. (2017). *A guide to ethics and moral philosophy*. Edinburgh University Press.
- Aldawood, H., Alashoor, T., & Skinner, G. (2020). Does awareness of social engineering make employees more secure? *International Journal of Computer Applications*, 177(38), 45–49. [CrossRef]
- Alizadeh, A., & Kurian, D. (2024). Introduction to ethical theories. In D. F. Russ-Eft, & A. Alizadeh (Eds.), *Ethics and human resource development: Societal and organizational contexts* (pp. 13–28). Springer International Publishing.
- Anderson, M., Anderson, S. L., & Armen, C. (2006). An approach to computing ethics. *IEEE Intelligent Systems*, 21(4), 56–63. Available online: www.computer.org/intelligent (accessed on 26 May 2025). [CrossRef]
- Andersson, L., Eriksson, J., Stillesjö, S., Juslin, P., Nyberg, L., & Wirebring, L. K. (2020). Neurocognitive processes underlying heuristic and normative probability judgments. *Cognition*, 196, 104153. [CrossRef] [PubMed]
- Ardichvili, A. (2022). The impact of artificial intelligence on expertise development: Implications for HRD. *Advances in Developing Human Resources*, 24(2), 78–98. [CrossRef]
- Ashok, M., Madan, R., Joha, A., & Sivarajah, U. (2022). Ethical framework for artificial intelligence and digital technologies. *International Journal of Information Management*, 62, 102433. [CrossRef]
- Augusto, L. M. (2021). From symbols to knowledge systems: A. Newell and H. A. Simon's contribution to symbolic AI. *Journal of Knowledge Structures and Systems*, 2(1), 29–62. Available online: <https://philpapers.org/rec/AUGFST-2> (accessed on 19 May 2025).
- Bandura, A. (1976). *Social learning theory*. Prentice Hall.
- Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes*, 50(2), 248–287. [CrossRef]

- Bangura, S., Duma, P. T., & Mthembu, N. A. (2025). Ethical considerations of implementing artificial intelligence in human resource management: A review. *International Journal of Business Ecosystem & Strategy*, 7(5), 274–281. [CrossRef]
- Bankins, S. (2021). The ethical use of artificial intelligence in human resource management: A decision-making framework. *Ethics and Information Technology*, 23(4), 841–854. [CrossRef]
- Banks, S. (2020). *Ethics and values in social work*. Bloomsbury Publishing.
- Beauchamp, T. L., & Bowie, N. E. (1979). *Ethical theory and business*. Prentice Hall.
- Beeri, I., Dayan, R., Vigoda-Gadot, E., & Werner, S. B. (2013). Advancing ethics in public organizations: The impact of an ethics program on employees' perceptions and behaviors in a regional council. *Journal of Business Ethics*, 112, 59–78. [CrossRef]
- Ben-Sasson, H., & Greenberg, R. (2023, September 18). 38TB of data accidentally exposed by Microsoft AI researchers. *Wiz.io*. Available online: <https://www.wiz.io/blog/38-terabytes-of-private-data-accidentally-exposed-by-microsoft-ai-researchers> (accessed on 19 May 2025).
- Bingham, C. B., & Eisenhardt, K. M. (2011). Rational heuristics: The 'simple rules' that strategists learn from process experience. *Strategic Management Journal*, 32(13), 1437–1464. [CrossRef]
- Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018, April 21–26). *It's reducing a human being to a percentage: Perceptions of justice in algorithmic decisions*. 2018 CHI Conference on Human Factors in Computing Systems (pp. 1–14), Montreal, QC, Canada. [CrossRef]
- Boekhorst, J. A. (2015). The role of authentic leadership in fostering workplace inclusion: A social information processing perspective. *Human Resource Management*, 54(2), 241–264. [CrossRef]
- Bogen, M. (2019, May 6). All the ways hiring algorithms can introduce bias. *Harvard Business Review*. Available online: <https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias> (accessed on 25 May 2025).
