Ethics and Information Technology (2021) 23:841–854
https://doi.org/10.1007/s10676-021-09619-6

ORIGINAL PAPER

The ethical use of artificial intelligence in human resource management: a decision-making framework

Sarah Bankins[1]

Accepted: 25 October 2021 / Published online: 25 November 2021

Abstract
Artificial intelligence (AI) is increasingly inputting into various human resource management (HRM) functions, such as sourcing job applicants and selecting staff, allocating work, and offering personalized career coaching. While the use of AI for such tasks can offer many benefits, evidence suggests that without careful and deliberate implementation its use also has the potential to generate significant harms. This raises several ethical concerns regarding the appropriateness of AI deployment to domains such as HRM, which directly deal with managing sometimes sensitive aspects of individuals' employment lifecycles. However, research at the intersection of HRM and technology continues to largely center on examining what AI can be used for, rather than focusing on the salient factors relevant to its *ethical use* and examining how to effectively engage human workers in its use. Conversely, the ethical AI literature offers excellent guiding principles for AI implementation broadly, but there remains much scope to explore how these principles can be enacted in specific contexts-of-use. By drawing on ethical AI and task-technology fit literature, this paper constructs a decision-making framework to support the ethical deployment of AI for HRM and guide determinations of the optimal mix of human and machine involvement for different HRM tasks. Doing so supports the deployment of AI for the betterment of work and workers and generates both scholarly and practical outcomes.

Keywords Artificial intelligence · Human resource management · Ethical AI · HRM and technology · Ethical task-technology fit · Human control

> It is only just dawning on us … that good workers can lose their jobs because of a poorly written algorithm. That AI can create new and hidden forms of discrimination. Edward Santow, Human Rights Commissioner (2016–2021), Australia

Introduction

The use of artificially intelligent technologies is influencing many aspects of our working lives. Artificial intelligence (AI) encompasses various interrelated technologies often underpinned by machine learning algorithms, whereby AI achieves set objectives via supervised (with human guidance) or unsupervised (machine autonomous) learning through analyzing large datasets (Walsh et al., 2019). Although most current forms of AI perform a restricted range of single domain functions ('narrow AI'), they can undertake some tasks better than humans, such as pattern recognition.

AI is increasingly automating and supporting various human resource management (HRM) functions, such as through: scheduling work (e.g., Uber, see Lee et al., 2015); screening job applicants' resumes and assessing video applications via verbal and body language analysis (e.g., Unilever, see Marr, 2018; Strohmeier & Piazza, 2015; Jia et al., 2018); and offering personalized career coaching (e.g., IBM, see Guenole & Feinzig, 2018b). This can generate many benefits through enhancing evidence-based decision making (Colson, 2019), improving the depth, diversity, and quality of applicant pools (Marr, 2018), and deepening personalization of HRM services (Guenole & Feinzig, 2018a).

However, as the opening quote shows, such benefits are not assured. Increasing AI use is "fueling anxieties and ethical concerns" regarding its trustworthiness (OECD, 2019, p. 3), particularly when its use impacts people's

✉ Sarah Bankins
sarah.bankins@mq.edu.au

[1] Macquarie Business School, Macquarie University, North Ryde Campus, Sydney, NSW 2109, Australia

livelihoods. This makes the appropriateness of applying AI to HRM a significant issue (Scholz, 2019; Tambe et al., 2019), as its use has already generated racially- and gender-biased outcomes in recruitment (Dastin, 2018), resulted in breaches of employee data (Starner, 2019), and compromised fair and just employee outcomes (for examples see Robert et al., 2020). The complexity of quantifying some human performance metrics (what makes a "good employee"?), the 'small' (rather than 'large') nature of HRM datasets, and imperatives for fairness and accountability in decision making all characterize HRM activities and so challenge the universal applicability of AI in this domain (Tambe et al., 2019, p. 17).

Despite these concerns, conceptual and practical guidance on how to maximize benefits and minimize harms when applying AI to HRM remain rare. While research in HRM and technology has importantly examined how AI can be applied to specific functions (Strohmeier & Piazza, 2015) and how big data use may impact HRM (Scholz, 2019), this literature is yet to fully grapple with the *ethical implications* of AI use for people management (for exceptions see Tambe et al., 2019; Robert et al., 2020). This leaves scope to look beyond what HRM functions AI can be used for, to instead detail salient factors relevant to the *ethical* use of AI in a HRM context.

To this end, I construct a decision-making framework to support the ethical deployment of AI for people management. This extends both the HRM literature, by examining the ethical implications of AI use in this function, and the ethical AI literature, by applying its principles to a specific context-of-use. I first draw on the ethical AI literature to overview the emerging global consensus toward five ethical principles underpinning the use of AI across contexts. The decision-making framework is then constructed to operationalize these principles in the HRM domain. To do this, I draw on the task-technology fit literature and conceptualize the notion of *ethical task-technology fit*. The framework centers on assessing: (1) particular technology characteristics that I term 'data and AI sensitivities'; and (2) what the AI is deployed to do through assessing task characteristics that I term 'task sensitivities'. Drawing on research on human control in automated systems, I then illustrate how variations in these sensitivities will drive differing levels of human control and involvement alongside the AI to help generate and sustain its ethical use. The framework does not aim to be highly prescriptive or universal, given the complex socio-technical context into which AI is often deployed and the evolving nature of the technology, but instead provides guidance on some key technology and task indicators to assess ethical task-technology fit. I conclude by applying the framework and discussing its theoretical and practical implications.

Ethical artificial intelligence: an overview

Ethical AI broadly refers to "the fair and just development, use, and management of AI technologies" (Bankins & Formosa, 2021, p. 60). Its study spans disciplines and examines issues ranging from identifying governing principles for AI development (Floridi et al., 2018), examining the legal and accountability implications of AI decision making (Doshi-Velez et al., 2017), the ethical implications of social robot use (Stahl & Coeckelbergh, 2016), and the efficacy of granting moral rights to AI agents (Formosa & Ryan, 2020). The field generally adopts a socio-technical stance (i.e., that social and technological contexts must be examined in tandem, Selbst et al., 2019) and takes the view that AI should be developed and applied by, with, and for people to aid the betterment of humans (e.g., Floridi et al., 2018).

Such goals have resulted in researchers, and indeed national governments, developing many principlist approaches to guide AI development and use. Recent reviews (e.g., Floridi & Cowls, 2019; Hagendorff, 2020; Jobin et al., 2019) suggest a significant overlap exists across these approaches that supports an emerging global consensus on such ethical principles. One such framework reflecting this emerging consensus, and on which I focus, is the AI4People framework (see Floridi et al., 2018), which synthesizes five ethical AI principles: beneficence; non-maleficence; autonomy; justice; and explicability. *Beneficence* means AI should benefit "people and the planet" through fostering human wellbeing and dignity and environmental sustainability (Floridi et al., 2018, p. 6). *Non-maleficence* refers to the development and use of AI that does not harm individuals and preserves individuals' privacy. This generally requires meeting the *autonomy* principle, by appropriately balancing decision making between humans, AI, or both, and that humans retain power to change such delegations (Floridi et al., 2018). The *justice* principle requires AI to promote fair and just outcomes, such as through eliminating bias and fostering diversity. *Explicability* facilitates accomplishment of all other principles by requiring AI to be intelligible (i.e., humans have some understanding of its operations) and accountable (i.e., responsibility for its use is clear) (Floridi et al., 2018).

While ethical AI principles are critical for broadly guiding technology development and use, they must also be operationalized for application in specific contexts (Aizenberg & van den Hoven, 2020). To help do this, I now turn to developing a decision-making framework to support the ethical application of AI in a HRM context.

