THREE ESSAYS ON HRM ALGORITHMS

THREE ESSAYS ON HRM ALGORITHMS: WHERE DO WE GO FROM HERE?

BY

MAGGIE (MINGHUI) CHENG, M.B.A.

A Thesis

Submitted to the Department of Human Resources Management

and the School of Graduate Studies

of McMaster University

in Partial Fulfilment of the Requirements

for the Degree of

Doctor of Philosophy

Doctor of Philosophy (2024) McMaster University

(Human Resources Management) Hamilton, Ontario, Canada

TITLE:	Three Essays on HRM Algorithms: Where Do We Go from Here?
AUTHOR:	Maggie (Minghui) Cheng Master of Business Administration City University of Hong Kong, China Bachelor of Arts in Journalism and International Communication Beijing Foreign Studies University, China
SUPERVISOR:	Dr. Rick D. Hackett
NUMBER OF PAGES:	xv, 148

LAY ABSTRACT

This thesis explores the use of advanced algorithms in Human Resource Management (HRM) and how they affect decision-making in organizations. With the rise of big data and powerful algorithms, companies can analyze various HR practices like hiring, compensation, and employee engagement. However, there are concerns about biases and ethical issues in algorithmic decision-making. This research examines the benefits and challenges of HRM algorithms and suggests ways to ensure fairness and ethical considerations in their design and application. By bridging the gap between theory and practice, this thesis provides insights into the responsible use of algorithms in HRM. The findings of this research can help organizations make better decisions while maintaining fairness and upholding ethical standards in HR practices.

ABSTRACT

The field of Human Resource Management (HRM) has experienced a significant transformation with the emergence of big data and algorithms. Major technology companies have introduced software and platforms for analyzing various HRM practices, such as hiring, compensation, employee engagement, and turnover management, utilizing algorithmic approaches. However, scholarly research has taken a cautious stance, questioning the strategic value and causal inference basis of these tools, while also raising concerns about bias, discrimination, and ethical issues in the applications of algorithms. Despite these concerns, algorithmic management has gained prominence in large organizations, shaping workforce management practices. This thesis aims to address the gap between the rapidly changing market of HRM algorithms and the lack of theoretical understanding.

The thesis begins by conducting a comprehensive review of HRM algorithms in HRM practice and scholarship, clarifying their definition, exploring their unique features, and identifying specific topics and research questions in the field. It aims to bridge the gap between academia and practice to enhance the understanding and utilization of algorithms in HRM. I then explore the legal, causal, and moral issues associated with HR algorithms, comparing fairness criteria and advocating for the use of causal modeling to evaluate algorithmic fairness. The multifaceted nature of fairness is illustrated and practical strategies for enhancing justice perceptions and incorporating fairness into HR algorithms are proposed. Finally, the thesis adopts an artifact-centric approach to examine the ethical implications of HRM algorithms. It explores competing views on moral responsibility, introduces the concept of "ethical affordances," and analyzes the distribution of moral responsibility based on different

types of ethical affordances. The paper provides a framework for analyzing and assigning moral responsibility to stakeholders involved in the design, use, and regulation of HRM algorithms.

Together, these papers contribute to the understanding of algorithms in HRM by addressing the research-practice gap, exploring fairness and accountability issues, and investigating the ethical implications. They offer theoretical insights, practical recommendations, and future research directions for both researchers and practitioners.

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank all those who have supported and contributed to the completion of my PhD thesis. Without their presence in my life, this journey would not have been possible.

First and foremost, I extend my deepest gratitude to my supervisor, Dr. Rick D. Hackett for his inquisitive enthusiasm, guidance, and unwavering support. Your expertise, guidance, and mentorship have been instrumental throughout this entire process. Your dedication to my academic growth, your insightful feedback, and your constant belief in my abilities, particularly during moments when I've doubted myself, have truly shaped the outcome of this thesis. I am grateful for the countless hours you have spent providing guidance and motivating me to achieve my best.

I would also like to thank the members of my thesis committee, Dr. Vishwanath V. Baba and Dr. Aaron Schat. Your knowledge, constructive criticism, and valuable insights have influenced the direction and quality of my research as well as who I have become today. I really enjoyed our intellectual conversations from different perspectives, and appreciate the time and effort you put into reviewing and evaluating my work.

A special acknowledgment goes to Dr. Richard D. Johnson, who served as my external examiner. Your feedback on my dissertation from an Information System perspective was enlightening and greatly appreciated. I'm also grateful to Dr. Milena Head for chairing my defence and providing encouragement throughout the process.

I am indebted to my colleagues and fellow researchers from DeGroote and other institutions, who have provided a supportive and intellectually stimulating environment throughout my PhD journey. I cherish the memories of the fun moments we

shared and am forever nostalgic for those times.

Furthermore, I would like to express my gratitude to McMaster University for providing the resources, facilities, and opportunities that have facilitated my research. The support from the department and university staff has been invaluable, and I am thankful for their assistance throughout my academic journey.

Lastly, I would like to acknowledge my close family for their love, support and understanding, without whom this dissertation may have been completed sooner but it would never have been the same journey. Now it's time for some new adventures.

LIST OF PUBLICATIONS

1. **Cheng, M.**, Hackett, R.D (2021). A Critical Review of Algorithms in HRM: Definition, Theory, and Practice, ***Human Resource Management Review***, 31(1), 100698. https://doi.org/10.1016/j.hrmr.2019.100698

2. **Cheng, M.** (forthcoming). From Counterfactual Fairness to Algorithmic Fairness: Building Principle of Equity in AI Management, ***Academy of Management Proceedings,*** 2024(1), 17746.

3. **Cheng, M.** (2020). Who is Responsible? Transparency, Affordances, and Accountability of Ethical Algorithms, ***Academy of Management Proceedings,*** 2020(1), 18388. 10.5465/AMBPP.2020.18388abstract

4. **Cheng, M.**, (2017). Causal Modeling in HR Analytics: A Practical Guide to Models, Pitfalls, and Suggestions, in Guclu Atinc (Ed.), ***Best Paper Proceedings of the Seventy-seventh Annual Meeting of the Academy of Management.*** https://doi.org/10.5465/AMBPP.2017.187

5. **Cheng, M.**, Hackett, R.D, Li, C (2018). In Search of a Language of Causality in the Age of Big Data for HR Analytics, ***Academy of Management Global Proceedings***, Surrey(2018), 170
https://doi.org/10.5465/amgblproc.surrey.2018.0170.abs

6. Turak, A., **Cheng, M.**, van Egdom, C., and van Gastle, G. (2019, June), Students and Instructors as Partners in Designing Labs: The Value of Conflicting Perspectives. in ***Discussions on University Science Teaching: Proceedings of the Western Conference on Science Education***, 2(1), 8039. https://ojs.lib.uwo.ca/index.php/wcsedust/article/view/8039

Contents

List of Figures

List of Tables

Chapter 1 PREFACE

BACKGROUND AND MOTIVATION

Despite Human Resource Management (HRM) traditionally being viewed as one of the least data-driven business functions (Bersin, 2012; Davenport, 2014; Martin et al., 2014), the availability of big data and algorithms has transformed the HR landscape. Major technological giants like Google, Microsoft, IBM, and LinkedIn have introduced software and platforms for analyzing HRM practices and outcomes related to hiring, compensation, employee engagement, and turnover management (Dignan, 2018; Meister, 2017; Walker, 2012; Walter, 2018). Examples of algorithmic applications include Deloitte's prediction of salespeople's performance based on resume errors (Bersin, 2013b), Xerox's identification of turnover risks based on personality types (Walker, 2012), and IBM's analysis of overtime work to predict employee attrition (Alexander, 2016). However, scholarly research has taken a cautious stance, questioning the strategic value (Angrave et al., 2016) and the causal inference basis (Cheng, 2017) of these tools. Concerns about bias, discrimination, and ethical issues surrounding AI applications in HR contexts (Buolamwini & Gebru, 2018; Peña et al., 2020; Vassilopoulou et al., 2023; Wang & Kosinski, 2018) have also been raised. Nonetheless, algorithmic management is playing an influential role in large organizations (Faraj et al., 2018; Gal et al., 2020; Kellogg et al., 2020), and the market for workforce analytics is projected to exceed $1 billion annually by 2022 (ZionMarketResearch, 2021).

Throughout the course of this thesis, the capabilities of algorithms utilized in HRM have witnessed substantial advancements, largely driven by the incorporation

of artificial intelligence (AI) techniques borrowed from various domains. This progression has enabled HRM practitioners to employ AI algorithms originally designed for managing vehicles to the management of human resources. Over the years, researchers have highlighted a knowledge gap, with practitioners often lacking awareness of impactful HR research findings (Deadrick & Gibson, 2009; Rynes et al., 2002, 2007). However, with the rapid advancement of algorithmic tools in HRM practices, concerns have emerged regarding the potential irrelevance of management researchers in the face of the expanding applications of complex modeling in the workplace (Phan et al., 2017). However, amidst these developments, a fundamental question remains unaddressed: as HRM researchers, who often possess limited technical expertise, what is our responsibility in the context of such algorithmic applications? What unique contributions can we offer? Moreover, considering the overarching challenge of ethics in AI, where should our focus lie moving forward? These critical inquiries underscore the need to critically examine the implications and ethical considerations surrounding the integration of AI algorithms in HRM, and to identify the role of HRM researchers in shaping and navigating this evolving landscape.

In summary, I advocates for the equal importance of the "human" aspect vis-à-vis understanding the algorithmic black box in the field of Human Resource Management (HRM). I emphasizes the need to "put the human back" into HRM while unraveling the intricacies of HRM algorithms. By recognizing and integrating both aspects, I contribute to the advancement of HRM research and algorithmic management practices to encourages interdisciplinary collaboration and highlights the challenges and opportunities in this rapidly evolving field.

THESIS OVERVIEW

The research outlined in this thesis has been published or is in the process of being published.

In light of the reverse research-practice gap, the first paper of this thesis serves as a comprehensive review by comparing the use of algorithms in HRM practice and scholarship. Through a thorough examination of high-quality academic research, trade journals, and popular press articles, I address key aspects of algorithms in HRM. First, I clarify the definition of algorithms in the HRM context and highlight their differences from traditional statistical approaches. I explore the unique features and value of algorithms in HRM decision-making processes. Second, I identify specific topics and issues within HRM that have attracted the interest of those using algorithmic approaches. By analyzing the existing literature, I uncover the areas where algorithms have been extensively applied, such as hiring, compensation, employee engagement, and turnover management. Additionally, I delve into pressing research questions that arise from the current state of the HRM algorithmic literature. Through critical analysis, areas that require further investigation are identified. Lastly, I aim to bridge the divide between scholarship and practice by examining the implications of our findings for both research-oriented and application-oriented databases. This review seeks to facilitate knowledge exchange, foster collaboration, and ensure the practical implementation of algorithmic HRM practices based on evidence-based research. By closing the gap between academia and management practice, I enhance the understanding and utilization of algorithms in HRM, ultimately benefiting organizations and their workforce.

The second paper focuses on the legal, causal, and moral issues associated with the use of algorithms in HR, which refers to software that augments HR-related decisions and automates HRM activities. Building upon the emerging discussion on fairness in algorithmic HR, this study contributes to the literature in several ways. Firstly, it presents a comprehensive list of potential sources of unfairness, with a specific emphasis on legal considerations, which aligns with previous research. Secondly, it critically compares the concept of counterfactual fairness, widely used in computer science, to alternative fairness criteria derived from the management and psychology literatures, such as equity theory and justice theory, highlighting both similarities and differences. Thirdly, it advocates for the use of causal modeling, specifically Structural Causal Modeling (SCM), to evaluate the fairness of HR algorithms. The paper demonstrates how causal modeling can be employed to assess legal compliance aspects related to HR algorithm use. It argues that grounding algorithms in well-established causal models is crucial to fully address the potential for discrimination. Fourthly, the study emphasizes the multifaceted and dynamic nature of fairness, extending beyond legal compliance to encompass moral considerations. It suggests strategies for employers to enhance justice perceptions regarding the use of HR algorithms, recognizing the challenges involved. Finally, the paper proposes a five-step model to promote fairness integration into HR algorithms, providing practical guidance for implementation. Through these contributions, I advance the understanding of algorithmic fairness in HR and provide actionable insights for both researchers and practitioners.

The third (and final) paper adopts an artifact-centric approach to explore the ethical implications of HRM algorithms. It examines the competing views on moral

responsibility in their design and use, one emphasizing accountability on humans and the other placing responsibility on the owners and users of the algorithms. By developing the concept of "ethical affordances" based on affordances theory, the paper investigates how the intentions of designers and perceptions of users shape the ethical outcomes and allocation of responsibility. Through the thought experiment method, a hypothetical case involving an HRM algorithm is analyzed, considering four forms of moral responsibility: culpability, fair communication, public accountability, and active responsibility. The paper demonstrates how different types of ethical affordances embedded in HRM algorithms, such as perceptible and real, perceptible but false, and real but hidden, lead to distinct distributions of moral responsibility among designers, users, and regulators. This ethical affordances framework provides a clear basis for analyzing and assigning moral responsibility to major stakeholders and contributes to the existing literature. The paper concludes by discussing theoretical and practical implications of the framework and identifying areas for future research.

This thesis also includes a research agenda outlining my future work and an afterword addressing other researchers who may be interested in conducting similar research.