- Bordage, G. (2009). Conceptual frameworks to illuminate and magnify. *Medical Education*, 43(4), 312–319. [CrossRef]
- Brase, G. L. (2014). Behavioral science integration: A practical framework of multi-level converging evidence for behavioral science theories. *New Ideas in Psychology*, 33, 8–20. [CrossRef]
- Brey, P., & Dainow, B. (2023). Ethics by design for artificial intelligence. *AI and Ethics*, 4, 1265–1277. [CrossRef]
- Calabretta, G., Gemser, G., & Wijnberg, N. M. (2017). The interplay between intuition and rationality in strategic decision making: A paradox perspective. *Organization Studies*, 38(3–4), 365–401. [CrossRef]
- Caldwell, C., & Karri, R. (2005). Organizational governance and ethical systems: A covenantal approach to building trust. *Journal of Business Ethics*, 58, 249–259. [CrossRef]
- Carpenter, R. E. (2021). Learning as cognition: A developmental process for organizational learning. *Development and Learning in Organizations: An International Journal*, 35(6), 18–21. [CrossRef]
- Chandra, S., Shirish, A., & Srivastava, S. C. (2022). To be or not to be... human? Theorizing the role of human-like competencies in conversational artificial intelligence agents. *Journal of Management Information Systems*, 39(4), 969–1005. [CrossRef]
- Chowdhury, S., Dey, P., Joel-Edgar, S., Bhattacharya, S., Rodriguez-Espindola, O., Abadie, A., & Truong, L. (2023). Unlocking the value of artificial intelligence in human resource management through AI capability framework. *Human Resource Management Review*, 33(1), 100899. [CrossRef]
- Chuang, S., & Graham, C. M. (2018). Embracing the sobering reality of technological influences on jobs, employment and human resource development: A systematic literature review. *European Journal of Training and Development*, 42(7–8), 400–416. [CrossRef]
- Claure, H., Kim, S., Kizilcec, R. F., & Jung, M. (2023). The social consequences of machine allocation behavior: Fairness, interpersonal perceptions and performance. *Computers in Human Behavior*, 146, 107628. [CrossRef]
- Clore, G. L., Schwarz, N., & Conway, M. (2014). Affective causes and consequences of social information processing. In R. S. Wyer Jr., & T. K. Srull (Eds.), *Handbook of social cognition* (pp. 323–418). Psychology Press.
- Crawshaw, J. R., Cropanzano, R., Bell, C. M., & Nadisic, T. (2013). Organizational justice: New insights from behavioural ethics. *Human Relations*, 66(7), 885–904. [CrossRef]
- Danysz, K., Cicirello, S., Mingle, E., Assuncao, B., Tetarenko, N., Mockute, R., & Desai, S. (2019). Artificial intelligence and the future of the drug safety professional. *Drug Safety*, 42, 491–497. [CrossRef]
- Dastin, J. (2018, October 11). Insight—Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters World*. Available online: <https://www.reuters.com/article/idUSKCN1MK0AG/> (accessed on 25 May 2025).
- Davies, G. B., & Brooks, P. (2017). Practical challenges of implementing behavioral finance: Reflections from the field. In H. K. Baker, G. Filbeck, & V. Ricciardi (Eds.), *Financial behavior: Players, services, products, and markets* (pp. 542–560). Oxford University Press.
- Elia, J. (2009). Transparency rights, technology, and trust. *Ethics and Information Technology*, 11, 145–153. [CrossRef]
- Foote, M. F., & Ruona, W. E. (2008). Institutionalizing ethics: A synthesis of frameworks and the implications for HRD. *Human Resource Development Review*, 7(3), 292–308. [CrossRef]
- Francolini, G., Desideri, I., Stocchi, G., Salvestrini, V., Ciccone, L. P., Garlatti, P., & Livi, L. (2020). Artificial intelligence in radiotherapy: State of the art and future directions. *Medical Oncology*, 37, 50. [CrossRef]

- Fryer, M. (2018). HRM: An ethical perspective. In D. G. Collings, G. T. Wood, & L. T. Szamosi (Eds.), *Human resource management* (pp. 98–116). Routledge.
- Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science*, 3(1), 20–29. [CrossRef]
- Gigerenzer, G. (2018). The bias bias in behavioral economics. *Review of Behavioral Economics*, 5(3–4), 303–336. [CrossRef]
- Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, 62, 451–482. [CrossRef]
- Gigerenzer, G., Reb, J., & Luan, S. (2022). Smart heuristics for individuals, teams, and organizations. *Annual Review of Organizational Psychology and Organizational Behavior*, 9, 171–198. [CrossRef]
- Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2), 627–660. [CrossRef]
- Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, 109(1), 75–90. [CrossRef] [PubMed]
- Goldstein, D. G., & Gigerenzer, G. (2008). The recognition heuristic and the less-is-more effect. *Handbook of Experimental Economics Results*, 1, 987–992. [CrossRef]
- Gutierrez, G. (2020). Artificial intelligence in the intensive care unit. In *Annual update in intensive care and emergency medicine 2020* (pp. 667–681). Springer. [CrossRef]
- Gürses, S., Troncoso, C., & Diaz, C. (2011). Engineering privacy by design. *Computers, Privacy & Data Protection*, 14(3), 25. Available online: <https://software.imdea.org/carmela.troncoso/papers/Gurses-CPDP11.pdf> (accessed on 6 June 2025).
- Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4), 55–66. [CrossRef]
- Han, J. W., Hoe, O. J., Wing, J. S., & Brohi, S. N. (2017, December 5–7). *A conceptual security approach with awareness strategy and implementation policy to eliminate ransomware*. 2017 International Conference on Computer Science and Artificial Intelligence (pp. 222–226), Jakarta, Indonesia. [CrossRef]
- Hatcher, T., & Aragon, S. R. (2000). Rationale for and development of a standard on ethics and integrity for international HRD research and practice. *Human Resource Development International*, 3(2), 207–219. [CrossRef]
- Heldal, F., Sjøvold, E., & Stålsett, K. (2020). Shared cognition in intercultural teams: Collaborating without understanding each other. *Team Performance Management: An International Journal*, 26(3/4), 211–226. [CrossRef]
- Hendrycks, D., Burns, C., Basart, S., Critch, A. C., Li, J. L., Song, D., & Steinhardt, J. (2021, May 3–7). *Aligning AI with shared human values* [Poster]. International Conference on Learning Representations, Vienna, Austria.
- Hertwig, R., & Hoffrage, U. (2013). *Simple heuristics in a social world*. Oxford University Press. [CrossRef]
- Hesselbarth, I., Alnoor, A., & Tiberius, V. (2023). Behavioral strategy: A systematic literature review and research framework. *Management Decision*, 61(9), 2740–2756. [CrossRef]
- Hibbert, P., & Cunliffe, A. (2015). Responsible management: Engaging moral reflexive practice through threshold concepts. *Journal of Business Ethics*, 127, 177–188. [CrossRef]
- Hjeij, M., & Vilks, A. (2023). A brief history of heuristics: How did research on heuristics evolve? *Humanities and Social Sciences Communications*, 10(1), 64. [CrossRef]
- Ho, M. K., & Griffiths, T. L. (2022). Cognitive science as a source of forward and inverse models of human decisions for robotics and control. *Annual Review of Control, Robotics, and Autonomous Systems*, 5, 33–53. [CrossRef]
- Hu, X. J., Pawirosetiko, J. S., Santuzzi, A. M., & Barber, L. K. (2024). Does your job shape your experience or interpretation of workplace telepressure? Exploring measurement invariance across occupational characteristics. *Computers in Human Behavior Reports*, 14, 100426. [CrossRef]
- Hunkenschroer, A. L., & Luetge, C. (2022). Ethics of AI-enabled recruiting and selection: A review and research agenda. *Journal of Business Ethics*, 178(4), 977–1007. [CrossRef]
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI-ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. [CrossRef]
- Jones, M., Butler, D., & Plenert, G. (2022). *Transform behaviors, transform results! Identifying and using behavioral indicators to drive sustainable change and improvement*. Productivity Press.
- Jones, T. M., Felps, W., & Bigley, G. A. (2007). Ethical theory and stakeholder-related decisions: The role of stakeholder culture. *Academy of Management Review*, 32(1), 137–155. [CrossRef]
- Krakowski, S., Luger, J., & Raisch, S. (2023). Artificial intelligence and the changing sources of competitive advantage. *Strategic Management Journal*, 44(6), 1425–1452. [CrossRef]
- Kurniawan, B., Marnis, Samsir, & Jahrizal. (2025). A conceptual framework for sustainable human resource management: Integrating green practices, ethical leadership, and digital resilience to advance the SDGs. *Sustainability*, 17(21), 9904. [CrossRef]
- Lakshmanan, R. (2023, September 19). *Microsoft AI researchers accidentally expose 38 terabytes of confidential data*. The Hacker News. Available online: <https://thehackernews.com/2023/09/microsoft-ai-researchers-accidentally.html> (accessed on 2 February 2026).