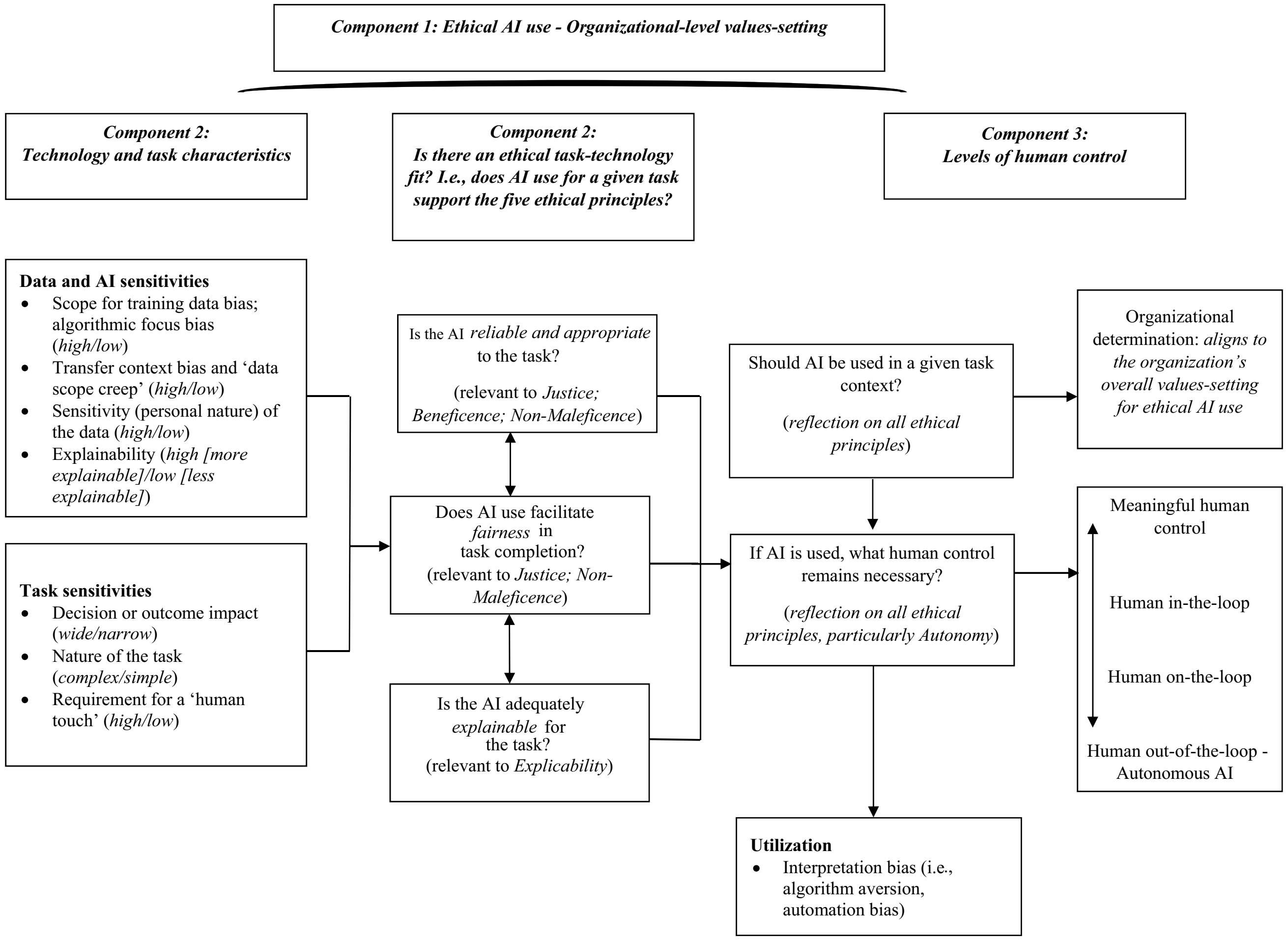
Component 1: Ethical AI use - Organizational-level values-setting
Component 2: Technology and task characteristics
Component 2: Is there an ethical task-technology fit? I.e., does AI use for a given task support the five ethical principles?
Component 3: Levels of human control
Data and AI sensitivities
Scope for training data bias; algorithmic focus bias (high/low)
Transfer context bias and 'data scope creep' (high/low)
Sensitivity (personal nature) of the data (high/low)
Explainability (high [more explainable]/low [less explainable])
Task sensitivities
Decision or outcome impact (wide/narrow)
Nature of the task (complex/simple)
Requirement for a 'human touch' (high/low)
Is the AI reliable and appropriate to the task?
(relevant to Justice; Beneficence; Non-Maleficence)
Does AI use facilitate fairness in task completion?
(relevant to Justice; Non-Maleficence)
Is the AI adequately explainable for the task?
(relevant to Explicability)
Should AI be used in a given task context?
(reflection on all ethical principles)
If AI is used, what human control remains necessary?
(reflection on all ethical principles, particularly Autonomy)
Organizational determination: aligns to the organization's overall values-setting for ethical AI use
Meaningful human control
Human in-the-loop
Human on-the-loop
Human out-of-the-loop - Autonomous AI
Utilization
Interpretation bias (i.e., algorithm aversion, automation bias)

Fig. 1 Ethical applications of artificial intelligence to HRM: a decision-making framework

Ethical applications of artificial intelligence in HRM: a decision-making framework

To identify salient factors influencing ethical AI use in HRM requires specifying the nature of different HRM tasks, the ethical risks associated with applying AI to such tasks, and then the extent to which humans should retain control over particular tasks. To do this, the decision-making framework constitutes multiple components (see Fig. 1). First, I outline the importance of organizational value-setting for AI use (component 1) as an overarching basis to situate the framework within its specific contexts-of-use, centered on organizational leadership for the governance of AI. I then draw on task-technology fit literature to develop the notion of ethical task-technology fit (component 2) and outline both technology and task characteristics (or 'sensitivities') that I argue will shape ethical AI deployment, with a focus on indicators of potential ethical risks, and generate key questions to assess ethical task-technology fit. I then connect the responses to requirements for ongoing human control in AI systems (component 3).

Framework component 1: organizational governance and value-setting for AI use

Technologies are implemented into often complex social systems that are shaped by the vision and values of organizational leaders (Hazy, 2006). Good practice AI implementation should begin with an organizational body, such as a senior leadership group, setting, monitoring, and adapting parameters for AI use (Andrews, 2019) and oversighting the collection and management of data (Guszcza et al., 2020). For example, Microsoft's leadership uses six principles (broadly mapping to the AI4People framework) to guide their use of AI: fairness (aligned to the *justice* norm); reliability and safety (aligned to the *beneficence* norm); privacy and security (aligned to the *non-maleficence* norm); inclusiveness (aligned to the *justice* norm); and transparency and accountability (aligned to the *explicability* norm) (Bankins & Formosa, 2021). Microsoft's Office of Responsible AI and Aether Committee then set and monitor compliance with these principles (Microsoft, n.d.). Such value-setting of a leadership body embeds an important tenet of ethical AI literature, that across the lifecycle of AI use multiple stakeholders should be involved in its management to support meeting the *beneficence*, *non-maleficence*, and *autonomy* principles.

The remainder of the framework operates on the basis that this value-setting creates an overarching context-specific frame for determinations of AI use in a given organization. I now turn to outlining specific technology and task characteristics related to the use of AI for HRM, that will inform assessments of ethical task-technology fit in this context and guide ethical AI deployment.

Framework component 2: assessing ethical task-technology fit

Task-technology fit refers to "the degree to which a technology assists an individual in performing his or her portfolio of tasks" (Goodhue & Thompson, 1995, p. 215). It is argued that assessments of fit will reflect technology characteristics, task characteristics, and factors influencing individual technology usage (termed 'utilization') and it has been shown that a better fit between task and technology enhances a worker's performance (Goodhue & Thompson, 1995; Spies et al., 2020). However, I argue that such assessments have implications not only for work performance but, when adapted toward an ethical lens, can also help identify the ways technology use may lead to unethical outcomes.