Chapter 2 A CRITICAL REVIEW OF ALGORITHMS IN HRM: DEFINITION, THEORY, AND PRACTICE

ABSTRACT

The recent surge of interest concerning data analytics in both business and academia has been accompanied by significant advances in the commercialization of HRM (Human Resource Management)-related algorithmic applications. A review of the literature uncovered 22 high quality academic papers and 122 practitioner-oriented items (e.g., popular press and trade journals). As part of the review, I draw several distinctions between the typical use of HRM algorithms and more traditional statistical applications. I find that while HRM algorithmic applications tend not to be especially theory-driven, the "black box" label often invoked by critics of these efforts is not entirely appropriate. Instead, HRM-related algorithms are best characterized as heuristics. In considering the implications of these findings, I note that there is already evidence of a research-practitioner divide; relative to scholarly efforts, practitioner interest in HRM algorithms has grown exponentially in recent years.

Keywords: algorithm, Big Data, HRM, heuristics, causality

INTRODUCTION

The attributes of volume, velocity, and variety associated with big data (Laney, 2001) cannot contribute to insightful decision-making without developing and applying proper algorithms. As workforce digitization is creating "ever increasing volumes of data" (George et al., 2014), algorithms are crucial to the interpretation of the data in a manner that has the potential to add value. In Human Resource Management (HRM) specifically, the increased datafication of HRM practices is calling attention to the development and application of advanced HRM algorithms.

HR used to be viewed as one of the least data-driven of all the business functions (Bersin, 2012; Davenport, 2014; Martin et al., 2014); however, the availability of big data and associated algorithms has drastically changed the HR landscape. Major technological giants including Google, Microsoft, IBM, and LinkedIn have all launched software or platforms that enable the analysis of HRM practices and outcomes, including those related to hiring, compensation, employee engagement, and the management of turnover (Dignan, 2018; Meister, 2017; Walker, 2012; Walter, 2018). Although HR-related algorithms have been developed and applied to small data sets, examples of big-data algorithmically-driven recommendations receive much of the publicity. Deloitte, for example, found that a lack of grammatical errors on a large dataset of resumes was predictive of the performance of salespeople to a greater degree than were academic grades (Bersin, 2013b). Similarly, Xerox found that personality types predict turnover, such that those identified as creative tended to stay longer than those regarded as being inquisitive (Walker, 2012). Data-driven analyses by IBM revealed that employees who worked overtime without rewards or promotion

were more likely to leave the organization (Alexander, 2016). Ultimately, the market for workforce analytics is projected to exceed 1 billion USD annually by 2022 (ZionMarketResearch, 2021). While the use of algorithms seems to be thriving in HRM practice, the scholarship concerning their use (including related analytical models) reflects a comparatively cautious, conservative outlook (Angrave et al., 2016; Cheng, 2017; Marler & Boudreau, 2017; Rasmussen & Ulrich, 2015). HRM researchers question, for example, the value of analytics-driven software for decision making; Angrave et al. (2016) concluded that there is little evidence to support the strategic value of these tools. Relatedly, Cheng (2017) warned that using analytical models without a strong basis for making causal inference is likely to result in spurious models that add little value to HRM practice. In all, some believe that the zealous embrace of algorithmic models in HRM practice will turn out to be a management fad (Angrave et al., 2016; Rasmussen & Ulrich, 2015).

The major aim of this paper is to provide a review of both high quality academic research as well as items in trade journals and the popular press concerning the use of algorithms in HRM. By doing so, I hope to help bridge the long-standing gap between academics and management practice (Bansal et al., 2012; Bartunek & Rynes, 2010; Ghoshal, 2005; Hambrick, 1994; Mowday, 1997; Pearce, 2004; Rousseau, 2006; Walsh et al., 2007). In the past, researchers have indicated a knowledge gap was caused by practitioners not being aware of impactful HR research findings (Deadrick & Gibson, 2009; Rynes et al., 2002, 2007). However, the recent rapid development of algorithmic tools in HRM practices leads to concerns that management researchers may be at risk of being left irrelevant in fast-growing workplace applications of complex modeling (Phan et al., 2017). As I discuss below, such concern is not unfounded. My analysis

takes this discrepancy from the realm of speculation (Bartunek & Rynes, 2010; Phan et al., 2017; Walsh et al., 2007) to empirical evidence.

As part of this review, I address several research questions. First, I seek to provide clarification concerning the definition of algorithms in the HRM context, including how their application differs from traditionally used statistical approaches. Second, I assess the extent to which there are particular topics or issues within HRM that have attracted the interest of those who use algorithmic approaches. Third, I identify several especially pressing research questions given the state of the HRM algorithmic literature. Finally, in line with the goal of helping to bridge the divide between scholarship and practice, I examine the degree to which the answers to these questions depend on whether I are dealing with a research-oriented or application-oriented database.

HRM ALGORITHM IN CONTEXT

As used in math, computer science, and related fields, an algorithm has a strict definition as an "unambiguous" specification in relation to problem solving (Boolos et al., 2002; Knuth, 1997; Rogers, 1967). Unambiguity typically refers to three criteria of clarity: (1) each step in the algorithm is clearly-identified; (2) the inputs and outputs of the algorithm are well-defined; and (3) the algorithm has a guaranteed end point that produces a correct result (Knuth, 1997; Rogers, 1967). The underlying algorithmic logic between the inputs and the outputs of algorithms can be roughly separated into two categories: deterministic and probabilistic (Cormen et al., 2009, p. 114-116, 123). The most studied type of algorithms in math and computer science assumes a deterministic relationship between the inputs and outputs, which means

that if an input A causes the output B, then A must *always* be followed by B. For instance, in one of the most famous algorithms, *the Traveling Salesman Problem (TSP)*[1], if the location of all cities were known, and the order that the salesman travel through each cities is fixed, the salesman will *always* travel the exact same total distance – not one mile more, not one mile less. The other type of algorithm is used to uncover probabilistic relationships between inputs and outputs, which means that the *occurrence of A increases the probability of B*. Informally, this type of algorithm is more often used when researchers are only exposed to imperfect knowledge of a real-life scenario, such as the relationship between smoking and lung cancer. An algorithm with a probabilistic nature employs a degree of randomness as part of its logic, therefore it does not guarantee correctness.

The computational nature of algorithms in HRM research is no different than those from math and computer science and can be either deterministic or probabilistic. For instance, finding the best solution for workforce scheduling is very similar to TSP – once I know the distances between destinations, the rules of scheduling among employees, and the sequence of services, the sales person will always spend the same time to travel the same distance. These optimization problems based on established deterministic causal relationships are usually at the centre of research in operations management field in business schools. On the other hand, most HRM researchers are interested in problems that are probabilistic in nature, for instance, whether conscientiousness increases the probability of better individual performance and how much effect it would have. The traditional regression models used in HRM research follow

[1]The traveling salesman problem (TSP) was first formulated in 1930 and is one of the most studied problems requiring optimization algorithms. It asks the following question: "Given a list of cities and the distances between each pair of cities, what is the shortest possible route that visits each city and returns to the origin city?" The problem is computationally difficult and has further applications in planning, logistics, or even DNA sequencing.

this probabilistic logic and are also algorithms in a broad sense. Since algorithms with a probabilistic nature do not guarantee correctness, researchers are particularly careful in applying them to make causal claims. Further discussion on this topic will be included later.

Both categories of algorithms have significant practical value for HRM practices, yet I believe that algorithms that are probabilistic in nature would be more meaningful to HRM research due to several considerations. First, these algorithms are closely-connected to the kinds of questions that HRM field tries to answer, such as recruitment, selection, turnover, and performance management. Second, algorithms with non-definitive answers results in a certain level of randomness, which requires judgment from their users. Third, as a result, decisions regarding such algorithms may challenge the existing ethical framework of HRM practices. Therefore, this article is aimed at understanding how HRM research and practice use and interpret algorithms with a probabilistic nature, as well as sparking discussion on the need for guidance for their application.

DATA AND PRELIMINARY ANALYSIS

Collecting Quality Sources for Review

Articles referring to algorithms in both HRM research and practice were identified using "integrative synthesis" review procedures (Rousseau et al., 2008), a systematic methodology to identify a comprehensive pool of literature. I searched two major multidisciplinary publication databases: ProQuest Databases and Web of Knowledge. ProQuest Databases covers multidisciplinary content from 90,000 authoritative

publishers and includes business, social sciences, communications, and engineering. Web of Knowledge, previously known as Web of Science, is a comprehensive research platform encompassing 256 disciplines, including science, social science, the arts, and humanities, reflective of more than 12,500 high impact journals, 170,000 conference proceedings, and 70,000 books. Together, these databases cover both the academic and non-academic domains of interest.

To address concerns from management scholars that evidence-based reviews are susceptible to underplaying the importance of the quality of evidence (Barends et al., 2014; Marler & Boudreau, 2017), I followed Barends et al. (2014)'s recommendation to evaluate each paper for the degree to which, for example, an explicit research question was formulated, followed by data collection and analyses designed to address the issue. I also followed the Marler and Boudreau (2017) approach by differentiating papers that were both peer-reviewed and members of the Journal Quality List (Harzing, 2018) for business or organizational contexts, from those that were not. The JQL is intended to help academics to identify outlets that meet reasonable scholastic standards (Harzing, 2018) and is often used by universities to evaluate publications for the purposes of tenure decisions (Marler & Boudreau, 2017). The JQL covers high-quality journals in multiple business-related disciplines with hundreds of those journals from Organisation behaviour/Studies, HRM, Industrial Relations, Psychology, and General Management. Regarding non-academic sources, I deliberately limited this review to sources that are part of the ProQuest Business Premium Collection such that sources with less credibility (e.g. company websites, individual blogs, etc.) were excluded.

HRM Research vs. Practice: The General Trend of Interest in Algorithms

The protocols of "human resource algorithm," "talent algorithm," and "workforce algorithm" were used in conjunction with the targeted databases. Also, in the search for research content I used the following keyword combinations, each paired with "algorithm": recruitment and selection; personnel training; performance; turnover; job satisfaction; pay and compensation; and salary; yielding 10 protocols containing "algorithm".

Regarding the research-based search, a preliminary pool of 536 articles was found, but 385 of them were ultimately eliminated because they did not involve JQL 2018 list outlets. Of the remaining 151 publications, 24 were excluded because the algorithms involved were either unrelated to HR (e.g. other business disciplines, and various subfields of computer science) or did not directly describe an algorithm for HR purposes (e.g., the study of peoples' perceptions of algorithms). The result was that 127 high-quality publications met the criteria for the review.

Figure 1(a) shows that the number of articles has grown steadily since 1974.

Insert Figure 1 about here.

Turning to the search of practitioner literature, the primary protocols as used in the research search (human resource algorithm, talent algorithm, and workforce algorithm) yielded 728 non-academic entries from ProQuest Business Premium Collection. These included newspapers, magazines, trade journals, as well as wire feeds,

blogs and, websites belonging to institutions. Given the large number of items uncovered by this initial effort, I elected not to use the additional protocols applied in the research-oriented search because I was confident that this initial effort sufficiently reflected the overall trends in practice.

As with the research-oriented findings, I confined the focus concerning practice to material that directly described the application of an algorithm to one or more HR functions. Thus, I first removed 144 articles that mentioned the algorithm-related keywords in general contexts. Second, 241 articles describing the use of algorithms outside of HR (e.g., finance, healthcare, media and fashion) were removed. Third, 35 articles that were general commentaries concerning the use of algorithms to HR, without any reference to specific HR functions, were removed. Lastly, I eliminated 86 overlapping items that reflected coverage of the same issue and/or event by multiple media events (e.g. XYZ company launches/kicks off/creates ABC algorithm). This resulted in 222 articles that described a computer algorithm that was specifically used in relation to one or more HR functions. Note that beyond the practitioner material I cite to illustrate various points in the paper, page constraints prevented us from providing a complete listing of this work here; a complete list of them is available from the author.

Figure 1(b) shows a notable surge in coverage of HR-related algorithms in trade journals and mass media after 2014. Moreover, a comparison of the trendlines shown in Figures 1(a) and 1(b) reveals that for many years research and practice-oriented interest in HR-related algorithms grew together, though in the past 5-10 years the practice content has grown at a much faster rate.

Analysis of the Literature by Subfields

I now examine the findings by HR function. Regarding the research-oriented literature, it is notable that 105 of the 127 scholarly articles (82.7%) focused on solving the challenges of increasing "manpower" efficiency by optimizing the allocation, routing, and scheduling of a workforce. As explained earlier, this research applies deterministic logic that yield exact answers as opposed to the algorithms that are probabilistic in nature and better connected with the HRM field. Examples include rearranging activities of operators working in transportation (airlines, trucks, railways, ships, etc.); or in systems requiring on-demand services (call centres, hospitals, gas stations, etc.). This literature concerning the optimal operation of equipment and the provision of services is thus most representative of the use of algorithms in operations management and industrial engineering contexts (Cardoen et al., 2010; Edwards & Holt, 2009; Ernst et al., 2004; Nof & Grant, 1991). Since, as explained earlier, the emphasis is on probabilistic algorithms, from this point forward, I focus on these applications, as reflected in the 22 remaining high-quality peer-reviewed papers. Also, irrespective of the algorithmic form involved, the interest is in the subset of HRM activities that view employees as investments with differentiated knowledge, skills, and abilities (Lado & Wilson, 1994; Noe et al., 2016; Schuler & Jackson, 1987; Schuler & MacMillan, 1984; Wright et al., 1994) and/or strategic partners in the business (Barney & Wright, 1998; Caldwell, 2008; Holbeche, 2009; Lemmergaard, 2009) who can be a source of sustainable competitive advantage (Lado & Wilson, 1994; Noe et al., 2016; Schuler & Jackson, 1987; Schuler & MacMillan, 1984), rather than deterministic algorithms, where the primary aim is to reduce labour costs.