- Lavanchy, M. (2018, November 1). Amazon's sexist hiring algorithm could still be better than a human. *Phys.org*. Available online: <https://phys.org/news/2018-11-amazon-sexist-hiring-algorithm-human.html> (accessed on 8 June 2025).
- Ledro, C., Nosella, A., & Dalla Pozza, I. (2023). Integration of AI in CRM: Challenges and guidelines. *Journal of Open Innovation: Technology, Market, and Complexity*, 9(4), 100151. [CrossRef]
- Lefkowitz, J. (2023). *Values and ethics of industrial-organizational psychology*. Routledge.
- Loi, M. (2020). *People analytics must benefit the people. An ethical analysis of data-driven algorithmic systems in human resources management* (pp. 1–56). Algorithmwatch. Available online: https://algorithmwatch.org/de/wp-content/uploads/2020/03/AlgorithmWatch_AutoHR_Study_Ethics_Loi_2020.pdf (accessed on 4 October 2025).
- Luan, S., Reb, J., & Gigerenzer, G. (2019). Ecological rationality: Fast-and-frugal heuristics for managerial decision making under uncertainty. *Academy of Management Journal*, 62(6), 1735–1759. [CrossRef]
- Manganello, F., Nico, A., Ragusa, M., & Boccuzzi, G. (2025). Testing the applicability of a governance checklist for high-risk AI-based learning outcome assessment in Italian universities under the EU AI act annex III. *Frontiers in Artificial Intelligence*, 8, 1718613. [CrossRef]
- Mattison, M. (2000). Ethical decision making: The person in the process. *Social Work*, 45(3), 201–212. [CrossRef] [PubMed]
- McGuire, D., Germain, M. L., & Reynolds, K. (2021). Reshaping HRD in light of the COVID-19 pandemic: An ethics of care approach. *Advances in Developing Human Resources*, 23(1), 26–40. [CrossRef]
- McWhorter, R. R. (2023). Virtual human resource development: Definitions, challenges, and opportunities. *Human Resource Development Review*, 22(4), 582–601. [CrossRef]
- Melé, D. (2012). The firm as a 'community of persons': A pillar of humanistic business ethos. *Journal of Business Ethics*, 106, 89–101. [CrossRef]
- Melé, D. (2019). *Business ethics in action: Managing human excellence in organizations*. Bloomsbury Publishing.
- Microsoft Corporation. (n.d.-a). Azure. Microsoft. Available online: <https://azure.microsoft.com> (accessed on 25 May 2025).
- Microsoft Corporation. (n.d.-b). Coordinated vulnerability disclosure. Microsoft Security Response Center. Available online: <https://www.microsoft.com/en-us/msrc/cvd> (accessed on 25 May 2025).
- Microsoft Corporation. (n.d.-c). Microsoft mitigated exposure of internal information in a storage account due to overly-permissive SAS token. MSRC, Security Research & Defense. Available online: <https://msrc.microsoft.com/blog/2023/09/microsoft-mitigated-exposure-of-internal-information-in-a-storage-account-due-to-overly-permissive-sas-token/> (accessed on 25 May 2025).
- Mohammed, A. Q. (2019). HR analytics: A modern tool in HR for predictive decision making. *Journal of Management*, 6(3), 51–63. Available online: <https://ssrn.com/abstract=3525328> (accessed on 2 June 2025). [CrossRef]
- Moore, R. (2021). The cultural evolution of mind-modelling. *Synthese*, 199(1), 1751–1776. [CrossRef]
- Nastase, C., Adomnitei, A., & Apetri, A. (2025). Strategic human resource management in the digital era: Technology, transformation, and sustainable advantage. *Merits*, 5(4), 23. [CrossRef]
- National Institute of Standards and Technology. (2024). *Artificial intelligence risk management framework: Generative artificial intelligence profile (NIST AI 600-1)*. U.S. Department of Commerce. [CrossRef]
- Njoto, S., Cheong, M., Lederman, R., McLoughney, A., Ruppner, L., & Wirth, A. (2022, November 10–12). *Gender bias in AI recruitment systems: A sociological-and data science-based case study*. 2022 IEEE International Symposium on Technology and Society, Hong Kong, China. [CrossRef]
- Oliveira, J., Murphy, T., Vaughn, G., Elfahim, S., & Carpenter, R. E. (2024). Exploring the adoption phenomenon of artificial intelligence by doctoral students within doctoral education. *New Horizons in Adult Education and Human Resource Development*, 36(4), 248–262. [CrossRef]
- Olson, M. H., & Ramírez, J. J. (2020). *An introduction to theories of learning*. Routledge.