How and for what purpose AI is implemented partly determines whether benefits or harms are generated from its use. For example, an algorithm autonomously tasked with determining welfare payments, without meaningful human oversight, and ultimately making inaccurate calculations is a deployment context that can generate harms (Braithwaite, 2020). Or AI tasked with assessing employee performance to input into, and potentially communicate, termination decisions raises questions regarding the transparency of data collection and the appropriateness of deploying the technology for such purposes (Obedkov, 2021). These, and other, examples show that technologies being used in ways that may exceed their capabilities, or being used in ways that inappropriately marginalize humans from the work process, can lead to harms.

Given this, the second plank of this framework is assessing what I term *ethical task-technology fit*, or the extent to which the use of AI in a given task context supports meeting the five ethical principles. To formulate this fit assessment, I adapt and build on the work of Sturm and Peters (2020), who contextualize task-technology fit to the specifics of AI. In line with the wider task-technology fit literature, the *technology characteristics* I identify are what I term 'data and AI sensitivities'. These focus on indicators associated with the scope for various biases, scope for changed context of deployment and data 'scope creep', the personal nature of the data used, and levels of AI explainability. The *task characteristics* I identify are what I term 'task sensitivities', focused on indicators associated with the scope of impact of a task, task complexity, and requirements for a 'human touch'. Assessments of these sensitivities will then provide the context for ethical task-technology fit calculations that I suggest will be guided by three key questions regarding AI reliability and appropriateness, fairness, and explainability (Sturm & Peters, 2020). As the focus of this framework is on identifying a range of technology and task sensitivities, or those aspects of AI and its deployment that have the potential to breach ethical principles, the ethical task-technology fit assessment is focused on surfacing potential ethical risks of AI use in HRM. I now outline these technology and task characteristics (sets of sensitivities).

Technology characteristics: data and AI sensitivities

Advancements in machine learning are driven, in part, by access to large datasets and advanced neural network modelling (Walsh et al., 2019). Ethically, this places scrutiny upon the quality and nature of the data used to generate AI outputs and how understandable (or explainable) those outputs are to humans. In a HRM context, I conceptualize these technology characteristics of AI and the data driving it as *data and AI sensitivities* along four indicators: the scope for data and computational bias (high/low); the scope for inappropriate deployment and data 'scope creep' (high/low); the personal nature of the data collected (high/low); and the explainability of AI output (high [greater explainability]/low [lower explainability]).

First, *scope for data and computational biases*[1] reflects a well-documented source of potential harm generated through machine learning methods: the potential for biased outcomes. In their taxonomy of algorithmic bias, Danks and London (2017) show that opportunities for such bias can stem from multiple sources, ranging from technical sources (to do with the design and computations of the technology) to social sources (to do with how the technology is used). Here, I focus on two sources of technical bias.

Training data bias occurs when the input data from which the AI learns is biased in some way and leads to model outputs that deviate from the population and/or leads to morally unjustifiable models (Danks & London, 2017). For example,

[1] There are competing arguments regarding whether, compared to human decision making, AI affords more or less biased decision outcomes (see Parker & Grote, 2020). In this paper I discuss both AI- and human-based biases and generally take the position that leveraging the strengths of technology- and human-based cognition can help mitigate the limitations of each, and that such human-AI collaboration will likely feature significantly in future workplaces (Daugherty & Wilson, 2018; Jarrahi, 2018).

training data may include: historical bias (replicating historical inequalities between groups); representation bias (underrepresenting certain groups); measurement bias (poor proxies represent variables of interest and/or data are measured poorly across groups); and aggregation bias (a single model does not adequately represent different groups) (Robert et al., 2020; Suresh & Guttag, 2020). Use of biased datasets ('bias-in') has clear ethical implications as it can result in, for example, outcomes that perpetuate past patterns of bias ('bias-out').

Issues of training data bias are particularly relevant for HRM. Historically, a range of minority groups have experienced systematic exclusion from full and equal participation in employment through explicit legislative or regulatory barriers and/or implicit attitudes and norms (Braddock & McPartland, 1987). While many of these explicit, and some of the implicit, barriers have been minimized or removed, many do remain along with their effects. Broadly, these include ongoing biases toward, and/or under representation of, employees from diverse racial and ethnic (i.e., non-caucasian) backgrounds, women, differently abled individuals, and employees from the LGBTQ+ community (Liebkind et al., 2016; Pizer et al., 2012). There are already examples of recruitment algorithms (i.e., that screen submitted applications) entrenching past biased recruitment decisions by unfairly excluding some groups for selection while privileging others (Dastin, 2018). "Sourcing algorithms" (i.e., that promote job ads) can also fail to show women technical or engineering jobs as potential applicants because the algorithm is trained on data showing mostly men holding such roles (Bogen, 2019).

Algorithmic focus bias occurs through "differential usage of information in the … training data" (Danks & London, 2017, p. 3). This occurs where information is used to train an AI that ought not be used, even when it is accessible, whether for moral, legal, or other reasons (Danks & London, 2017). Such bias may not even be evident unless data are scrutinized by decision makers. For example, Xerox collected data capturing job applicants' commuting times, which then predicted that employees with quicker commutes were more likely to be retained (Weber & Dwoskin, 2014). However, Xerox managers determined that using such data could unfairly disadvantage (i.e., exclude) job applicants from minority group neighborhoods (which may be further from the place of employment). This type of bias highlights that "there's some knowledge that you gain that you should stay away from when making a hiring decision" (Weber & Dwoskin, 2014). Overall, the scope for these types of bias in certain HRM functions, such as recruitment, increases opportunities for unethical outcomes and particularly breaches the *beneficence, non-maleficence,* and *justice* principles.

Second, the *context of deployment and opportunity for data* '*scope creep*' reflect the potential for social (or user-driven) sources of bias in AI use. *Transfer context bias* refers to an AI being designed for use in one context but then being deployed in a different one (Danks & London, 2017). This potentially voids guarantees of AI accuracy and may lead to misleading or harmful outcomes (Selbst et al., 2019). For example, there exists the potential for roughly similar technical solutions to be deployed across different settings, such as applying an algorithm designed to generate automated risk assessments in a judicial context toward generating automated hiring outcomes (Selbst et al., 2019). Relatedly, how technology is used in organizations can change over time. As the Xerox example demonstrates, data collected for one seemingly harmless purpose (i.e., understanding job applicant commute times) can be re-purposed to search for other relationships between variables that may generate discriminatory outcomes (i.e., linking commute times to retention and potentially disadvantaging minority group job applicants). Selbst et al., (2019, p. 65) label this a "ripple effect" trap, where a technology deployed into a socially dynamic context, such as an organization with changing leadership groups, can lead to new (potentially unethical) uses and interpretations of that technology (Selbst et al., 2019). This shows that the potential for AI's inappropriate deployment and data 'scope creep' in HRM functions increases opportunities for unethical outcomes by potentially breaching the *non-maleficence* principle through generating harms and negative *justice* implications.