Regarding the practice literature, this review revealed a similar pattern to that found with the academic literature. Of the 222 articles, 78 were scheduling, allocation and/or routing related, while 22 described HR-related automation, including automatic form filling and audio transcription. Thus, the subsequent focus was on the remaining 122 articles (a listing of the excluded articles is available from the author).

COMPARISON OF THE USE OF ALGORITHMS BETWEEN HRM RESEARCH AND PRACTICE

Research Articles

In the 22 remaining journal articles, algorithms are applied across many different HR concerns, in outlets that span multiple disciplines including management, industrial relations, operations management, economics, information systems, and statistics. Table 1 displays the outlets along with the topic(s) of interest, and the stated purpose of HRM algorithm.

Insert Table 1 about here.

At the individual-level of analysis, algorithms are widely used in the description and prediction of work-related attitudes such as job satisfaction (Aouadni & Rebai, 2017; Becker & Ismail, 2016; Grilli & Rampichini, 2007; Hsiao et al., 2015,?; Kuron et al., 2016; Liu et al., 2015; Somers & Casal, 2009). Relatedly, algorithms have also been developed to predict motivation (Canós-Darós, 2013) and employee turnover (Koch & Rhodes, 1981; Wang et al., 2017). In recruitment and selection,

an algorithm has been used to predict future performance outcomes when a quota is imposed for minority group hiring (Kroeck et al., 1983). In training and performance management, algorithms have been used to rank the importance of HR capabilities against developmental needs (Lin & Hsu, 2010), and to predict competency gaps in the management of software engineers (Colomo-Palacios et al., 2014).

At the macro-level, algorithms have been used to describe how ownership separations and acquisitions influence the composition of HR of organizations (Boudreau & Berger, 1985). They have also been used to reduce HR overhead (Strub et al., 1994), to optimize HR-related investments (Gutjahr, 2011), and to analyze the alignment between various HR practices and the strategic capabilities of small and medium-sized enterprises (Fabi et al., 2009).

Other applications include efforts to predict labour force participation (Hall et al., 2004). Bidding and arbitration behaviours in final offer arbitration have been modeled (Gerchak et al., 2004; Swartz, 2003). Hatfield and Milgrom (2005) endeavour to match contracts between individuals and organizations (Hatfield & Milgrom, 2005).

HRM Research vs. Practice: Divergence by HR Function

Articles from trade journals and the popular media also reference the use of algorithms across a variety of HR functions. To compare the topics targeted in these literatures, I grouped the studies in terms of commonly referred to HRM concerns. For example, the Job Attitudes category includes algorithm studies related to job satisfaction, motivation, engagement, and happiness, while Recruitment and Selection includes algorithmic resume screening, selection algorithms, and online job matching. In Figure 2, I use percentages to compare the literature pools given the large

discrepancy in the total number of articles in favour of practice over research.

Insert Figure 2 about here.

Figure 2 reveals that use of algorithms in areas such as Performance Management and Turnover draws attention from both researchers and practitioners, whereas in areas such as Job Attitudes, Collective Bargaining, Labor Participation, and Strategic HRM, there has been little or no interest among practitioners, relative to researchers. In contrast, the HRM functions of Recruitment and Selection, Training and Development, and Compensation have attracted more interest from practitioners than researchers. Accordingly, I now turn to a consideration of the research opportunities that exist in relation to these three areas of practitioner interest.

Research Opportunities

In the course of reviewing both the scholarly and application-oriented literatures by HR function, several areas especially in need of research became apparent. I now consider some of these possibilities.

Recruitment: Blind Hiring Algorithms This review reveals a growing trend in the use of "blind" screening algorithms to eliminate unconscious human bias by removing demographic markers from application materials, as their disparate influence on decision makers has been shown, for example, with regard to ethnicity (Kang et al., 2016) and gender (Moss-Racusin et al., 2012). While research designed to find viable interventions with a demonstrable positive impact are rare, I found media

reports (e.g., Erts, 2018) concerning the use of automated text processing software to reduce the cost and effort of conducting blind resume screening. In addition to fostering the blind review of candidates, there are algorithmic techniques that help remove gender bias in job listings. Textio, a software developer, analyzes the wording in the job listings that may inadvertently attract one gender over another (Silverman & Gellman, 2015). Their analyses, for example, suggests that use of the term "rock star" may attract more males than females, and that the phrase "high performer" should be used instead.

Training and Development: A Bottom-up, Self-driven Training System The analysis of the practice-oriented literature revealed use of bottom-up training algorithms that empower employees to make decisions concerning the training content required for their jobs and/or make suggestions concerning training needs to their employer. As described by Walker (2012), statistics gathered from current and former Google employees are used to inform managers of likely training needs at various points in their career (Walker, 2012). Vencat 2006 describes a platform that employees at Cisco used to distribute videos, including content from YouTube channels, to promote learning across teams. At Whirlpool, a digital platform allows engineers to immediately create interactive webcast tutorials to correct product flaws which can be shared with other employees across 70 countries (Vencat, 2006). Among other opportunities, the research community has yet to address the impact of bottom-up, self-driven approaches to on-the-job training enabled partly by algorithmic platforms.

Compensation Relative to other HR concerns, the topic of compensation has been a neglected area of research (Gupta & Shaw, 2014). For example, a meta-analysis

of research concerning the impact of financial incentives on job performance found only 39 studies over a 40-year timeframe (Jenkins Jr. et al., 1998); this relative lack of research interest remains unchanged (Gupta & Shaw, 2014). Relatedly, I did not find any research reflective of the practitioner-oriented trend of applying algorithms to designing compensation systems. For example, Google has been using a predictive algorithm to reduce attrition by making timely, flexible adjustments to compensation packages (Silverman & Gellman, 2015). In the United Kingdom, large banks are evaluating a pay system based on multilevel modelling that captures regional variations to attract and retain talent in different locations (Times, 2000). More broadly, HelloWallet offers an online diagnostic algorithm to compare employer-offered salary and benefits against open data sources from government (HelloWallet, 2012). From a research perspective, the impacts and effectiveness of these efforts are open questions.

INVESTIGATING THE NATURE OF HRM ALGORITHMS

Are HRM Algorithms Black Boxes?

In addition to variations between the research and practitioner literatures regarding areas of relative interest, the nature of algorithms themselves tend to be portrayed differently. It is notable that the term "black box", while uncommon in the research literature, is widely adopted by practitioners, especially those who have a general discomfort that these applications are producing solutions that are mystical in nature (Boulton, 2017; Johnson & Ruane, 2017; Pasquale, 2015; Wilson & Daugherty, 2017). Wilson, Daugherty, & Morini-Bianzino (2017), for example, note that the "black box"

nature of sophisticated algorithms made many executives uneasy, especially when the resulting recommendations conflicted with conventional wisdom. Relatedly, Johnson and Ruane (2017) note that hidden bias in an algorithm cannot easily be detected since "you cannot simply read the code to analyze what is happening." Boulton (2017) also warns against blind trust in algorithms generally, including in HRM contexts, because if "you're making decisions that impact peoples' lives you'd better make sure that everything is 100 percent."

In comparing the literature pools, the relative differences concerning references to a black box might be partially because the practitioner-oriented articles rarely discuss the details of the algorithms involved. In comparison, due to differences in mission and readership, the algorithms in the research-oriented pool were typically clearly-defined. The information provided usually consisted of detailed guidelines generated by humans and communicated to computers via coding. This is important since any missteps in the process will result in software failures. To address the level of clarity and transparency, I reviewed each of the 22 research papers for the degree to which the analysis protocols were clearly-defined with regard to the: (1) independent variables and dependent variables used, (2) methods of processing unstructured data (e.g. graphs/video/sound) involved; (3) goal of estimation/optimization (minimization or maximization) used, and (4) sequence and priority of calculations.

First, most research articles clearly stated the independent and dependent variables used. For example, the most complex model used 78 factors reflective of eight categories as antecedents to predict employee motivation (Canós-Darós, 2013). Two papers employed historical data to predict the future decision-making of arbitrators, such that the algorithms involved could accommodate all previous events if necessary.

In terms of dependent variables, most papers used only a single dependent variable though some used as many as three (Fabi et al., 2009; Hsiao et al., 2015). The simplest algorithm in the pool used two independent variables (home life and work attitude) to predict a single dependent variable, job satisfaction (Liu et al., 2015). One exception to the high level of clarity concerning the variables involved was Lin and Hsu (2010) who used the generic term "Decision Support System" instead of disclosing the details of their algorithm. Collectively this review revealed that when researchers use the term "algorithm" in place of traditional models, the algorithm typically includes more independent variables than is common in traditional HR research.

Second, none of the articles in the pool involved unstructured data. Nonetheless, there is an increasing trend to using unstructured data in fields such as social psychology. For instance, computerized-text analysis methods, including Linguistic Inquiry and Word Count (LIWC), were created and validated to count words in psychologically meaningful categories (Tausczik & Pennebaker, 2010). Related methods are widely used in marketing (Ludwig et al., 2013), strategic management (Crilly et al., 2016; Nadkarni & Chen, 2014), and health management (Monrouxe et al., 2014). Image processing represents another new method of unstructured data analysis adopted by social psychologists. Wang and Kosinski (2018), for example, used deep neural networks to predict human sexual orientation from facial images shown on a dating website. Use of wearable sensors to measure geo-spatial data for social network analysis is also an emerging field in management research (Chaffin et al., 2017; Tonidandel et al., 2018). Use of these methods in HRM scholarship is either absent or nascent.

Third, although the goals of algorithm estimation or optimization varied across studies, they were typically clearly stated. Several used algorithms to find the best

fitting model to describe their data (e.g., Becker & Ismail, 2016; Canós-Darós, 2013; Colomo-Palacios et al., 2014; Hall et al., 2004; Hsiao et al., 2015; Koch & Rhodes, 1981; Somers & Casal, 2009) and for maximizing prediction. Some used algorithms for clustering (Fabi et al., 2009; Kuron et al., 2016) or generating simulated datasets and assessing proposed models (Boudreau & Berger, 1985; Kroeck et al., 1983). Others aimed to minimize the sum of measurement errors as part of a construct measurement, e.g. measuring multiple facets of job satisfaction (Aouadni & Rebai, 2017).

Fourth, the sequence or priority of operations in the algorithms used was also typically detailed. This includes the order of calculation in applications of Neural Networks or Genetic Algorithms (Aouadni & Rebai, 2017; Colomo-Palacios et al., 2014; Somers & Casal, 2009), two-step clustering algorithms (Fabi et al., 2009; Kuron et al., 2016), iterations of estimating likelihood functions (Gerchak et al., 2004; Swartz, 2003), the parameters and steps used in a simulation (Boudreau & Berger, 1985; Kroeck et al., 1983), the selection of parameters or models following a specific preference, e.g. using correlations instead of R^2, assigning smoothing parameters from large to small, and matching parameters to their weighting in the population (Becker & Ismail, 2016; Hall et al., 2004; Hsiao et al., 2015). Only two studies (Lin & Hsu, 2010; Strub et al., 1994) did not detail the mathematical decision-making embedded in their algorithms.

Given the above, I suggest that the "black-box" perception associated with using algorithms is primarily a reflection of the techniques used to process complex data. Factors contributing to the complexity of algorithms include the translation of graph/video/sound into binary variables, the automatic clustering of data points, the automatic assignment of various weights to large numbers of variables, and the

generation of a global solution despite local heterogeneity. Thus, a complete understanding of the underlying computational complexity associated with some algorithms requires detailed knowledge from a variety of fields, including engineering, mathematics, and/or computer science. Importantly, since most management researchers lack this background, the "black box" label is appropriate to a limited extent, underscored by the lack of basic information concerning the algorithms described in the practice literature.

The Role of Theory is Deemphasized in the use of HRM Algorithms

In the research I reviewed, algorithms largely take the place of traditional statistical models, typically without highlighting the differences involved. This is important because the use of traditional statistical approaches such as multiple regression (Cohen et al., 2003) are intimately linked with the desire to test a theory. Specifically, a complete theory has at least four essential "building blocks" – factors (variables, constructs, concepts), mechanisms (causal relationships), rationale (underlying psychological, economic, or social dynamics), and contextual conditions (who, where, when) (Dubin, 1978; Whetten, 1989). Hence, the statistical approaches traditionally used in HRM research have been chosen to align with the testing of theory reflected by "boxes (constructs)" and "arrows (causal relationships)", and the underlying rationale and boundary conditions being sufficiently discussed. In comparison, theory is downplayed in HR-related algorithmic applications; underlying complex calculations take on the role of model-building instead. Thus, for example, with the exception of Becker and Ismail (2016) who specify that their study is intended to assess an existing model of job attitudes (Hult, 2005), most of the research in the database

emphasizes *what* their algorithms are capable of doing (e.g. handling complex antecedents, dealing with a biased sample, processing categorical or ordinal variables, predicting nonlinear relationship between variables) and/or their mathematical deduction. Theory-related discussion was either minimal or non-existent, especially in studies involving more than 10 antecedents.