- Ortega-Bolaños, R., Bernal-Salcedo, J., & Ortiz, M. G. (2024). Applying the ethics of AI: A systematic review of tools for developing and assessing AI-based systems. *Artificial Intelligence Review*, 57, 110. [CrossRef]
- Pleskac, T. J., & Hertwig, R. (2014). Ecologically rational choice and the structure of the environment. *Journal of Experimental Psychology*, 143(5), 2000–2019. [CrossRef] [PubMed]
- Prikshat, V., Malik, A., & Budhwar, P. (2023). AI-augmented HRM: Antecedents, assimilation and multilevel consequences. *Human Resource Management Review*, 33(1), 100860. [CrossRef]
- Raihan, A. (2023). A comprehensive review of artificial intelligence and machine learning applications in energy sector. *Journal of Technology Innovations and Energy*, 2(4), 1–26. [CrossRef]
- Reese, S. D. (2023). Writing the conceptual article: A practical guide. *Digital Journalism*, 11(7), 1195–1210. [CrossRef]
- Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27, 169–192. [CrossRef]
- Rodgers, W., Murray, J. M., Stefanidis, A., Degbey, W. Y., & Tarba, S. Y. (2023). An artificial intelligence algorithmic approach to ethical decision-making in human resource management processes. *Human Resource Management Review*, 33(1), 100925. [CrossRef]
- Royakkers, L., Timmer, J., Kool, L., & Van Est, R. (2018). Societal and ethical issues of digitization. *Ethics and Information Technology*, 20, 127–142. [CrossRef]

- Salancik, G. R., & Pfeffer, J. (1978). A social information processing approach to job attitudes and task design. *Administrative Science Quarterly*, 22, 224–253. [\[CrossRef\]](#)
- Sales, B. D., & Lavin, M. (2000). Identifying conflicts of interests and resolving ethical dilemmas. In B. D. Sales, & S. Folkman (Eds.), *Ethics in research with human participants* (pp. 109–128). American Psychological Association.
- Sanderson, C., Douglas, D., Lu, Q., Schleiger, E., Whittle, J., Lacey, J., & Hansen, D. (2023). AI-ethics principles in practice: Perspectives of designers and developers. *IEEE Transactions on Technology and Society*, 4(2), 171–187. [\[CrossRef\]](#)
- Sane, M. G., Kumar, V. R., & Ger, A. (2025). Enhancing AI adoption efficiency in enterprises: The role of infrastructure readiness and decision-making. *Journal of Marketing & Social Research*, 2, 678–684. [\[CrossRef\]](#)
- Scherer, M. U. (2015). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harvard Journal of Law and Technology*, 29(2), 353–398. [\[CrossRef\]](#)
- Schoenherr, J. R. (2022). *Ethical artificial intelligence from popular to cognitive science: Trust in the age of entanglement*. Routledge.
- Schraw, G., Dunkle, M. E., & Bendixen, L. D. (1995). Cognitive processes in well-defined and ill-defined problem solving. *Applied Cognitive Psychology*, 9(6), 523–538. [\[CrossRef\]](#)
- Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin*, 134(2), 207–222. [\[CrossRef\]](#)
- Shefrin, H., & Statman, M. (2003). The contributions of Daniel Kahneman and Amos Tversky. *The Journal of Behavioral Finance*, 4(2), 54–58. [\[CrossRef\]](#)
- Shneiderman, B. (2020). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems*, 10(4), 1–31. [\[CrossRef\]](#)
- Solomon, R. C. (1994). The corporation as community: A reply to Ed Hartman. *Business Ethics Quarterly*, 4(3), 271–285. [\[CrossRef\]](#)
- Stilgoe, J., Owen, R., & Macnaghten, P. (2020). Developing a framework for responsible innovation. In A. Maynard, & J. Stilgoe (Eds.), *The ethics of nanotechnology, geoengineering, and clean energy* (pp. 347–359). Routledge.
- Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*, 61(4), 15–42. [\[CrossRef\]](#)
- Tanantong, T., & Wongras, P. (2024). A UTAUT-based framework for analyzing users' intention to adopt artificial intelligence in human resource recruitment: A case study of Thailand. *Systems*, 12(1), 28. [\[CrossRef\]](#)
- Taylor, A., & Taylor, M. (2014). Factors influencing effective implementation of performance measurement systems in small and medium-sized enterprises and large firms: A perspective from contingency theory. *International Journal of Production Research*, 52(3), 847–866. [\[CrossRef\]](#)
- Taylor, P. L. (2020). Dispatch priming and the police decision to use deadly force. *Police Quarterly*, 23(3), 311–332. [\[CrossRef\]](#)
- Tegarden, D. P., & Sheetz, S. D. (2003). Group cognitive mapping: A methodology and system for capturing and evaluating managerial and organizational cognition. *Omega*, 31(2), 113–125. [\[CrossRef\]](#)
- Textor, C., Zhang, R., Lopez, J., Schelble, B. G., McNeese, N. J., Freeman, G., & de Visser, E. J. (2022). Exploring the relationship between ethics and trust in human–artificial intelligence teaming: A mixed methods approach. *Journal of Cognitive Engineering and Decision Making*, 16(4), 252–281. [\[CrossRef\]](#)
- Thibault, P. (2004). *Agency and consciousness in discourse: Self-other dynamics as a complex system*. A&C Black.
- Thiel, C. E., Bagdasarov, Z., Harkrider, L., Johnson, J. F., & Mumford, M. D. (2012). Leader ethical decision-making in organizations: Strategies for sensemaking. *Journal of Business Ethics*, 107, 49–64. [\[CrossRef\]](#)
- Tudor, K. (2023). Critical heuristics in psychotherapy research: From 'I-who-feels' to 'we-who-care—And act'. In K. Tudor, & J. Wyatt (Eds.), *Qualitative research approaches for psychotherapy* (pp. 115–132). Routledge.
- Tursunbayeva, A., Di Lauro, S., & Pagliari, C. (2018). People analytics—A scoping review of conceptual boundaries and value propositions. *International Journal of Information Management*, 43, 224–247. [\[CrossRef\]](#)
- Walther, J. B. (2008). Social information processing theory. In D. O. Braithwaite, & P. Schrodt (Eds.), *Engaging theories in interpersonal communication: Multiple perspectives* (pp. 391–452). Routledge.
- Weiskopf, R., & Hansen, H. K. (2023). Algorithmic governmentality and the space of ethics: Examples from 'people analytics'. *Human Relations*, 76(3), 483–506. [\[CrossRef\]](#)
- Welch, R. V., & Dixon, J. R. (1994). Guiding conceptual design through behavioral reasoning. *Research in Engineering Design*, 6, 169–188. [\[CrossRef\]](#)
- Whelehan, D. F., Conlon, K. C., & Ridgway, P. F. (2020). Medicine and heuristics: Cognitive biases and medical decision-making. *Irish Journal of Medical Science*, 189, 1477–1484. [\[CrossRef\]](#) [\[PubMed\]](#)
- Whittlestone, J., & Clarke, S. (2022). AI challenges for society and ethics. In J. B. Bullock, Y. Chen, J. Himmelreich, V. M. Hudson, A. Korinek, M. M. Young, & B. Zhang (Eds.), *The oxford handbook of AI governance* (pp. 45–64). Oxford University Press. [\[CrossRef\]](#)
- Worden, S. (2003). The role of integrity as a mediator in strategic leadership: A recipe for reputational capital. *Journal of Business Ethics*, 46(1), 31–44. [\[CrossRef\]](#)

- Yorks, L., Rotatori, D., Sung, S., & Justice, S. (2020). Workplace reflection in the age of AI: Materiality, technology, and machines. *Advances in Developing Human Resources*, 22(3), 308–319. [[CrossRef](#)]
- Yunos, Z., Ab Hamid, R. S., & Ahmad, M. (2016, July 13–15). *Development of a cyber security awareness strategy using focus group discussion*. 2016 SAI Computing Conference (pp. 1063–1067), London, UK. [[CrossRef](#)]
- Zhu, L., Xu, X., Lu, Q., Governatori, G., & Whittle, J. (2022). AI and ethics—Operationalizing responsible AI. In F. Chen, & J. Zhou (Eds.), *Humanity driven AI: Productivity, well-being, sustainability and partnership* (pp. 15–33). Springer. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.