Third, the various tasks associated with managing people often necessitates collecting *personal data*, such as address, age, gender, and medical histories. Advancements in AI are also expanding the ability to collect more types of data, ranging from biometric data (e.g., Solon, 2018), to health and wellbeing data (Ajunwa et al., 2016), to facial and voice data in recruitment submissions (Marr, 2018). For some HRM tasks, capturing personal data may be necessary (even legislated) and not overly invasive, such as collecting bank details for payroll purposes. In other cases, the collection of personal data may be desirable to reach a particular outcome (such as optimizing a wellness initiative), but it is not necessary. In these latter cases, organizations must balance the collection of increasingly personal forms of data (which has employee privacy implications, potentially breaching the *non-maleficence* principle) with the value generated from doing so (potentially supporting the *beneficence* principle). The collection of personal data may also have implications for employees' *autonomy*, through opportunities (or not) to opt in or out of providing their data generally, and to opt in or out of providing their data for the purposes of influencing their behaviors. For example, the collection of biometric data, related to one's health and body, can feed into AIs that 'nudge' employees toward certain behaviors, such as optimizing the movements of warehouse workers (Solon, 2018). Organizations must also consider whether increasing the collection of personal data may facilitate AI data 'scope creep' or even data bias (e.g., are there relevant differences in biometric data for older and younger workers?). Therefore, the ability to collect increasingly personal forms of data to support HRM functions increases opportunities for unethical applications of AI, potentially breaching *non-maleficence* and *autonomy* principles.

Finally, *explainability* reflects a unique feature of machine learning (particularly unsupervised forms of it): that its computations can be opaque or 'blackbox' (Carabantes, 2020). This reflects a trade-off, whereby more complex neural network modelling can generate greater predictive power but human insights into that modelling diminish. This has ethical implications, particularly in sensitive use cases of the technology where understanding and explaining the basis for AI-informed decisions is required for moral and/or legal reasons. This reflects that explainability has a particularly close link to the type of task AI is being used for, and so closely relates to the task sensitivities discussed next, and requirements for explainability will vary between tasks. For example, an AI that provides training and development recommendations may not need to account for how it generated and ranked those opportunities. However, understanding how an AI identifies which employees will have their employment terminated becomes more critical to ensure procedural (demonstration of fair processes) and distributive (demonstration of equity in resource distribution) forms of justice, and to perhaps meet legislative requirements. Therefore, the 'blackbox' nature of many AI systems (low explainability) increases opportunities for unethical outcomes by potentially breaching *explicability* (AI should be intelligible), *justice* (clarity of outcomes to promote fairness), and *non-maleficence* (related to the potential for harm from unintelligible decisions) principles.

Task characteristics: task sensitivities

The ethical deployment of AI to various tasks within HRM functions, and the appropriate extent of human control over the technology, will also be shaped by several task sensitivities (the terms 'task' and 'decision' are used interchangeably) that I conceptualize along three indicators: the scope of the decision impact (wide/narrow); the nature of the task (simple/complex); and whether the task requires a 'human touch', through human–human interaction and/or the need for uniquely human skills (high/low).

Different HRM decisions have different levels of *decision impact*. The impact of some activities will be narrow, such as providing an irrelevant training recommendation, with limited ethical implications. Other decision outcomes may generate wider and more significant impacts, such as failing to offer a job applicant a position based on unfairly discriminatory selection. This compromises the individual's wellbeing through limiting employment opportunities (breaching the *beneficence* and *non-maleficence* principles), but also contributes to systemic societal biases and restricts workforce diversity (breaching the *justice* principle). This shows that where AI is used for tasks where errors may generate significant harms for individuals, organizations, and society more broadly, this heightens the risk of breaching ethical principles.

HRM tasks vary in their *nature and complexity*. Broadly, decisions can be viewed as relatively routine (i.e., are simple, repetitively made, and a clear process exists for making them) or more complex (i.e., multiple considerations are required, they are novel, and ambiguity exists in how to approach them). Evidence suggests that simpler tasks are generally best placed to be automated, such as through AI, as they are "most easily understood, optimized, and codified" (Gibbs, 2017, p. 2). Whereas attributes of more complex tasks can restrict, and sometimes exclude, the use of technology for undertaking them. This is because a machine learning-based AI requires an "algorithmic frame", or a set of outputs (i.e., outcomes to be achieved) and inputs (i.e., data) to construct a model (Selbst et al., 2019, p. 60). Compared to simple tasks, complex tasks are often difficult to frame. In terms of outputs, more complex tasks can be challenging to codify. For example, it can be difficult to define a 'good employee' for the purposes of recruitment (as noted by Tambe et al., 2019). Such an outcome is dependent on many indicators that vary across different roles. Universally defining a 'good employee' may also marginalize employee groups with non-standard work histories or those with knowledge, skills, and abilities that may not 'fit' universal definitions. In terms of inputs, more complex tasks may not be wholly or partly reducible to mathematical terms (Greene, 2019). For example, some indicators of a 'good employee' may not be easily quantifiable, such as the ability to get on well with others or expressing organizational citizenship. Inappropriately applying AI models to tasks that are difficult or inappropriate to frame in codified terms may particularly breach *non-maleficence* (e.g., potential for harm)*, justice* (e.g., potential for unjust outcomes)*,* and *explicability* (e.g., poorly specified models lacking accountability) principles.

Further, rare (or 'exceptions') cases are often not well represented in the datasets required for machine learning. This means the ability of AI to model and then generate predictions for such cases is limited. For example, the unique and nuanced nature of offering career planning advice to an employee with a complex work history may be poorly suited to the use of AI solely, as the outcome will likely depend on assessments of the individual case rather than attempting to generalize from other cases that may poorly represent it (Binns, 2020), with the latter approach potentially breaching the *justice* principle. Where such cases exist, it may be preferable and appropriate to have human–human interaction (a task sensitivity indicator discussed next) to capture nuanced information from the employee. This is what Crawford (2013) calls utilizing 'small data', or data collected with small groups or individuals, which may better suit the task's purposes overall but could also complement the use of 'big data'. Overall, more complex tasks are generally not routinely done, may involve ambiguity, and may involve dealing with unique employee circumstances that require case-by-case handling to avoid breaching the ethical principles noted and support the *beneficence* principle.

HRM tasks can involve handling sensitive issues *that require a 'human touch'*, such as those related to performance, work-life balance, and health and wellbeing, among others. Evidence suggests that such tasks continue to require the use of at least some uniquely human judgment and social skills that remain beyond AI's capabilities (Colson, 2019; Gibbs, 2017). I suggest that the need for a 'human touch' in such tasks can take two main forms. First, certain tasks can demand human–human interaction in their execution. Studies show that in circumstances of fully automated decision making (including for HRM decisions) the denial of human interaction can degrade individuals' justice perceptions, or the perceived fairness of a decision, generating feelings of impersonal treatment, devaluation, and dehumanization (such as "being reduced to a percentage") (Binns et al., 2018, p. 1; Lee et al., 2019, p. 16; Lee, 2018). Ensuring human–human interaction in such tasks will help support employee wellbeing (*beneficence* principle) and feelings of being treated with value and dignity (*justice* principle). Second, a 'human touch' may manifest through requiring uniquely human skills, such as context-specific judgment, to execute tasks and particularly complex ones. As noted earlier, this may occur when the individual merits of a case must be assessed and human value judgments must be made that can't easily be specified through AI modelling (see Binns, 2020; Colson, 2019). Empirical evidence supports the idea that individuals view some tasks, particularly those related to HRM, as requiring uniquely human skills. For example, where algorithms allocate work, workers can feel that "human abilities, emotion, and motivation" are not accounted for (Lee et al., 2015, p. 5). Algorithms can also be viewed as "incapable" of undertaking job candidate selection or performance evaluation because "they lack human intuition, only measure quantifiable metrics, and cannot evaluate social interaction or handle exceptions" (Lee, 2018, p. 12). Such findings reflect concerns that an overreliance on AI gives primacy to that which is quantifiable and may lead to the diminishment of that which is not (Selbst et al, 2019). Such concerns link to breaches of *beneficence*, *justice*, *explicability*, and *autonomy* principles, the latter two particularly through inappropriate delegation of tasks between humans and machines.