In comparison, in applying traditional HRM models, researchers first form causal hypotheses derived from theory. The hypotheses typically consist of a causal description between theoretical constructs, for instance, job satisfaction, turnover intention, and turnover behaviour. Some of the constructs may be directly observable (e.g. turnover—whether the employee left the organization); others are not (e.g. turnover intention). As such, researchers attempt to operationalize the unobservable constructs to bridge theoretical constructs with observable measurement. Only then can they collect observational data based on justified observable measurement, make statistical inferences, and draw statistical conclusions concerning the statistical significance of effect sizes. These statistical conclusions are essentially associations with strong theoretical causal support and are eventually interpreted as research conclusions for making practical or policy recommendations. In sum, the major source of inferring causality in traditional HR research is through theoretical discussion such that much of the research relies heavily on theory to support causal arguments.

In some of the research I reviewed predictive modeling, defined as a statistical model or data-mining algorithm for the purpose of predicting new or future observations (Shmueli, 2010), was the explicit goal. For instance, the stated aim of Hall, Racine, and Li (2004) was to improve the mathematical predictive power of categorical variables using nonparametric methods, using female participation in the workforce

as an example. Colomo-Palacios et al. (2014) used Artificial Neural Networks to predict the competency gaps in key management personnel when the relationships among variables of interest were nonlinear. Hsiao and his colleagues (2015) applied Fuzzy Set Qualitative Comparative Analysis to predict happiness-at-work and the job performance of hospitality employees. A crucial feature among all these predictive models is the temporal forecasting of the dependent variables involved. Nonetheless, predictive modeling has been criticized as "atheoretical" or "unacademic" and is sometimes disregarded for the purposes of theory testing (Shmueli, 2010). Often, the debate concerns "whether prediction per se is a legitimate objective of economic science, and also whether observed data should be used only to shed light on existing theories or also for the purpose of hypothesis seeking in order to develop new theories" (Feelders, 2002, p. 174). Importantly, many statisticians emphasize the value of statistical prediction (Findley & Parzen, 1998; Friedman, 1997), noting that observed variables can have greater relevance in estimation than artificially constructed ones (Geisser, 2017). In the end, predictive modeling based on more powerful calculation capabilities to analyze large, rich datasets may give rise to new hypotheses and help uncover new causal mechanisms useful in theory testing (Shmueli, 2010).

Some of the research I found was descriptive. In descriptive modeling, theoretical discussion is either absent or informally characterized (Freedman, 2009). Relative to predictive modeling, where the aim is to improve predictive accuracy, descriptive modeling is focused on fully capturing the associations among relevant variables and fitting a regression model. Strictly speaking, descriptive models aim to present data in a succinct manner, but not for the purpose of causal inference or prediction per se. For instance, Hsiao et al.'s (2015) study involving the use of an algorithm with

four conditions or antecedents that identify and characterize a high performing hospitality employee is descriptive. In fact, none of the antecedents involved (including whether the frontline employees are happy at work; whether they work well with other employees, whether they ever cause peer conflicts, or arrive to work on time) were subject to a rigorous causal analysis or theoretical discussion as part of the algorithmic estimation. In other words, I can infer from the dataset that people in a specific organization who fit a set of criteria happen to be high performers, but I cannot say with certainty that people in the target organization who fit these criteria will be high performers in the future, nor can I conclude that if I help them improve on one or more of the antecedents (e.g. warn them to avoid conflict with co-workers, encourage them to arrive at work on time), that improved performance will result.

In summary, most of the algorithm-related HR research is either descriptive or predictive, and hence cannot be classified as theory-driven. Causal testing as suggested by theory is not evident in the reviewed journal articles. Rather, the mechanisms connecting the variables involved are presumed unknown. Discovery without *a priori* assumptions dominates this literature.

HRM Algorithms Are Heuristics

Algorithms used in HRM might best be characterized not as "black boxes" but as "glass boxes" because they are reflective of some, but not all of the components associated with a theory. That is, these algorithms usually consist of a large list of clearly-defined variables, drawing out certain predictive or descriptive "patterns" without defining their causal direction. Most of the research either predicts variables of interest to the HR function (e.g., competence gaps; employee happiness; labour

participation, turnover, etc.) or provides a descriptive estimation of them (e.g., job satisfaction, alignment between HRM and strategic capabilities, etc.), without addressing matters of causality. Only in one paper was it noted that the level of prediction achieved might vary considerably as a function of changes in the historical data used (Colomo-Palacios et al., 2014); yet even here there was no explicit statement of boundary conditions. As such, much of the algorithmic HR research approximates a theory, while attempting to maximally account for the phenomenon in question from incomplete sources of information. Hence, most of the research applications found in my review are heuristics, which by definition, are approaches to problem solving that offer sufficient, practical solutions that are not necessarily optimal or perfect (Kahneman et al., 1982).

Heuristics are a compromise of two criteria, i.e., the need to use simple methods requiring less resources, and the need to sufficiently distinguish between good and bad choices (Pearl, 1984). For humans, they relieve the cognitive load associated with decision making (Kahneman et al., 1982); for computers, they reduce computational time. Examples include the educated guess, intuitive judgments, rules of thumb, or common sense (Pearl, 1984). While heuristics work well under many circumstances, they are relatively simplistic in nature, which can result in bad choices. For instance, a rule of thumb is that white mushrooms are edible, yet some are actually highly toxic and could be deadly to humans. In HRM, stereotyping and racial profiling are examples of heuristics.

Computer software reflective of heuristics can also generate erroneous judgements or trade-offs to, for example, emphasize computational speed and efficiency. Recommendation algorithms used by major news websites may capture a pattern of reading

baseball stories and start to "push" news items regarding various baseball teams to the user, without knowing that the user is interested only in one team. Most importantly, as aforementioned, HRM algorithms of a probabilistic type are not exempt from errors and biases. For example, in an organization with a dominant demographic group (e.g. young, Caucasian, males), an algorithm programmed to predict "a good candidate" based on the past performance of employees may inappropriately favor the demographics of the existing workforce.

Despite the possible biases associated with the application of heuristics (e.g., Kahneman et al., 1982), some researchers (Gigerenzer & Goldstein, 1996; Gigerenzer & Todd, 1999) highlight their positive aspects, i.e., that in an uncertain world, smart heuristics or rules of thumb help us adapt to the environment by contributing to better decisions with less effort. To illustrate, Gigerenzer and Selten (2002) identified situations in which "less is more" by illustrating the value of several heuristics as part of a "fast and frugal" toolbox. In developing the toolbox, they created a series of algorithms in which less knowledge can make for better prediction. The hypothetical "secretary problem" is an example, in which applicants are randomized and interviewed one by one. To maximize the chance of selecting the best applicant, the first 37% of the applicants should be sampled, followed by selecting the first candidate thereafter who is regarded as better than all the previous ones (Seale & Rapoport, 1997; Todd & Gigerenzer, 2003). Though this heuristic likely could not be used in practice given legal and ethical constraints, it illustrates the counterintuitive argument that some simple algorithms can outperform more complicated ones despite the reduced effort, such that simple heuristics can "make us smart (Gigerenzer & Todd, 1999)."

In fully considering the findings of my review, I propose the following definition of an HRM algorithm:

HRM algorithms are computer programs of a heuristic nature[2] *that use economical input of variables, information, or analytical resources to approximate a theoretical model, enabling an immediate recommendation of screening, selection, training, retention, and other HR functions.*

ADVANCING HRM THEORY & PRACTICE USING ALGORITHMS AS HEURISTICS

Given the view that HRM algorithms are most appropriately regarded as heuristics, I now consider some of the related implications with regard to inferring causality in research and avoiding adverse impact when HRM algorithms are used in practice, and in theory building.

Algorithms and Causality Issues in HRM

As discussed earlier, it is important to acknowledge that HRM algorithms are predominantly predictive or descriptive. Therefore, unlike empirically-tested theories grounded in causal inference, these algorithms should not be used in isolation to directly inform decision-making. Confidence in decision making can be fostered by gaining an understanding for previously uncovered patterns among variables to avoid inferences based upon spurious relationships and endogeneity issues. Two common

[2]Please note that here by heuristics I refer to *how* researchers use algorithms instead of *what kind* of algorithms they are using. I am aware of the distinction in math and computer science on exact algorithm, heuristic algorithm, and approximation algorithm, however I am not referring to heuristic algorithm here.

causes of endogeneity are: (1) confounding variable(s) that mask the true underlying cause of the relationship between independent and the dependent variables; and (2) reverse causality involving variables in the model. Importantly, manipulating independent variables based upon spurious findings will not necessarily result in the anticipated effects on the outcome variables of interest. A classic example of the impact of a confounding variable involves the positive correlation between city-based ice cream sales and the rate of swimming pool drownings, with more sales and drownings linked with high summer temperatures. Reducing ice cream sales during summer months would have no impact on drownings.

There is potential for some relationships uncovered by HRM algorithms to be an unsound basis for decision-making. For instance, at Deloitte, Bersin (2013a) used textual analysis to identify several factors correlated with success among sales professionals, one of which was the lack of typographical or grammatical errors on resumes. Nonetheless, Cheng (2017) cautioned that a selection policy based on this variable is risky. This results from the implication that typos and grammatical errors are a suitable proxy for a latent quality that has a causal bearing on future sales performance. This may be an inappropriate inference in multinational contexts involving English-as-second-language applicants.

Antonakis and colleagues (2010) suggested six methods for inferring causality in non-experimental settings. Two of the more broadly applicable options involve using statistical adjustments and quasi-experiment designs. Statistical adjustment or measuring and controlling for all possible causes of y, is the simplest way to help ensure causality inferences (cf. Angrist & Krueger, 1999). However, while it is relatively

easy to control for some variables (e.g., employee demographics) it is virtually impossible to rule out all the possible causes of variance in y. Variations in childhood experience, for example, may contribute to developing individual emotional processes and personality, but it is not viable to design an algorithm for employee selection that would account for this variation.

The use of quasi-experiment designs to test for causality lacks the element of random assignment to the treatment or the control group that characterizes the gold standard of experimental design. To help deal with this shortcoming, Antonakis et al. (2010) suggest that simultaneous-equation models, regression discontinuity, difference-in-differences models, as well as Heckman selection models, be used to establish causality when certain assumptions are met. Table 2 briefly describes the essence of each of the highly recommended approaches.

Insert Table 2 about here.

Algorithms and Adverse Impact in HRM Practice

Understanding that HRM algorithms are essentially heuristics is crucial to avoiding the potential for adverse impact that they may introduce to HRM practice. For instance, it would be fair to say that the algorithms that produce fast recommendations concerning selection likely involve stereotyping or profiling. Hall, Racine & Li (2004), for example, demonstrated a non-parametric method in which demographic variables were included as independent variables (e.g. age, number of children). To guard against the replication of historical biases and the potential for adverse impact against protected classes, policies are needed that delineate boundaries around how

these algorithms can be used to inform decision-making. There is also the potential for adverse impact even when demographics are not directly involved. For example, in Hsiao et al. (2015), a person who never arrives late to work fits the top performer criteria, but it may be that such a person is especially likely to be single or married without young children.

Interestingly, in 2016, the European Union (EU) ratified the General Data Protection Regulation (GDPR) intended to protect EU residents guaranteeing them more control over their personal data and in terms of its movement, collection or processing by an organization (Tankard, 2016; Wachter, 2018). One aspect of the regulation requires that any algorithms used for decision making must be explainable by those who engineered them (Kean, 2018). Aligned with the spirit of the GDPR, I propose that not only the collection of data related to individuals be protected, but also that HR-related algorithms be regulated such that employers be required to: (1) disclose to applicants/employees the nature of any decision-making made solely by algorithms; (2) ensure the right of individuals to contest the outcome of decisions based solely on algorithms; and relatedly, (3) have sufficient expertise to address challenges to algorithmic decision making.

Algorithms and HRM Theory Building

While algorithms do not typically allow for strong causal inferences, this does not mean that they are necessarily irrelevant to developing and verifying HRM theory-based models (Shmueli, 2010; Shmueli & Koppius, 2011). They can, for example, provide the foundation for new hypotheses, uncover new measures, suggest improvements to existing models, and help in assessing the explanatory power of theories. Using

Hsiao et al. (2015) as an example, it might be hypothesized that conscientiousness (arriving on time) interacts with collegiality (avoid conflict, work well with others) to ultimately enhance performance. In the process, the extent to which on-time arrival reflects conscientiousness in the workplace could be evaluated. New measures and new hypotheses, in turn, may result in improvements to predictive modeling of job performance. Predictive modeling can also offer a straightforward way to compare competing theories by comparing their predictive power and thereby informing future research.

Importantly, it is also possible to extract considerable value from descriptive or predictive algorithms to inform decision-making. Specifically, as implied earlier, algorithm-based research can be viewed as an exploratory step in the quest to establish causality, and/or as a post-hoc effort to assess causal theories. Over time, I should look to establish methodological standards regarding the use and interpretation of algorithms based on a clear understanding of the types and functions of the various algorithmic models. Triangulation between causal and non-causal modeling can be valuable in this regard. It should be possible, for example, to develop a weighted algorithm that draws from various theories of employee performance and turnover (cf. Harrison, Virick, & William, 1996) and develop new measures for constructs such as conscientiousness and honesty that are less susceptible to social desirability or faking. Given that I located only 22 high-quality research papers, there are obviously many remaining opportunities for scholarly work.

DISCUSSION AND CONCLUSION

Research is required to better define the nature and characteristics of the range of algorithms used, especially in HRM practice. Also, as implied earlier, the findings in this regard are likely to have regulatory implications from a public policy perspective, especially as the popularity of algorithmic applications in HRM practice grows. For example, to what degree is adverse impact associated with HRM algorithms used in practice? If it is a problem, to what degree can the biases be reduced? What level of accuracy would be considered "good enough" and does the answer depend on the specific HR issue under consideration? Are there variables that should not be included in algorithmic models intended to predict human behaviours, e.g., ancestry information that ends up in the public domain? At present, questions of this nature are not attracting much attention, but there is potential for that changing. In a manner analogous to the pressures social media companies such as Facebook are facing tied to violations of privacy that were contrary to stated policy (Gallagher, 2018), biased algorithms may conflict with stated organizational diversity goals. Moreover, research has already begun concerning the steps that organizations may need to take in order to increase the perceived authenticity of algorithms (Jago, 2019).

As documented earlier, resulting from the ready availability of the tools required, there is a research opportunity concerning HR-related algorithms that tap into unstructured data. The gaps include, but are not limited to, the analysis of text, images, spatial data, voice, and video. There is a need to evaluate the extent to which these sources of data can inform HRM decision making. HRM researchers are in danger of falling behind other fields in this regard, and possibly HR practice as well. Hitachi, for

example has been using social badges or sensors to monitor their employees' mood and social interaction for years, and there are claims in the practitioner literature that 60% of companies are already using applicant tracking systems that apply textual analysis techniques in recruitment and selection (Bersin, 2013a; Bersin et al., 2016). The implications of these and other practices need to be evaluated from a research perspective. Textual analysis, for example, could be used to extract multiple variables that may be relevant to recruitment and selection, while spatial data may be useful in analyzing social networks for team building and facilitating knowledge sharing.

The lack of research in this regard is consistent with the overall finding that the growth of scholarly interest in HR-related algorithms pales in comparison to the surge of applications in HR practice. As such, I have the beginnings of another wide gap between HRM research and practice. By targeting the research opportunities identified in this review, researchers can seize the opportunity to close the academic-practitioner divide concerning use of HR-related algorithms and avoid having their contributions disparaged as "arcane" (Walsh et al., 2007) and "ceremonial" (Bartunek & Rynes, 2010).

Chapter 3 FROM COUNTERFACTUAL FAIRNESS TO ALGORITHMIC FAIRNESS: BUILDING PRINCIPLE OF EQUITY IN AI MANAGEMENT

ABSTRACT

The issue of fairness associated with the use of algorithms in HRM has only recently gained significant scholarly attention. This paper contributes to this field in several ways. First, potential sources of unfairness are identified, with emphasis on legal considerations. Second, the concept of counterfactual fairness, commonly used as a criterion for fairness in computer science, is critically compared with alternative principles rooted in equity theory and justice theory from management and psychology. Third, this paper presents a case for using causal modeling to fully evaluate the fairness of HR algorithms. Fourth, and relatedly, I illustrate how causal modeling can usefully be applied to evaluate several aspects of legal compliance. Fifth, in line with the call for HR practitioners to build a deeper insight concerning the construct, I contend that fairness is a multifaceted, dynamic concept, entailing the consideration of moral issues, beyond legal compliance per se. Potential ways to enhance perceptions of justice concerning the application of algorithms for HR decision-making by employers are suggested. Finally, in concert with offering a broadened definition of algorithmic fairness, a five-step model is proposed to help ensure its integration into the use of HR algorithms.

Keywords: AI management; business ethics; algorithmic fairness; algorithms; HR analytics

INTRODUCTION

Recent decades have witnessed major advances of artificial intelligence (AI) applications in computer science, engineering, and health sciences (Frutos-Pascual & Zapirain, 2017; Mellit et al., 2009; Vaishya et al., 2020). Automated algorithms have been used in tasks including the game of Go, among others (Frutos-Pascual & Zapirain, 2017; Silver et al., 2016), medical diagnoses and classification, including cancer imaging and COVID-19 screening (Esteva et al., 2017; Vaishya et al., 2020), autopiloted driving (Hecht, 2018), quantum cryptography (Aerts & Czachor, 2004), new material selection (Jahan et al., 2010; Zhou et al., 2009), and customized recommendations for individual internet users (Lee et al., 2002; Liang et al., 2008). Notably, these tools have been improved significantly over time in terms of their accuracy and efficiency, thereby informing human decision-making and fostering paradigm changes in various disciplines.

Increasingly AI is being used to facilitate decisions that can have a huge impact on the everyday lives of people, such as predictive policing and court decisions (Brantingham et al., 2018; Medvedeva et al., 2020; Shapiro, 2017), loan lending (Gorton & Pennacchi, 1995; Sachan et al., 2020), insurance-related determinations (Dhieb et al., 2020; Lamberton et al., 2017), and evidence-based management(Eapen et al., 2023). Despite these achievements, a major concern is that AI models tend to replicate and/or amplify human biases, including in the context of HR decision making (Vassilopoulou et al., 2023), which, depending on jurisdiction, involves several legally

protected attributes, including race or ethnic origin (Peña et al., 2020), gender (Buolamwini & Gebru, 2018), and sexual orientation (Wang & Kosinski, 2018). For example, Amazon's automatic HR tools were found to favour male job candidates over females due to biased historical data leading to lower scores for female candidates and penalization for using terms such as "women's" in resumes (Dastin, 2018). Also in HR, Google's ad-targeting algorithm proposed higher salaries in executive openings for men relative to women (Datta et al., 2015; Simonite, 2015). The use of AI models in HR contexts has also posed ethical issues involving data privacy, labour relations, and user perceptions. In particular, AI tools tracking employees' performance have been criticized as intrusive workplace surveillance (Moore et al., 2018), abusive of employee rights (Kellogg et al., 2020), that create a false sense of trustworthiness (infallibility) among users (Glikson & Woolley, 2020; Vassilopoulou et al., 2023). These issues are of great importance, despite the fact that the ethical considerations associated with the use of AI can come across as an afterthought among practising managers (Hao et al., 2023). Moreover, what algorithms do in the background is still often a black box to the affected employees, and sometimes even to the employer.

This paper concerns a range of legal, causal and moral issues tied to the use of algorithms in HR, which refers to "software that operates on the basis of digital data to augment HR-related decisions and/or automate HRM activities" (Meijerink et al., 2021). As noted by Cheng and Hackett (2021) and Vassilopoulou et al. (2023), the multifaceted issue of fairness associated with their use has only recently begun to receive serious consideration. As such, I add to the literature in several ways. First, a list of potential sources of unfairness, with emphasis on legal considerations is offered; it reflects many of the same considerations highlighted by Vassilopoulou et al. (2023).

Second, in line with the call of Vassilopoulou et al. (2023), that we not be lulled by scientism into a sense of complacency in the use of use of HR algorithms, I critically compare the concept of counterfactual fairness, widely used in computer science (Pearl & Mackenzie, 2018; Pessach & Shmueli, 2020) as the criterion for fairness, to alternative rules from the management and psychology literatures, based on equity theory, equity (e.g., Adams & Freedman, 1976) and justice theory (Cohen-Charash & Spector, 2001; Newman et al., 2020; Viswesvaran & Ones, 2002). As I will detail, there are both similarities and important differences among these criteria. Third, I present a case for using causal modeling[1] to fully evaluate the fairness HR algorithms. Fourth, and relatedly, I illustrate how causal modeling and related analyses can usefully be applied to evaluate several aspects of legal compliance (e.g., Walsh, 2023) as related to HR algorithm use. I contend that algorithms be grounded in well-established causal models, as the possibility of discrimination is not fully considered without their application. Fifth, in line with the call for HR practitioners to build a deeper insight concerning the construct (Vassilopoulou et al., 2023), I argue that fairness is a multifaceted, dynamic concept (Cook & Hegtvedt, 1983; Jones & Skarlicki, 2013), that entails the consideration of moral issues, beyond legal compliance per se. I also suggest some ways employers can improve justice perceptions[2] concerning the use HR algorithms, a matter that has proven to be challenging (Newman et al., 2020). Finally, in concert with offering a broadened definition of fairness, I propose a five-step model to help ensure that fairness is built into HR algorithms.

[1] Sometimes referred to as Structural Causal Modeling (SCM), which establishes the mathematical basis for studying complex causal relationships (Pearl, 2009, 2010) that are not restricted by the assumptions of normal distribution and linear relationships. Note that SCM is not equivalent to structural equation modeling (SEM), and SEM can be seen as a special case of SCM.

[2] A large research interest has emerged around understanding the perception of algorithmic decisions (c.f. Lee 2018), and it is worth noting that this dissertation does not investigate this subfield as it takes solely an artifact-oriented lens.

SOURCES OF UNFAIRNESS IN MANAGEMENT ALGORITHMS

As summarized in Table 3, researchers from various fields have identified sometimes overlapping sources of potential bias in algorithms, especially those based on machine learning (Chouldechova & Roth, 2018; Cowgill & Tucker, 2020; Martínez-Plumed et al., 2019; Pessach & Shmueli, 2020). Many of the sources in this non-exhaustive list, which reflects the work of others (e.g. Vassilopoulou et al., 2023), involve unfairness tied to legal issues that ultimately also have moral implications. It does not include biases originating with programmers unaware of their own unconscious discriminatory tendencies (Floridi et al., 2015; Mittelstadt et al., 2016; Pessach & Shmueli, 2020), or technical design considerations where for example, HR algorithmic platforms which may discriminate against certain individuals because of the specific ways users must engage with it (Vassilopoulou et al., 2023).

1. Causal issues (historical data input)

a. Non-uniform noise: Biases already ingrained in the datasets collected, originating from, for example, biased measurement processes, historically biased internal and external human decision making, erroneous reporting. Machine learning algorithms are essentially designed to replicate these biases.

b. Missing data: Missing values or sample/selection biases can result in datasets that are not representative of the target population.

2. Legal issues (algorithmic pathways and optimization objectives)

a. Proxy variables: As certain characteristics including race, gender and age, cannot legally be used in decision making, proxy characteristics, which on their face are legal, may be used in their place resulting in bias.

b. Algorithmic feedback loops: Actions derived from an algorithm may result in the contamination of future training data when the predictions are codified as "ground truth" (Cowgill, 2019; Cowgill & Tucker, 2020). For example, judges in the legal system may have been unduly influenced by "high risk of recidivism" labels generated by "COMPAS" algorithms, resulting in additional detention time to defendants (Cowgill, 2019), which in turn may have led to small increases in two-year re-arrest rates. Such "self-fulfilling prophecies" can occur as the influence of originally contaminated data input is amplified over time as the algorithm continues to be used.

c. Conflicting objectives: Organizations may have pre-exiting notions concerning the types of employees that "fit", which operate to exclude protected groups. Moreover, the interests of society overall can conflict with those of minorities. To the extent that such conflicts are programmed into algorithmic objectives to minimize aggregated prediction errors, majority groups benefit relative to minorities.

3. Moral issues: Algorithmic objectives

a. Conflicting objectives between organizations and their employees: When algorithms are applied to manage crowd-work (e.g. online freelancing platforms like Fiverr or Upwork) or work-on-demand (e.g ride-hailing services like Uber or Lyft), the underlying working conditions associated with the platform may deviate from those more favourable to the employees, as established for example, via collective bargaining (Birgillito & Birgillito, 2018). For example Uber's use of algorithms to oversee drivers, may result in unfair treatment of drivers stemming from algorithmic parameters such as, earnings of the ride,

surge pricing for large events, or time to wait for the passengers to maximize profits, irrespective of the working conditions.

Insert Table 3 about here.

ALGORITHMIC FAIRNESS IN COMPUTER SCIENCE VS. MANAGEMENT

It has been argued that fairness concerns associated with the use of HR algorithms have not been adequately addressed partly because of scientism, i.e., "the unfounded prioritization of scientific method over and above other moral and reasoned arguments" (Vassilopoulou et al., 2023). Indeed, in considering this possibility there is a need to critically consider the concept of counterfactual fairness, which is widely used in the computer science literature (Chiappa, 2019; Kusner et al., 2017). It refers to the fairness of a prediction or decision made by a model, when counterfactual or alternative situations are considered (Pearl & Mackenzie, 2018; Pessach & Shmueli, 2020). Counterfactual fairness involves the evaluation of whether an algorithmic prediction or decision would have been the same *if* the protected characteristics of the target individual being predicted or decided upon were different. Importantly, from this perspective, for an algorithm to be considered counterfactually fair, the same decision should be generated even when the characteristics of the person are changed in the counterfactual world. Thus, counterfactual fairness is a theoretical condition in which an individual with sensitive attribute(s) would be treated exactly the same way in the counterfactual world in which those attribute(s) are swapped with the

dominant ones. By sensitive attributes, I am referring to any individual characteristics that may prompt discrimination or unfair treatment. These include, not only all prohibited grounds (e.g., race, ethnicity, gender, and age; Walsh, 2023) that could be used as the basis of a legal claim of discrimination, but also other characteristics that could result in the unfair treatment, including variables tied to socio-economical background, such as commuting distance and care-giving responsibilities. A widely used example used in computer science to illustrate the counterfactual fairness concept involves a recruitment case at Berkeley University (Buolamwini & Gebru, 2018; Chiappa, 2019) wherein the admission algorithm would be regarded as being fair to females if it gave a female the same probability of admission in a counterfactual world in which the applicant were male.