Assessing ethical task-technology fit: key questions

Examining the various technology (data and AI) and task sensitivities now generates key questions regarding the ethical fit of AI to given tasks. As identified earlier, I take ethical task-technology fit to be an assessment of whether the use of AI for a given task supports meeting the five ethical principles. I adapt Sturm and Peters' (2020) indicators of task-AI fit to formulate these questions and apply them directly toward addressing the ethical implications of AI deployment, to which I argue they are also relevant. These questions are not intended to be mutually exclusive, and indeed overlap, nor exhaustive, but collectively they generate an overarching assessment of whether the use of AI in a given task context carries ethical risks.

The first question refers to *whether the use of AI is reliable and appropriate to the task*. This will be informed by data and AI sensitivities, such as the scope for bias and potentially poor levels of explainability (which may both reduce reliability) and the need for personal data collection (which may require the collection of data deemed inappropriate/too invasive), and task sensitivities, such as a high decision impact, task complexity, and the need for human interaction (which may all make AI use inappropriate for the task). The second question refers to issues of fairness, specifically *whether the use of AI facilitates fairness in task completion*, with a more explicit focus than the first question on opportunities for harms or injustices to occur. For example, various technology and task sensitivities could breach the *justice* principle, such as the scope for bias, context of deployment and data 'scope creep' concerns, and a lack of explainability and human interaction where the task is deemed to require them. The third question focuses on *whether the AI's outputs are adequately explainable for the task*. While explainability is also reflected in the prior questions, it stands alone here to align with Floridi et al.'s (2018) positioning of *explicability* (of which explainability is an important component) as a principle underpinning all others. Here, task sensitivities such as decision impact and complexity will particularly help determine the levels of explainability required for appropriate task completion. Overall, the 'higher' the data and AI sensitivities (which incorporates lower explainability) and the higher the task sensitivities (which incorporates wider impact and higher complexity), the less likely it is that AI's deployment will be reliable/appropriate, fair, and explainable, which reduces ethical task-technology fit. Conversely, the 'lower' each set of sensitivities are (which incorporates higher explainability, narrower decision impact, and lower task complexity), the more likely it is that AI's deployment will be reliable/appropriate, fair, and explainable, which generates better ethical task-technology fit.

Following assessments of ethical task-technology fit, organizations must then determine whether AI will be utilized for particular tasks and, if so, in what ways. This determination will also be informed by the organizational value-setting discussed earlier, as it provides an overall frame guiding such decisions. For example, this value-setting can help determine when AI will and will not be used, in recognition that "the best solution to a problem may not involve technology" (Selbst et al., 2019, p. 63). For example, an available technology may not allow for any human understanding of its outputs (related to the third question above), and so an organization may choose to never implement such a technology given their task context (while also meeting the *explicability* principle).

If ethical task-technology fit exists, a key question then becomes: in what way and to what extent will humans retain control of an AI system? AIs, like many technologies, are imperfect and for all their benefits and computational power can still exhibit brittleness and generate errors (Lohn, 2020). In some instances their autonomous deployment will continue to support meeting the ethical principles. But in other instances, even where ethical task-technology fit exists, some degree of human control will still be required to further enhance that ethical fit and best leverage the capabilities of both humans and AI. This means that ethical AI deployment extends beyond only assessing technology and task sensitivities to further include appropriate and ongoing human control and oversight (McCoy et al., 2019). It is to these questions I now turn.

Framework component 3: what role for humans? Assessing needs for human control

Determining optimal balances of machine use and human control to support ethical AI deployment is a complex issue. Determining the specifics of such control will often involve identifying the tasks of a work process and assessing the nature of human control needed within and across those tasks, particularly where they are inter-related (see Heikoop et al., 2019 for an autonomous vehicle example). Therefore, my aim here is not to provide a prescriptive account of the level of human control required in every possible circumstance, but to instead first outline a taxonomy of levels of human control and then suggest how these link to technology (data and AI) and task sensitivities and assessments of ethical task-technology fit.

McCoy et al. (2019) suggest that some degree of ongoing human involvement alongside machines is required for performance- and responsibility-related reasons. In terms of performance, humans still retain a "cognitive comparative advantage" (Langlois, 2003, p. 167) over machines in a range of areas, such as dealing with "novel or atypical inputs", adding heterogeneity to automated systems, and morally contextualizing decisions aligned to human ethical judgments and ethical considerations that remain challenging to algorithmically program (McCoy et al., 2019, pp. 4–5). In terms of responsibility, ethical and legal lines of accountability can blur when autonomous AI is used. This means operationalizing the balance of human–machine control links to all ethical principles, but particularly to how the *autonomy* (the delegation of decision making to humans versus machines) and *explicability* (the accountability for and explanation of AI actions) principles are enacted.

Literature on human control in automated systems generally presents a taxonomy of control that varies from extensive and demonstrable human control and accountability across an entire work process, to full machine autonomy (perhaps with some limited human oversight). This taxonomy often extends from meaningful human control (the highest level of human control) to humans then being in-, on-, or out-of-the-loop (ending at the weakest level of human control). While definitions of meaningful human control (MHC) are difficult to pin down, broadly "it connotes that in order for humans to be capable of controlling—and ultimately responsible for—the effects of automated systems, they must be involved in a non-superficial or non-perfunctory way" (McCoy et al., 2019, p. 2). Santoni de Sio and van den Hoven (2018) suggest this involves a 'tracking' condition (the system should respond to human moral reasoning and relevant aspects of its environment) and a 'tracing' condition (that one or more human agents can meaningfully understand how the system operates and be fully accountable for it). Requirements for MHC are most prevalent in high-risk contexts, such as autonomous weapons use, but it remains applicable in other sensitive contexts-of-use such as the management of people.

Human in-the-loop (HITL) control means that an AI can, for example, undertake analysis and reach a decision but it cannot autonomously take action, it can only execute following human approval (Walsh et al., 2019). Here, the human can also help identify and correct machine "misbehavior" and be accountable for it (Rahwan, 2018, p. 7). When a human is on-the-loop (HOTL) it means that an AI can, for example, undertake analysis, reach a decision, and then also execute that decision without human approval. However, the execution of that decision is oversighted by a human (i.e., the human knows the AI's actions) and the human can override those actions (i.e., stop or change them) as appropriate. Full automation occurs when a human is out-of-the-loop (HOOTL), where an AI executes actions with no human input or interaction. These levels of human control can then be linked to different technology and task sensitivities.

Generally, the decision-making framework suggests that the higher the technology and task sensitivities overall, which reduces ethical task-technology fit, the greater the imperative to have higher levels of human control exerted upon the system. This reflects an 'additive effect', whereby when there is more than one indicator of 'high' task sensitivity (i.e., a wide decision impact and a highly complex task) and 'high' data and AI sensitivity (i.e., scope for bias and poor explainability), then the case becomes stronger for more meaningful forms of human control to be retained (i.e., MHC or human in-the-loop control). However, when the sensitivities are 'moderate', a human-on-the-loop (focused on monitoring and oversight) may be more appropriate, and where the sensitivities are not relevant or 'low', then more minimal or no human involvement (i.e., a human out-of-the-loop) is likely reasonable. It may also be that organizational value-setting involves 'weighting' each sensitivity indicator to guide determinations of levels of human control.[2]

There may also be tensions across sensitivities that organizations must resolve through 'on balance' assessments. For example, the use of AI to automate payroll processes constitutes a fairly codifiable task (the nature of the task is simple) and there is generally no imperative to have a human communicating standard payroll outcomes (the need for a 'human touch' is low), which both support affirmative responses to reliability and fairness questions when assessing ethical task-technology fit. But if the fortnightly payment of employees is compromised this could have significant implications for workers (the task impact is wide). In such cases, organizations must assess the risk of AI failure (i.e., reliability) and organizational tolerance of it. If the organization assesses such risks as very low, then a human out-of-the-loop may be appropriate to balance the risk of harm with the other ethical principles that the AI's use may support. However, if the organization assesses the risk of harm as too high, a human on-the-loop to approve system actions may be more appropriate.