Popular as the concept of counterfactual fairness is, questions about some of its assumptions have been raised. For instance, Chiappa (2019) contends that *individual choice* should also play a role in assessing counterfactual fairness. For example, as applied to the Berkeley case, female applicants were more likely to be rejected because they tended to apply to departments with lower acceptance rates. This type of additional consideration yields to the concept of *path-specific* counterfactual fairness; it states that a decision is fair to an individual if it coincides with the determination that would have been made in a counterfactual world in which the protected attribute, along the unfair pathways, were different (Chiappa, 2019). The implication is that the admission decision would be fair for females if it would have remained the same had the candidates along the pathways been male.

Others question whether counterfactual fairness for all groups can be achieved simultaneously. For example, in comparing several criteria used to evaluate the fairness

of recidivism prediction instruments, Chouldechova and Roth (2018) demonstrate that all the criteria cannot be simultaneously satisfied when recidivism prevalence differs across groups. They also found that if a type II error for any observation is prevalent and differs across groups, e.g., if more black offenders are caught than white offenders, the finding of empirical fairness may be misleading due to the missing data.

The Berkeley example from the computer science discipline can be generalized so that the applicability of the counterfactual fairness concept to HR decisions can be evaluated. Thus, for example, to the extent that job candidates or incumbents possess the same job-related characteristics (e.g., knowledge, skills, abilities, and opportunities and performance), their algorithmic evaluation should be the same irrespective of any sensitive characteristics, such as their race or gender. It is notable that this counterfactual understanding of fairness has some similarities with equity (Adams & Freedman, 1976)) and organizational justice theories from management and psychology (Cohen-Charash & Spector, 2001; Newman et al., 2020). First, in all these theories, the fairness concept is grounded in a comparison between a central individual and others in the same pool, just as counterfactual fairness compares a person with minority characteristics to a dominant group. Thus, for example, equity theory (Adams & Freedman, 1976) compares a person's perception of their contribution and benefits to others considered as comparators, while the distributive justice component of organizational justice compares the perceived outcomes of a central individual to that of others (Cohen-Charash & Spector, 2001). Second, all three theories focus mainly (counterfactual fairness and equity theory) or partially (organizational justice) on comparisons involving either the objective or subjective distribution of tangible outcomes. Third, these theories all emphasize the importance of a certain level of

impartiality, and the absence of favoritism or discrimination. Similarities notwithstanding, management-based theory differs from counterfactual fairness in significant ways, as I document below.

"Objectivity" and "Universality"

First, relative to management theory, counterfactual fairness appears to assert a fairness standard from a more "objective" perspective. This objectivity lies in the fact that counterfactual fairness is relatively detached from the perceptions of individuals, thereby offering a universal standard. Even so, individual perceptions enter counterfactual fairness via the world view of programmers (Bryson, 2020; Martin, 2019) or the AI research community as a whole. Moreover, HR algorithmic decisions may not typically be perceived as objective by employees, who may view them as inappropriately reductionist in nature (Newman et al., 2020).

Second, while counterfactual fairness aims to construe a pre-determined, clear, and universal standard concerning the meaning of fairness, the discussion is largely limited to the AI community, such that the degree to which the standard is truly universal is unclear. In comparison, equity (Adams & Freedman, 1976) and organizational justice (Cohen-Charash & Spector, 2001) theory *do not necessarily* hold that perceptions concerning fairness are universal; instead, they are viewed as largely dependent on individual perceptions and preferences. As such, universality is implied only to the extent that the data are aggregated across employees to yield an average perception of those in the targeted organization.

Comparing the Principles of Equality, Equity, and Justice

Counterfactual fairness, by definition, is based on an equal distribution of resources (e.g. admission probability, performance evaluation) among different minority/majority groups; it is intuitive to understand and provides a potentially progressive framework to ensure that the treatment of minority groups is at least no worse than that of the majority group. The notion of a counterfactual world also provides a platform for investigating the objective treatment of the protected groups separately. In practice, for programmers, this equalness is easily quantifiable and executable. While there are debates concerning which mathematical equalness standard (e.g. equal error or any other parameters) should be applied to assess the algorithm, (Chouldechova & Roth, 2018), the quantifiability of equalness enables the programmers to test, evaluate, and model a method of resource distribution to ensure the outcomes of the algorithms are counterfactually fair. In essence, counterfactual fairness conforms to the principle of equality, which refers to the idea that everyone is entitled to the same rights, opportunities, and resources, regardless of their individual circumstances or needs. Notably, while the focus is on the equal distribution of resources the concept *fails to* consider the specific needs of individuals, including the accommodations they might be legally or morally entitled to. Instead, the predictions or outcomes of the model are driven *purely by* counterfactual scenarios.

In comparison to counterfactual fairness, in organizations that attend to the concepts of equity (Adams & Freedman, 1976) and organizational justice (Cohen-Charash & Spector, 2001), managers look to reasonably accommodate the specific needs of individuals to the extent possible. Specifically, as both equity and organizational justice

theory center on the subjective fairness or justice perceptions of individuals, it is often necessary to consider unique circumstances and the tailoring of accommodations to achieve fairness and justice. For example, since equity theory (Adams & Freedman, 1976) suggests that individuals compare their inputs and outcomes to others to judge whether they are receiving an equitable share, they may advocate for accommodations that match their sense of fairness. Similarly, organizational justice involves at least three components. i.e., procedural justice (evaluation of the fairness of organizational policies, procedures, and decisions), distributive justice (the allocation of resources and rewards), (e.g., Newman et al., 2020; Viswesvaran & Ones, 2002 and interactional (the fairness of interpersonal interactions) justice (Cohen-Charash & Spector, 2001). As such, various accommodations may be necessary to ensure these forms of justice are achieved for all employees. In comparison, counterfactual fairness does not focus on accommodation; moreover, based on the current definition, it would likely be a major challenge for programmers to incorporate some of the varying considerations associated with principle of equity, for example, those involving the historical disadvantages of some minority groups, systemic discrimination, discriminatory organizational policies and decision-making, and toxic organizational culture.

Moral Agents

Implementation of counterfactual fairness involves both policy makers and programmers acting as moral agents, setting the rules for allocating resources. In comparison, policy makers and *direct managers* have the role of fostering equity and justice theory outcomes. In comparison to counterfactual fairness, the implementation of equity and justice constructs may be more nuanced, in that when employees

raise concerns about unfair treatment, they may collaborate with management to come up with a fair solution. Notably, the lack of human involvement in fully automated algorithmic HR decisions contributes to a lack of perceived justice and lowered organizational commitment among those impacted by the decisions (Newman et al., 2020).

These differences concerning moral agents have several implications: First, there is little analysis or discussion in the literature concerning the rules that govern counterfactual fairness. This is a crucial gap as these rules are foundational to ensuring fair decision-making by machine learning models. Second, in comparison to the role of moral agents associated with counterfactual fairness, equity theory, and justice theory suggest that an *overall reexamination* of the procedures and methods used to ensure fairness and justice in decision-making may be required. For example, there may be a need to involve various stakeholders (e.g., employees and customers) more deeply than is typically the case. Third, in some instances, the accommodations required to ensure fairness and justice in decision-making may result in the need to changes to policies, procedures, and/or systems. Table 4 summarizes the side-by-side comparisons involving counterfactual fairness, equity theory, and justice theory.

Insert Table 4 about here.

CAUSAL MODELING AND OBJECTIVITY

While there are a wide range of potential roadblocks (Vassilopoulou et al., 2023), the relative objectivity of algorithms designed by programmers has *the potential* to

facilitate less biased decision-making because unlike typical human decision making, algorithms are designed to follow predetermined rules that do not allow personal biases or subjective judgment. Moreover, even if algorithms are trained based on biased data, counterfactual fairness can be used to uncover discrimination. For example, John-Mathews, Cardon, and Balagué (2022) apply Bontalski and Thevenot's (1983) framework to describe the organizational decision-making processes managed by algorithms as "reality tests" in which the data input into an algorithm constitutes *reality*, while the choices or scores transformed from the initial situations are *algorithmic tests*. To ensure such algorithmic tests based on input data do not ignore the "world," that is, causal relationships that underlie potential unfairness, John-Mathews et al. (2022) contend that "fairness metrics" need to be incorporated to ensure that the tests reflect both reality and the world.

In line with John-Mathews et al. (2022), I recommend that before applying fairness metrics to evaluate the algorithmic tests, we also draw from the three levels of causal inference from Pearl and Mackenzie's *The Book of Why* (2018) to improve how algorithmic tests map both the world and the reality. Indeed, causal inference is foundational to counterfactual concept as applied in computer science. In their book, Pearl and Mackenzie (2018) use a causal ladder analogy to articulate three levels of causality in algorithms. The first level is association, i.e., a correlation between two variables, that may or may not be causal; that is, if we manipulate one of the correlated variables, the other would not necessarily be influenced. The second level in the ladder is intervention, which is satisfied if it were shown that manipulating one variable influences the other. The third level, which typically does not get considered, is counterfactual (Pearl & Mackenzie, 2018), which refers to the idea that if variable

A *is not* manipulated, then variable B would not be influenced. This corresponds to the "but for" legal concept (Spellman & Kincannon, 2001) and is the most stringent test of a causal relationship. Since most HR algorithms can be placed, in terms of causality, between levels one and two, the analysis that follows focuses mostly on level two. I contend that attending to all three levels of the causal ladder would allow us to design algorithms that not only recognize unfairness in data but also unfairness in the algorithmic pathways that are typically embedded in HR systems.

Level one: Association

A statistical association between sensitive attributes and valued outcomes is the first step of the recognizing unfairness. Grounded in the ladder of causality (Pearl & Mackenzie, 2018), in the case of sensitive attributes especially, it is crucial to remember level one relationships are not necessarily causal, but that further analyses must be performed as a starting point in the evaluation of potential unfairness. This step is crucial, since as Pearl and Mackenzie (2018) among others point out, most of the rapid advances in machine learning systems, including self-driving cars, speech recognition systems, and even deep-learning algorithms operate almost entirely at the level one associational mode. At this level, the machine learning systems are capable of identifying patterns and correlations within data, but they do not understand the underlying causal relationships between the variables. This means that these systems can make accurate predictions based on the available data, but they may not be able to explain how or why these predictions were made, hence limiting the ability to have a deep understanding of the complex systems.

Level Two: Intervention

Many would contend that algorithms based purely on associations are not sufficient to inform HR decision making. The second step of the ladder, intervention or manipulating the independent variable to demonstrate an effect on the outcome variable, can be facilitated using causal modeling to assist in examining potential unfairness; as suggested by statisticians (Freedman, 2009): the purpose of the analysis is to understand the impact of interventions. This level of causation is comparable to the causal inference drawn from quasi-experiments that lack a strict control group, which deals with questions such as "What happens if I change a certain variable?" or "What would have happened if a certain intervention had been applied?" Importantly, most of the data used in HR analytical software are collected under naturally occurring conditions, which do not meet the quasi-experimental standards (Cheng & Hackett, 2021) Even so, as Cheng & Hackett (2021) summarized, two of the six methods for inferring causality in non-experimental settings (cf. Antonakis et al., 2010) have potential applicability in the HR algorithmic context. These are statistical adjustments and quasi-experimental designs.

Statistical adjustment or measuring and controlling for all possible causes of y, is the simplest way to help ensure causality inferences (cf. Angrist & Krueger, 1999). Even so, while it is relatively easy to control for some common variables such as employee demographics, it is virtually impossible to rule out all the unknown or unmeasured factors that may cause variance in y. For instance, childhood experience may contribute to emotional regulation and personality, but it is not viable for employers to model these variables for selection or assessment purposes.

Quasi-experimental designs offer another option for improving the quality of causal inferences. Quasi-experiments can be seen as a way to perform interventions

in observational studies, and thus are closely related to the second level of causality defined by Pearl's framework. They are an improvement over studies based purely on level one associations but lack random assignment to the treatment and control groups that characterize the gold standard of traditional experimental designs (Antonakis et al., 2010). To help address this shortcoming, Antonakis et al. (2010) suggest that when certain assumptions are met, causality could still be established using alternative statistical techniques, such as simultaneous-equation models, regression discontinuity, difference-in-differences models, and Heckman selection models. These techniques allow researchers to control for potential confounding variables that can arise in non-randomized studies, and to make inferences about causality. Therefore by using quasi-experiments, researchers can make causal claims about the effects of certain interventions, which can inform the development of algorithms and policies in management and HRM.

Level Three: Counterfactual

The third step in the use of causal modeling for unfairness recognition is based on the counterfactual fairness concept. At the counterfactual level, causal inference can be used to: (1) assess whether intentional and direct discrimination against a protected class, referred to as disparate treatment under employment laws (Walsh, 2023), is legitimate or illegitimate; (2) identify corrective steps that can be taken in the face of illegitimate disparate treatment; and (3) identify the accountable parties if accommodations concerning the disparate treatment are not undertaken.

Causal modeling can facilitate the identification of legitimate disparate treatment, that is driven, for example by forces associated with a Bona Fide Occupational Qualification (BFOQ) (Walsh, 2023), which, under U.S. law refers to a standard or

criterion that permits employers to make discriminations based an otherwise prohibited classification, because it is reasonably required to perform the job. For instance, it may be that a church employee must be a member of the denomination to fulfill the duties of the job. Kilbertus et al. (2017) introduce a similar concept, a "resolving variable" to describe a legitimate causal mechanism that resembles BFOQ in algorithms. A resolving variable must: (1) fully mediate the existing relationship between the sensitive attribute and the outcome variable, thus tying it to the ability to do the job; and (2) generally be accepted as a legitimate requirement[3] of the job. In other words, even if there is a correlation between a sensitive attribute and an outcome variable (i.e. disparate treatment), a variable that fully mediates the relationship and is a job requirement may be considered a BFOQ and thus the discrimination may be considered legal.