Utilization

Discussions of human control should recognize that the behaviors of social actors shape technology use (Selbst et al., 2019). The task-technology fit literature recognizes this through the notion of 'utilization', which forms the final plank of this framework. Utilization captures how the "characteristics of the individual … could affect how easily and well he or she (or they) will utilize the technology" (Goodhue, 1995, p. 216). When humans work alongside AI their behaviors and biases toward it can generate unintended harms (breaching *beneficence* and *non-maleficence* principles), particularly when the AI is not used as designed (Selbst et al., 2019; Suresh & Guttag, 2020). While many individual characteristics are relevant to understanding technology 'utilization', I focus here on two human biases that could particularly undermine the ethical use of AI: algorithm aversion and automation bias (Bahner et al., 2008; Prahl & Van Swol, 2017).

Automation bias exists when individuals rely too heavily on automated systems, failing to sufficiently oversight them or intervene when errors occur (Bahner et al., 2008). This can stem from the belief that data analytics "always reflect objective truth" (Crawford, 2013, p. 1). Algorithm aversion refers to the "irrational discounting of automation advice" (Prahl & Van Swol, 2017, p. 691), whereby individuals rely more heavily on human-generated advice over algorithmically-generated advice, even when the latter is seen to outperform the former (Dietvorst et al., 2015). For this decision-making framework, such biases mean that even when humans are expected to be meaningfully involved alongside AI systems, for example when their unique skills are needed, they may not activate those skills (through automation bias) or they may inappropriately disregard the AI's output (through algorithm aversion).

These outcomes could then undermine meeting various ethical principles. For example, where organizations have assessed task sensitivity and determined the appropriate human–machine balance for task completion, disrupting that through diminished or excessive human involvement changes this balance and risks breaching the *autonomy* principle. Algorithm aversion may generate unintended harms through reintroducing the potential for other human cognitive biases to interfere with a task, while automation bias may result in failure to complement AI output with necessary human judgment (each breaching the *non-maleficence* and *justice* principles). The change in human–machine balance, either toward more use of human skills or more reliance on machine skills, may also blur accountability for outcomes (breaching the *explicability* principle).

To mitigate such issues, organizations must align employees' actual AI use with its intended use. This can occur at the organizational value-setting level to include explicit guidance on how employees interact with the AI and rely on it, which particularly supports meeting the *autonomy* and *explicability* principles. This also raises issues regarding the

[2] It should be noted that the framework itself does not 'quantify' or 'weight' each sensitivity indicator, nor suggest one is more important than another. However, differing organizational value-setting may influence how each indicator is perceived and assessed in different contexts.

training and skill development of workers operating alongside AI, which are canvassed further below.

Applying the framework: ethical AI for HRM in action

The framework is now applied to two hypothetical examples focused on AI use in HRM: (1) in health and wellness programs; and (2) in performance management. Both examples are situated in the context of a professional, knowledge-intensive organization and assume that the organization has determined to implement AI for these HRM functions.

The use of AI is transforming the nature of employee health and wellbeing programs by extending the scope of health data extracted from employees and the range of services provided (Ajunwa et al., 2016). I utilize a relatively basic example here. The sedentary, screen-focused nature of much professional work has documented adverse health effects over time (Dunstan et al., 2012). The hypothetical organization wants to help remedy this issue by replacing stationary desks with sit-stand models and by employing an AI-driven program that prompts employees to move during the day and reduce screen time. To do this, via an 'app' employees can choose to install on their devices, the program collects data on how long employees sit during the day, when they log into and off their work devices, and how often they use programs such as email. The app also collects data on employees' age, gender, weight, ethnicity, and any pre-existing health conditions to tailor its prompts. Then, guided by evidence-based health recommendations, the program prompts employees to stand, move, and focus away from the screen at regular intervals.

In terms of data and AI sensitivities, such an AI is collecting *moderately personal data*. While organizations collect some employee demographic data, this generally does not extend into pre-existing health conditions or daily movement patterns. Organizations also tend to collect data on employees' use of work-related devices (i.e., log on/off times), but perhaps not how long computer programs are open. There would appear to be a *low opportunity for bias*, assuming that the AI accurately personalizes recommendations for different groups where necessary. The *opportunity for transfer context bias and data 'scope creep' is arguably high*, as data related to device and program usage for wellness purposes could foreseeably be transformed for performance monitoring purposes. Levels of *explainability* will depend on the algorithm's design, but I will assume *explainability is moderate* and sufficient to the task (i.e., it meets expectations for user understanding). Overall, this suggests *moderate data and AI sensitivity*. In terms of task sensitivities, the computations for prompting an individual to sit-stand or reduce screen time appear reducible to a codified process, suggesting that the *nature of the task is fairly simple*. Assuming the AI accurately prompts individuals, and because employees may opt out of participating, the *outcome impact is narrow*. As technologies such as wearable fitness monitoring devices are increasingly used by individuals to support their wider health and wellbeing, there is also likely a *low need for a 'human touch'* in the prompting to make minor physical changes throughout the workday. Overall, this suggests *nil-low task sensitivity*.

Taken together, the three ethical task-technology fit questions would be answered in the affirmative (reliable/appropriate, fair, and explainable), where the only key ethical risk is the potential for data 'scope creep'. This suggests that a human out-of-the-loop approach to human control would be sufficient for the day-to-day use of the app, but with organizational oversight and approval (i.e., be in-the-loop for) changes to the use of the collected data beyond the current scope.

Performance management is a common HRM function and generally involves "managing employee efforts based on measured performance outcomes" (den Hartog et al., 2004 p. 557). One broadly drawn task within this function is collecting employee performance data, communicating with employees about their performance, and providing feedback. In terms of data and AI sensitivities, *scope for bias is likely moderate* as there are competing considerations here. Some evidence suggests that metric-based performance management (which AI can facilitate) can help reduce bias driven by human errors (Marr, 2017). However, there remains a risk of transfer context bias, algorithmic focus bias and data 'scope creep', depending on the types of information collated and connected by the algorithm (i.e., is it inappropriately using demographic or other data to find connections to performance that are morally questionable?). The *personal nature of the data is likely moderate*, depending on how visible performance metrics are across the organization. As with the previous example, explainability will depend on the algorithm's design, but compared to the first example demands for *explainability will likely be high* given the task sensitivities (potentially wide decision impact and complexity, discussed next) and the potential for regulatory requirements for employers to make performance management decisions transparent. Overall, this suggests *moderate data and AI sensitivity*. In terms of task sensitivities, the hypothetical organizational context means it is likely that performance measurement is *complex* and includes non-codifiable indicators, generating a potentially *wide impact* on the employee if data is incorrectly collected and analyzed. The often sensitive nature of managing performance means the *need for a 'human touch' is likely also high*. Overall, this suggests *high task sensitivity*.

Taken together, the three ethical task-technology fit questions would likely be answered 'maybe'. There are multiple ethical risks, particularly related to the potential for transfer context algorithmic focus biases (a data and AI sensitivity indicator) and high task sensitivity on each of its indicators. This suggests that moving closer to more meaningful human control would be required, particularly for oversighting the types of data the AI uses (mitigating transfer context bias and data 'scope creep'), adequately understanding the AI's outputs (mitigating explainability risks), appropriately contextualizing those outputs with other employee-specific circumstances (in recognition of task complexity and decision impact), and for conveying decisions to employees (affording meaningful human interaction).

Affordances and limitations of the framework

The framework serves to ground broad, normative ethical principles related to AI design and deployment within a specific context-of-use. By outlining a range of technology and task sensitivities relevant to the use of AI within HRM, it can guide organizational decision makers toward key questions to assess the ethicality, and particularly the ethical risks, of using AI in this domain given the nature of the technology and the tasks to which it is applied (i.e., ethical task-technology fit). This then supports assessments of optimal human-AI mixes.

However, the framework does not account for every contextual aspect relevant to understanding how a technology will impact the complex socio-technical system into which it is deployed. A range of intra- and extra-organizational factors will also play a role, alongside many and varied human behaviors that shape use of or resistance to a technology. Further, as the framework is largely assessing risks to ethical AI deployment, it doesn't fully account for the efficiency and other benefits AI may bring. These should also be detailed within a given context and integrated with the framework presented here. Similarly, ethical assessments must also balance the trade-offs in meeting different ethical principles (Davis, 1995). For example, collecting very personal data may improve selection prediction accuracy, but this trades-off applicant privacy. Such issues must also be resolved in the framework's application, and the organization's value-setting could guide such trade-offs. Finally, the discussion of human control remained, by design, broad. There are many potential combinations of human and machine input into a given task and more detailed specifications of this will need to be grounded in the context-of-use (e.g., see Heikoop et al., 2019; Ficuciello et al., 2019).

Implications for research and practice

For scholarship, the framework offers a range of future research directions. A key area is operationalizing and examining the enactment of the various components of the framework, for example by empirically examining the decision-making and value-setting processes of organizational actors in determining AI implementation for HRM. Further, it will be important to identify how (and whether) the proposed sensitivities occur across various HRM tasks and how organizational actors balance assessments of them, how they potentially weight their importance, and whether other sensitivities exist. To nuance component three of the framework (regarding human control), it will be important to examine how each level of control is enacted in practice in this context and how issues of human control relate to the task and technology sensitivities proposed.

The increasing use of AI also has implications for the types of work and skills that HRM practitioners will undertake and require into the future. While such work has begun (see Scholz, 2019), the increasing interaction and collaboration between humans and machines has implications both for work design and employees' skill sets, which requires empirical investigation. While this framework, through indicators such as task complexity and the need for a 'human touch', suggests the types of work that should be retained by humans, the in- and on-the-loop distinctions also foreground skills in understanding the AIs being used, the data powering them, and knowing how to best interpret their outputs (Daugherty & Wilson, 2018; Fleming, 2020).

Practically, the framework offers practitioners pathways for better understanding the nature of the AIs they are implementing and the decision points relevant for doing so in ways that support ethical implementation. The framework extends beyond broad prescriptions that AI should be implemented 'for good', to instead guide practitioners toward answering questions such as 'what does utilizing AI for good in my organization look like and what should I consider?' through assessments of ethical task-technology fit. While developed for a HRM context, the framework also has broader applicability to other types of decision and task settings.

Conclusion

This paper developed a decision-making framework to support the ethical use of AI for HRM. This contributes to research at the intersection of HRM and technology by extending it further toward examining the ethical implications of AI use in this context and how ethical implementation can be supported through considered combinations of human and machine involvement in tasks. The framework also contributes to the ethical AI literature by providing concrete applications of its broad principles in a defined context-of-use. Overall, there remains work to be done to better understand where and when AI is best deployed in a HRM context to ensure it supports positive and functional workplaces.

Funding The author has no relevant financial or non-financial interests to disclose.

Declarations

Conflict of interest The author has no conflicts of interest to declare that are relevant to the content of this article.

References

Aizenberg, E., & van den Hoven, J. (2020). Designing for human rights in AI. *Big Data & Society*. https://doi.org/10.1177/2053951720949566

Ajunwa, I., Crawford, K., & Ford, J. S. (2016). Health and big data: An ethical framework for health information collection by corporate wellness programs. *The Journal of Law, Medicine & Ethics, 44*(3), 474–480.

Albert, E. T. (2019). AI in talent acquisition: A review of AI-applications used in recruitment and selection. *Strategic HR Review, 18*(5), 215–221.

Andrews, L. (2019). Public administration, public leadership and the construction of public value in the age of the algorithm and 'big data.' *Public Administration, 97*, 296–310.

Bahner, J. E., Hüper, A.-D., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies, 66*(9), 688–699.

Bankins, S., & Formosa, P. (2021). Ethical AI at work: The social contract for artificial intelligence and its implications for the workplace psychological contract. In M. Coetzee & A. Deas (Eds.), *Redefining the psychological contract in the digital era: Issues for research and practice* (p. 55). Springer.

Binns, R. (2020). Human judgment in algorithmic loops: Individual justice and automated decision-making. *Regulation & Governance*. https://doi.org/10.1111/rego.12358

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). It's reducing a human being to a percentage": Perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI conference on human factors in computing systems—CHI '18* (pp. 1–14).

Boden, M. A. (2016). *AI: Its nature and future*. Oxford University Press.

Bogen, M. (2019, May 6). All the ways hiring algorithms can introduce bias. *Harvard Business Review*. https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias

boyd, d, & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662–679.

Braddock, J. H., & McPartland, J. M. (1987). How minorities continue to be excluded from equal employment opportunities: Research on labor market and institutional barriers. *Journal of Social Issues, 43*, 5–40.

Braithwaite, V. (2020). Beyond the bubble that is Robodebt: How governments that lose integrity threaten democracy. *Australian Journal of Social Issues, 55*(3), 242–259.

Carabantes, M. (2020). Black-box artificial intelligence: An epistemological and critical analysis. *AI & Society, 35*, 309–317.

Colson, E. (2019). What AI-driven decision making looks like. *Harvard Business Review*. https://hbr.org/2019/07/what-ai-driven-decision-making-looks-like

Crawford, K. (2013, April 1). The hidden biases of big data. *Harvard Business Review*. http://blogs.hbr.org/2013/04/the-hidden-biases-in-big-data

Danks, D. & London, A. (2017). Algorithmic bias in autonomous systems. In: *Proceedings of the twenty-sixth international joint conference on artificial intelligence* (Marina del Rey, CA: IJCAI, 2017) (pp. 4691–4697). https://doi.org/10.24963/ijcai.2017/654.

Dastin, J. (2018, October 11). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Daugherty, P. R., & Wilson, H. J. (2018). *Human+Machine: Reimagining work in the age of AI*. Harvard Business Press.

Davis, R. B. (1995). The principlism debate: A critical overview. *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine, 20*(1), 85–105.

den Hartog, D. N., Boselie, P., & Paauwe, J. (2004). Performance management: A model and research agenda. *Applied Psychology: An International Review, 53*(4), 556–569.

Dietvorst, B., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Schieber, S., Waldo, J., Weinberger, D., & Wood, A. (2017). Accountability of AI under the law: The role of explanation. arXiv:1711.01134.

Dunstan, D. W., Howard, B., Healy, G. N., & Owen, N. (2012). Too much sitting—A health hazard. *Diabetes Research and Clinical Practice, 97*, 368–376.

Ficuciello, F., Tamburrini, G., Arezzo, A., Villani, L., & Siciliano, B. (2019). Autonomy in surgical robots and its meaningful human control. *Paladyn, Journal of Behavioral Robotics, 10*(1), 30–43.

Fleming, M. (2020, March 24). AI is changing work—And leaders need to adapt. *Harvard Business Review*. https://hbr.org/2020/03/ai-is-changing-work-and-leaders-need-to-adapt

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.8cd550d1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society. *Minds and Machines, 28*(4), 689–707.

Formosa, P., & Ryan, M. 2021. Making moral machines: Why we need artificial moral agents. *AI & Society*, *36*(3), 1–13.

Gibbs, M. (2017). How is new technology changing job design? *IZA World of Labor*. https://doi.org/10.15185/izawol.344

Goodhue, D. L., & Thompson, R. L. (1995). Task-technology fit and individual performance. *MIS Quarterly, 19*(2), 213–236.