The data required to examine the legitimacy of a proposed BFOQ and/or a resolving variable may be less-than-ideal to meet the requirement. That is, even though the underlying aim of counterfactual fairness, from a theoretical perspective, is to establish a standard of fairness that will be accepted by all involved, Herington (2020) warns that the assumptions underlying a fair algorithm may not be sustainable due to matters such as the role of historical injustice (i.e. assuming that the association between a protected class and an outcome variable is legitimate despite historical structural disadvantages), unmodelled injustice (e.g. race *is* what causes black people are arrested by an explicitly racist police officer), and rectified injustice (i.e. assuming that the affirmative action programs that are set up to rectify historical injustice are discriminatory against the dominant groups) from an unfair world. Relatedly, Kilbertus et al. (2017) notes that it can be challenging to make a case justifying disparate

[3]Notably, not all cases of full mediation are likely to be regarded as legitimate.

treatment because readily available datasets may not be sufficient to demonstrate full mediation of the associations between sensitive attributes and outcome variables, leaving systemic discrimination as a possible mechanism. Thus, if disparate treatment is unresolved, it may be necessary to review and redesign HR policies, practices, and decision-making processes to, for example, break the norms tied to the dominant culture.

When variables resolve the existing associations but do not fit the legal definition of a BFOQ, causal analyses may be helpful in identifying the actual cause(s) of discrimination so that proper intervention and accountability can be established. For instance, if in a dataset, race involving door-to-door salesmen is significantly associated with performance outcomes, and if the variable "diversity of neighbourhood" resolves the association, the lower performance for black, indigenous and people of colour (BIPOC) salesmen may reflect racism among neighbourhood residents. Thus, a workable intervention may be to match BIPOC salespeople with relatively more diverse neighbourhoods. If after the intervention the association is unresolved or only partially resolved neighbourhoods, it may be concluded that the source of discrimination is customers or other employees despite the employer acting in good faith, such that subsequent HR decisions would less likely be regarded as discriminatory. Notably, the moral responsibility of the employer, which may differ from the legal obligation, will be considered later in the paper.

In practice, a theoretical causal model used to establish counterfactual fairness has several limitations. First, it is extremely challenging to build a perfect model, i.e., one that involves no error variance. Second, causal models represent truth, only to the extent that the mechanisms associated with bias are not reflected in the

originating data set. Third, the models in and of themselves do not address issues of morality. The precision of the model, for example, may be compromised where fairness considerations are incorporated because fairness itself implies preferential treatment to minority groups, who are in fact likely to perform less well due to historical disadvantage and systemic barriers. Such trade-offs are less statistical issues than moral ones. Thus, both legality and morality are crucial components of fairness determinations.

LEGAL REQUIREMENTS AND THE PRINCIPLE OF EQUITY

While some computer scientists may suggest that a model that truthfully represents causation should be considered fair algorithmically, in management practices, an algorithm that is causal does not ensure fairness given the legal context most employers face (Walsh, 2023). I contend that causal algorithms are appropriately viewed as mirrors of management systems that may or may not involve unfairness. Relatedly, the aim of causal algorithms should be to depict not only *how* fairness is distributed among various minority groups, but also *why* and via *what mechanisms* it influences individuals within the system. Accordingly, in this section, I provide a step-by-step guideline for evaluating the legal compliance of a management algorithm, that reflects the underlying values and priorities that inform the law in Canada (and to a large extent, the United States, the United Kingdom, Australia, New Zealand, and many European countries). In recognition of the fact that many jurisdictional differences and legal technicalities exist, the discussion is purposely generic, but still useful in terms of articulating the major considerations. Finally, the steps outlined below recognize the parallels between the three stages of causality and HR legal compliance

issues concerning the use of HR algorithms.

Step 1: Association → Assessing Direct Discrimination and Indirect Discrimination

As referred to earlier, one legally recognized from of discrimination is disparate treatment where an employer intentionally treats members of a protected class differently than others (Walsh, 2023). Decisions driven solely by membership in a protected classes may be evident in algorithms when they are linked to negative management decisions. In comparison, disparate impact, another legally recognized form of discrimination is *unintentional*, and occurs when HR practices act on a negative prediction of the model which is correlated with membership in a protected class (Walsh, 2023). Many cases of disparate impact involve proxy discrimination, where the model uses information related to a sensitive attribute, that in turn, is related to a valued outcome, such as performance. Proxy discrimination may be tied to HR practices (e.g., implicit biases among managerial/HR staff, etc.), systemic barriers (e.g., misogynous organizational culture; racially discriminatory traditions), and/or historical disadvantages (e.g., minority groups that are under-represented, under educated, or under socially-supported). In any case, the underlying reasons for a variable having a disproportionate effect on a protected class need to be analyzed and understood; if, in practice, decisions are based on the variable in question, allegations of illegal discrimination could follow.

Step 2: Intervention → Finding Resolving Variables

Resolving variables can be mathematically fair, yet importantly, such fairness may not meet legal compliance requirements. Employers in many jurisdictions may, for example, be required to offer reasonable accommodations, or to act affirmatively

by setting quotas (Walsh, 2023) to address historical disadvantages and systemic barriers. Thus, in comparison to the counterfactual fairness approach typically taken in computer science, it is crucial to note that the many underlying legal concepts were *not established* with the idea that everyone be treated in the same way. Instead, the aspiration is that some individual unfairness to majority groups in the present will help ensure social fairness in the long-term, similar to the policies and actions of affirmative action.

When evaluating algorithmic fairness, it is important to take into account the requirement for reasonable accommodation. Such consideration goes *beyond* the scope of just counterfactual fairness yet is often overlooked. Generally, reasonable accommodation refers to any change to the hiring process, job content, or the work environment that allows an individual who is otherwise qualified to perform the essential functions of the target job. The intent is that accommodations be reasonable, i.e., that they do not create an undue hardship for the employer (Walsh, 2023). Accordingly, there are at least two implications for HR algorithms. First, the algorithm itself may need to be modified to accommodate certain individuals. For instance, if a woman has taken a year away from outside work due to pregnancy, leaving a gap in her career path, her performance may need to be adjusted to accommodate her legal right to be treated equally. Second, any existing data used as a basis for algorithmic determinations in the future must be adjusted to reflect the accommodation, especially regarding the assessment of performance. In all, both the algorithm and the data must be adjusted to help ensure legal compliance.

Step 3: Evaluating BFOQ Variables

As discussed earlier, discrimination is legally permissible when the resolving variable is a BFOQ (Walsh, 2023). Importantly, BFOQ exceptions are defined very narrowly, typically requiring instead that employers make reasonable accommodations, which resolving variables do not consider. Thus, again, a feature needs to be built into the algorithm to reflect the BFOQ legal context. Depending on the jurisdiction, it must adjust the performance evaluations of individuals in various relevant scenarios, including, pregnancy, disability, grieving, and ageing. Second, accommodation guidelines must be established and be made transparent to employees; they should be open to discussion and be included in collective agreements as applicable. Third, when reasonable accommodations cannot offset the barriers involved, the legal responsibility of the employer is to improve the relevant working conditions in a gradual manner. Fourth, an assessment and auditing system should be in place to evaluate the fairness of these algorithms in these regards.

ALGORITHMIC FAIRNESS AS A DYNAMIC CONCEPT

In relation to many of the issues raised above, some employers regard legal compliance as the minimum threshold, and strive for a more progressive management policy. These strivings take us beyond the legal realm and contend that the microlevel conceptualization that underlies much of the research concerning equity (Adams & Freedman, 1976) and organizational justice (Cohen-Charash & Spector, 2001) theory could benefit from the addition of a more macro level perspective (e.g., Cook & Hegtvedt, 1983) that encompasses elements of both the sociomateriality (Vassilopoulou et al., 2023) and the ensemble (Kim et al., 2021) perspectives of HR management technology, but goes beyond them. In this sense, the Cook and Hegtvedt (1983) perspective

aligns directly with Vassilopoulou et al. (2023) who suggest that HR professionals working with algorithms need to be open to deepening their perspective concerning the concept of fairness.

Cook and Hegtvedt (1983) contend that justice is not merely a property of individual decisions or outcomes, but that it is also shaped by the larger social, economic, and political system in which it is embedded. Specifically, they articulate a macro justice perspective, which entails the examination of the structural and contextual factors that influence the distribution of resources and opportunities within a society. This includes the distribution of power and resources as well as social norms and values. Use of their framework can provide valuable insights concerning how justice is experienced and perceived by different groups and can help identify ways to promote more equitable and just outcomes. In all, they caution against oversimplifying the concept of justice, which they view as a multifaceted, complex, and embedded by social context. Below, I apply their framework to identify external factors that potentially impact HR algorithmic models.

Changes in demographics and other characteristics of the population: Macro-level phenomena, such as the influx of immigrant workers, increased female participation, and increases in disability rate post-pandemic, can result in underrepresentation in the data used to train or test HR algorithmic models. If the dataset used to train or test HR algorithmic models does not include enough data points from these groups, then the model may not accurately represent the characteristics and behaviours of these groups. Any of these changes in demographics and other characteristics of the population alone or in combination could easily impact model accuracy. If the data becomes more diverse or includes new features, the model may

need to be retrained or adjusted to take these changes into account.

Changes in the model's assumptions or assumptions about the data: External changes likely affect the assumptions algorithmic models use concerning the data and/or relationships. For instance, a machine learning model trained pre-COVID-19 to predict the likelihood of employee turnover, would miss the advent and phase-out government-sponsored financial aid programs, exacerbated child-care challenges, and changes in preferred working environment, among others. Model re-training based on new data and/or other adjustments (e.g., modifying the model's architecture, fine-tuning its hyperparameters, or using different pre-processing or feature engineering techniques) would be required to, for example, reflect changes in the turnover intention distribution. Reworking the entire model may be required.

Changes in the goals or objectives of the model: Changes in the external environment can impact the goals or objectives of a model, which can in turn affect its performance. For example, a model designed to predict retention of only the highest performing employees may become less effective due to changes in the labour market or a shift in strategic goals toward accelerated growth. In such cases, the model may need to reweigh its goals or objectives, potentially sacrificing some accuracy in predicting employee performance in order to prioritize retention.

Changes in the model's environment or context: Changes to a model may be necessitated, for example, by a shift in geographic location or the use of different hardware. For instance, if a model was developed to predict employee performance based on their qualifications and past on-site performance, adjustments may be required should the employer shift to a work from home approach and interactions over virtual environment. As a result, the data distribution for employee performance

shifts, and the model assumptions about the data may no longer hold. Retraining of the model, modifying its architecture, fine-tuning the parameters, or using different processing or feature engineering techniques may be needed to adjust to the new performance profile.

MANAGING DYNAMIC CHANGE IN JUSTICE PERCEPTIONS

Notably, the examples of external change used above are simplifications for many employers who typically face many types of change *simultaneously*. Moreover, the examples do not capture the *rate of change*, which can be especially challenging in building fairness into machine learning systems. From a micro-justice perspective, researchers have also pointed out how the perceptions of fairness can change over time in response to various factors. Jones and Skarlicki (2013), for example, proposed a dynamic model of organizational justice, which suggests that fairness perceptions are influenced by experience, current events, as well as future expectations, which all interact on an ongoing basis in complex ways. As such, they contend that organizations must be ready to address the dynamic nature of fairness perceptions as they significantly impact employee attitudes and behaviour (Cohen-Charash & Spector, 2001; Vassilopoulou et al., 2023). Jones and Skarlicki (2013) suggest that employers can improve fairness perceptions by being transparent and consistent in their decision-making processes, providing opportunities for employees to voice, and by addressing issues and/or conflicts as they arise. Below I illustrate how their recommendations have implications for HR algorithmic fairness.

Being transparent

As discussed earlier, biases in HR algorithms can arise from the data used in the training stage, which may not represent the full diversity of examples and inputs that the model is expected to handle. For instance, if the data used for training is heavily skewed towards a particular demographic, the model may perform well with regard that group, but poorly on others. As implied earlier, the use of diverse training data can help to reduce model bias by exposing it to a wide range of examples and inputs. In turn, this is likely to lead to more accurate generalizations and the avoidance of over generalizations. Being transparent about the data used includes the size, quality, and diversity of the data set, as well as any preprocessing or filtering that may have been applied. Even so, Newman, Fast, and Harmon (2020) found that increasing transparency concerning the factors used in algorithms failed to ease perceptions of unfairness across a wide range of HR decisions. Instead, people preferred meaningful human involvement over purely automated processes, due to their belief that algorithms cannot apply to qualitative factors. In other words, people still prefer human involvement in HR decisions even if the algorithm is transparent and the factors used are clear.

Being consistent

Being consistent means ensuring that algorithmic management models remain fair and unbiased over time. As reflected throughout this paper, regular evaluation and monitoring of models is required to detect changes in performance that involve issues of fairness. Some key elements of this strategy include conducting bias audits,

evaluating the model performance for specific subgroups, and monitoring for changes in the data that may lead to unintended bias. For example, a bias audit entails the systematic review of predictions and decisions of the model to identify systematic problems. Subgroup performance evaluation involves examining the accuracy, fairness, and other performance metrics of the model when applied to the targeted group as a function of different types of inputs. Finally, monitoring for changes in the data entails regularly checking the training data and other input data to ensure that they remain representative of the population and are not subject to bias or changes that produce unintended consequences.