Greene, T. (2019, February 7). Why the criminal justice system should abandon algorithms. *The Next Web*. https://thenextweb.com/artificial-intelligence/2019/02/07/why-the-criminal-justice-system-should-abandon-algorithms/

Guenole, N. & Feinzig, S. (2018a). The business case for AI in HR: With insights and tips on getting started. *IBM Smarter Workforce Institute*. https://public.dhe.ibm.com/common/ssi/ecm/81/en/81019981usen/81019981-usen-00_81019981USEN.pdf

Guenole, N., & Feinzig, S. (2018b). Competencies in the AI era. *IBM Smarter Workforce Institute*. https://www.ibm.com/downloads/cas/ONNXK64Y

Guszcza, J., Lee, M.A., Ammanath, B., & Kuder, D. (2020). Human values in the loop: Design principles for ethical AI. *Deloitte Insights*. https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/design-principles-ethical-artificial-intelligence.html

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines, 30*, 99–120.

Hazy, J. K. (2006). Measuring leadership effectiveness in complex socio-technical systems. *Emergence: Complexity and Organization, 8*(3), 58–77.

Heikoop, D., Hagenzieker, M., Mecacci, G., Calvert, S., Santoni De Sio, F., & van Arem, B. (2019). Human behaviour with automated driving systems: A quantitative framework for meaningful human control. *Theoretical Issues in Ergonomics Science, 20*(6), 711–730.

Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems—CHI '19* (pp. 1–16). https://doi.org/10.1145/3290605.3300830

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons, 61*(4), 577–586.

Jia, Q., Guo, Y., Li, R., Li, Y., & Chen, Y. (2018). A conceptual artificial intelligence application framework in human resource management. *ICEB 2018 Proceedings (91)*.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399.

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. *Academy of Management Annals, 14*(1), 366–410.

Langlois, R. N. (2003). Cognitive comparative advantage and the organization of work. *Journal of Economic Psychology, 24*(2), 167–187.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society, 5*(1), 205395171875668.

Lee, M. K., Jain, A., Cha, H. J., Ojha, S., & Kusbit, D. (2019). Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. In *Proceedings of the ACM on human-computer interaction*, *3*(CSCW) (pp. 1–26).

Lee, M. K., Kusbit, D., Metsky, E., & Dabbish, L. (2015). Working with machines: The impact of algorithmic, data-driven management on human workers. In *Proceedings of the 33rd annual ACM SIGCHI conference*, Seoul, South Korea (pp. 1603–1612). ACM Press.

Liebkind, K., Larjab, L., & Brylka, A. (2016). Ethnic and gender discrimination in recruitment: Experimental evidence from Finland. *Journal of Social and Political Psychology, 4*(1), 403–426.

Lohn, A. (2020). Estimating the brittleness of AI: Safety integrity levels and the need for testing out-of-distribution performance. arXiv:2009.00802

Marr, B. (2018). The amazing ways Unilever uses artificial intelligence to recruit and train thousands of employees. *Forbes*. https://www.forbes.com/sites/bernardmarr/2018/12/14/the-amazing-ways-how-unilever-uses-artificial-intelligence-to-recruit-train-thousands-of-employees/?sh=286750f56274

Marr, B. (2017). The future of performance management: How AI and big data combat workplace bias. *Forbes*. https://www.forbes.com/sites/bernardmarr/2017/01/17/the-future-of-performance-management-how-ai-and-big-data-combat-workplace-bias/?sh=3fbb173e4a0d

McCoy, L., Burkell, J., Card, D., Davis, B., Gichoya, J., LePage, S., & Madras, D. (2019). *On meaningful human control in high-stakes machine-human partnerships*. UCLA School of Law, Science, and Evidence (PULSE), California Digital Library: University of California.

Microsoft. (n.d.). *Responsible AI*. https://www.microsoft.com/en-us/ai/responsible-ai

Obedkov, E. (2021). Xsolla fires 150 employees using big data and AI analysis, CEO's letter causes controversy. *Game World Observer*. https://gameworldobserver.com/2021/08/04/xsolla-fires-150-employees-using-big-data-and-ai-analysis-ceos-letter-causes-controversy

OECD. (2019). *Artificial intelligence in society*. OECD Publishing. https://doi.org/10.1787/eedfee77-en.

Parker, S. K. & Grote, G. (2020). Automation, algorithms, and beyond: Why work design matters more than ever in a digital world. *Applied Psychology: An International Review*. https://doi.org/10.1111/apps.12241

Pizer, J. C., Sears, B., Mallory, C., & Hunter, N. D. (2012). Evidence of persistent and pervasive workplace discrimination against LGBT people: The need for federal legislation prohibiting discrimination and providing for equal employment benefits. *Loyola of Los Angeles Law Review, 45*, 715–780.

Prahl, A., & Van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting, 36*, 691–702.

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology, 20*(1), 5–14.

Robert, L. P., Pierce, C., Marquis, L., Kim, S., & Alahmad, R. (2020). Designing fair AI for managing employees in organizations: A review, critique, and design agenda. *Human-Computer Interaction*. https://doi.org/10.1080/07370024.2020.1735391

Santoni de Sio, F., & van den Hoven, J. (2019). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*. https://doi.org/10.3389/frobt.2018.00015

Scholz, T. M. (2019). Big data and human resource management. In J. S. Pederson & A. Wilkinson (Eds.), *Big data: Promise, application and pitfalls* (pp. 69–89). Edward Elgar.

Selbst, A., boyd, d., Friedler, S., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In: *ACM conference on fairness, accountability, and transparency* (FAT* 2018).

Solon, O. (2018, February 1). Amazon patents wristband that tracks warehouse workers' movements. *The Guardian*. https://www.theguardian.com/technology/2018/jan/31/amazon-warehouse-wristband-tracking

Spies, R., Grobbelaar, S., & Botha, A. (2020). A scoping review of the application of the task-technology fit theory. In: *Lecture notes in computer science* (Vol. 12066, pp. 397–408). Springer.

Stahl, B. C., & Coeckelbergh, M. (2016). Ethics of healthcare robotics: Towards responsible research and innovation. *Robotics and Autonomous Systems, 86*, 152–161.

Starner, T. (2019, December, 2019). AI can deliver recruiting rewards, but at what legal risk? *Human Resource Executive*. https://hrexecutive.com/ai-can-deliver-recruiting-rewards-but-at-what-legal-risk/

Strohmeier, S., & Piazza, F. (2015). Artificial intelligence techniques in human resource management—A conceptual exploration. In C. Kahraman & S. Çevik Onar (Eds.), *Intelligent techniques in engineering management* (Vol. 87, pp. 149–172). Springer.

Sturm, T. & Peters, F. (2020). The impact of artificial intelligence on individual performance: Exploring the fit between task, data, and technology. *ECIS 2020 Research Papers*. https://aisel.aisnet.org/ecis2020_rp/200

Suresh, H., & Guttag, J. V. (2020). A framework for understanding unintended consequences of machine learning. arXiv:1901.10002 [Cs, Stat]

Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review, 61*(4), 15–4228.

Walsh, T., Levy, N., Bell, G., Elliott, A., Maclaurin, J., Mareels, I., & Wood, F. (2019). *The Effective and ethical development of Artificial Intelligence* (p. 250). ACOLA. https://acola.org/wp-content/uploads/2019/07/hs4_artificial-intelligence-report.pdf

Weber, L., & Dwoskin, E. (2014, September 29). Are workplace personality tests fair? *The Wall Street Journal*. http://www.wsj.com/articles/are-workplace-personality-tests-fair-1412044257

World Economic Forum. (2018). *The future of jobs report 2018*. Centre for the New Economy & Society. http://www3.weforum.org/docs/WEF_Future_of_Jobs_2018.pdf

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.