Incorporate voice

Incorporating the voice of employees, which is crucial to perceptions of procedural justice (e.g., Lind et al., 1990) helps ensure that algorithmic management systems are fair and just for these direct stakeholders. The design of algorithmic HR systems should take into consideration the perspectives of all stakeholders, including employees who will be impacted by its predictions and decisions. One option for incorporating employee voice is through direct engagement and consultation. Employers can solicit input and feedback from employees using, for example, focus groups, and surveys conducted by third parties. Third-party audits to ensure an unbiased review of the algorithm can yield an independent perspective concerning system fairness and performance. Areas in need of improvement from the perspective of various stakeholders that were not evident to the employer can potentially be identified. The algorithms can thus be retrained to incorporate various feedback, their parameters adjusted to be more accurate, and new features added to afford new functionality.

Address conflicts

Encouraging stakeholder engagement and collaboration can be helpful in addressing issues or conflicts that arise during the implementation of an algorithmic HR system. By engaging employees, customers, and other stakeholders one can better understand the full range of needs and perspectives. Relatedly, creating a fair, transparent appeals process, accessible to all stakeholders, can help address any issues arising as a result of decisions made by the algorithmic system.

In sum, to foster a dynamic perception of algorithmic fairness, employers should use diverse training data, conduct bias audits, including the evaluation of performance by specific subgroups. Moreover, employees need to be transparent in their representation of training data, be consistent in evaluating and monitoring the model, incorporate the voice of key stakeholders, engage their collaboration, and incorporate their feedback, including that obtained from a well-designed transparent appeals process.

BUILDING HR ALGORITHMIC FAIRNESS: A PROCESS MODEL

Definition of HR Algorithmic Fairness

In consolidating the various perspectives discussed in this paper, I offer the following definition of algorithmic fairness:

Algorithmic fairness in management contexts refers to a set of processes ensuring the proactive removal of inappropriate social barriers through the identification of unfairness based on causal inference, the accommodating of personal needs grounded in legal compliance, system implementation based on the principle of equity, and regular system reevaluation that incorporates the feedback of key stakeholders.

Relatedly, Figure 3 depicts a five-step process model that incorporates the major considerations concerning the evaluation of HR algorithmic fairness.

Insert Figure 3 about here.

A Five Step Process Model for Building Algorithmic Fairness

First, the use of causal modelling is proposed to uncover existing biases, including the analysis of the processes that generated it. The first step in addressing existing employer biases involves the identification of the variables that comprise the causal model, such as demographic

characteristics and those tied to organizational policies or practices. Second, a careful consideration of all other variables that may be related to the outcome of interest should be undertaken (e.g., age, gender, education level) that could potentially confound the relationships involving the variables of interest. Third, data relating to potentially confounding variables should be obtained so that their impact on variables of interest can be assessed and controlled, using regression analysis and/or causal modeling. Finally, the findings associated with the initial analysis should be carefully evaluated in relation to, for example, the potential influence of resolving

variables. Additional analyses or adjustments can be made as required, to control for their influence.

Second, a multi-step assessment concerning legal requirements can be undertaken. This begins with research concerning laws that apply to jurisdictions, including the extent to which the employer is subject to additional requirements, for example, as a government contractor (Walsh, 2023). Given the range and depth associated with this step, consulting with legal counsel (either in-house or external) is likely required. Next, guidelines concerning accommodations need to be developed that are clear and transparent to all employees, via regular communication, employee handbooks, and collective agreements as applicable. These should address, for example, those in need of accommodations should first contact the organization, and the extent to which the employer requires documentation of the need (Casey, 2023). The guidelines once established, should be integrated into the algorithm to ensure that the appropriate evaluations and adjustments are made. This may involve programming the algorithm to recognize and respond to specific circumstances as required by law to ensure that the relevant guidelines are consistently and objectively followed.

The third step in building algorithmic fairness is to plan for accommodations based on both the legal requirements and the moral values of the employer. One of the key objectives of AI researchers is to ensure the alignment of AI systems with human values. However, in a world characterized by diverse backgrounds, resources, and beliefs, the selection of principles to govern AI becomes a challenging task. To address such complexity, researchers from Google DeepMind proposed to operationalize the application of John Rawls' concept of the Veil of Ignorance (Weidinger et al., 2023). Based on the thought experiment of the Veil of Ignorance (Rawls, 1999, p. 118),

participants are allowed to choose principles without any knowledge of their own position in society. The study findings indicate that participants when shielded from awareness about their own circumstances, exhibited a higher frequency of selecting and endorsing principles that prioritize the well-being of the most disadvantaged individuals. This suggests that the Veil of Ignorance holds promise as a viable process for the selection of principles to govern real-world applications of AI, above and beyond the legal requirements.

The fourth step concerns implementing the principle of equity in the algorithm and testing it to ensure it effectively promotes fairness. This component of the process should begin by communicating the principle of equity and its goal to all employees, who should be involved in its development and implementation. Relevant policies and procedures to foster its implementation should be subject to employee input and feedback, via, for example, focus groups and surveys. Channels of communication should be established, for example, through dedicated email, to allow for ongoing employee feedback. Aided perhaps by specific goal setting, progress relating to the principle of equity should be monitored by collecting data concerning the representation and treatment of various groups of employees.

As a final step in promoting fairness (see Figure 3), the algorithm should be regularly reviewed and updated in response to changes in the data, objectives, and the external environment. This may involve reconsideration of the strategic resources, capabilities, and goals of the organization in tandem with the needs of employees, and the repetition of some of the earlier steps in the process. In any case, it is crucial to hold individuals accountable for their actions in relation to the algorithm. Thus, in both this and the earlier steps, supporting organizational structures and governance

are required to help ensure effectiveness (cf. Vassilopoulou et al., 2023).

FURTHER DISCUSSION AND CONCLUSION

This paper offers a conceptual framework to build understanding concerning a highly controversial topic, the concept of algorithmic fairness in HR management. In doing so, I contrast the computer science and management perspectives concerning the construct to broaden its definition by incorporating causal, legal, and moral components. I also present a five-step model intended to help ensure HR algorithmic fairness.

The associated traceability of HR algorithms in management provides an excellent opportunity for both researchers and practitioners to review and identify potential systemic discrimination in management systems. While from a strictly legal perspective there may be sufficient language to qualitatively describe unfair discrimination, the crux of governance is in quantifying it in a manner that allows algorithms to assist in its effective management and eradication.

Chapter 4 TOWARDS ASSIGNING MORAL ACCOUNTABILITY IN THE DESIGN AND USE OF HRM ALGORITHMS: ETHICAL AFFORDANCES AND MORAL DELEGATION

ABSTRACT

I identify two competing views, grounded largely in the computer science and management disciplines respectively, concerning who should bear much of the moral responsibility associated with the design and use of HRM algorithms. Using an artifact-centric approach, I apply affordances theory to develop the construct of *ethical affordances* that describe various properties of algorithms which significantly affect the distribution of moral responsibility among the designers, users, and regulators involved. I present a thought experiment to illustrate how, relative to the existing literature, the application of my ethical affordances framework provides a much clearer basis for analyzing and assigning moral responsibility (i.e., culpability, fair communication, public accountability, and active responsibility) among the major stakeholders. Finally, I outline some theoretical and practical implications associated with the framework, including some areas in need of future research. This includes a call for the establishment of a comprehensive legal framework and certification system to help ensure accountability in the design and use of HRM algorithms.

Keywords: Algorithm; ethical affordances; accountability; transparency; artificial intelligence

INTRODUCTION

Algorithmic management has rapidly grown to play an influential role in managing workforces in large organizations (Faraj et al., 2018; Gal et al., 2020; Kellogg et al., 2020; Parent-Rocheleau & Parker, 2022; Robert et al., 2020). Computer-based algorithms are used to evaluate credit scores, the risks associated with providing insurance, and in recent years, HRM decisions including hiring, compensation, employee engagement, and the management of turnover (Dignan, 2018; Meister, 2017; Walker, 2012; Walter, 2018). Early examples of HRM algorithms include Deloitte's "grammatical errors on resumes" (Bersin, 2013b), Xerox's "creative personality type" (Walker, 2012), and IBM's "working overtime without rewards or promotion" (Alexander, 2016). Much more complex algorithmic management systems, like Uber, Upwork, and Deliveroo, are now able to coordinate and assign tasks to many workers, provide evaluation and feedback, and maximize the profit of organizations across different industries (Bucher et al., 2021; Kellogg et al., 2020; Meijerink et al., 2021; Schildt, 2017; Tambe et al., 2019). The scholarship concerning their use, including related analytical models, reflects a cautious, conservative outlook (Angrave et al., 2016; Cheng et al., 2018; Marler & Boudreau, 2017; Rasmussen & Ulrich, 2015). While some of those systems have incorporated human input and interventions, many have the capacity to make fully automated decisions that are criticized as "human-out-of-the-loop systems" (Danaher, 2016, p. 246). Some management researchers have questioned the ethical and practical implications of these algorithms; Angrave et al. (2016), for example, concluded there is little evidence to support the strategic value of these tools. Relatedly, Cheng (2017) warned that these analytical models typically lack a strong

basis for making causal inferences and that the resulting recommendations could be discriminatory against protected minority groups. Furthermore, algorithms are criticized as discriminatory through replicating the existing biases in the data they are trained on, producing inaccurate or misleading content, and creating ethical dilemmas based on opaque and unaccountable decisions (Neale, 2023). Even so, contrary to early perspectives that the zealous embrace of algorithmic models in management practice would turn out to be a fad (e.g., Angrave et al. (2016); Rasmussen & Ulrich (2015), projections for the size of the workforce analytics market exceed 3.5 billion USD annually (ZionMarketResearch, 2021).

Much of the initial research on algorithms and management focuses on examining these algorithmic systems without unpacking what is inside the black box (Basukie et al., 2020; Evans & Kitchin, 2018; Galière, 2020; Malik et al., 2022; Meijerink & Bondarouk, 2023). The conceptualization of algorithms in such research is centered on: (a) *the context of the algorithms*, such as the regimes of control that govern and discipline the big data systems (Evans & Kitchin, 2018) or the intervention of managers and platform workers on the algorithmic management platforms (Galière, 2020); (b) *the discrete capabilities of the artifact*, such as capabilities of AI-enabled bots in carrying out a range of HRM tasks (Malik et al., 2022) or the duality of HRM algorithms to both restrain and enable worker autonomy and value (Meijerink & Bondarouk, 2023); or (c) *the dependent variable such as unintended negative effects including algorithm bias* (Basukie et al., 2020). While valuable, research of this nature often treats the notions of this technology as fixed and unified, without providing detailed descriptions or theoretical explanations for the algorithm. Orlikowski & Iacono (2001) argue that researchers have assigned greater theoretical significance

to the contextual factors within which the technology operates, yet the technology itself tends to fade from view, being taken for granted or presumed unproblematic once constructed. The term "artifact" is widely used in the field of information systems to refer to such technology, as the collection of material and cultural elements assembled in a socially recognizable manner, which usually takes the form of hardware and/or software (Orlikowski & Iacono, 2001). Many researchers have raised concerns over algorithms from an artifact-centric perspective, including transparency issues, labeling algorithmic recommendations as a "black box" that lacks explanation (Boulton, 2017; Johnson & Ruane, 2017; Pasquale, 2015; Wilson & Daugherty, 2017). Relatedly, Johnson and Ruane (2017) note that bias hidden in an algorithm cannot easily be detected as users "cannot simply read the code to analyze what is happening." Boulton (2017) warns against blind trust in algorithms generally, including in management contexts, as he argues that the underlying modeling must be accurate when the decisions impact peoples' lives. Martin (2019) argues that software designers bear the primary responsibility for the moral implications associated with algorithmic recommendations, as programmers can redelegate (shift) the associated moral obligations to various stakeholders (e.g., via the use of default settings). Martin's view reflects the concerns of many that important information embedded in the software has bearing on its ethical application but is often hidden from users.

While many computer scientists also share the view that accountability should be pinned on humans instead of the artifacts, others argue that the owners and users of the artifacts should be held solely responsible for their ethical use (Bryson, 2018; Bryson & Kime, 2011; Bryson & Theodorou, 2019). To lend clarity to the debate, I use an artifact-centric approach to analyze the properties of the HRM algorithms and

examine the ethical implications associated with various decision making scenarios. First, I elaborate on the two competing views, grounded largely in the computer science and management disciplines respectively, concerning who should bear the moral responsibility[1] associated with the design and use of HRM algorithms. Second, based on affordances theory (Gaver, 1991), I develop the construct of *ethical affordances* to describe properties of algorithms that vary in terms of the intentions of the designers and the perceptions of users; these in turn have ethical impacts for the individuals being evaluated and on the assignment of ethical responsibility for the outcomes. Third, I apply the thought experiment method (Gendler, 1998, 2011; Kornberger & Mantere, 2020) to develop a hypothetical case involving the use of an HRM algorithm, in which four forms of moral responsibility (i.e., culpability, fair communication, public accountability, and active responsibility; Santoni de Sio & Mecacci, 2021) are considered. Fourth, I show how various types of ethical affordances commonly designed into HRM algorithms (i.e., perceptible to users and real; perceptible to users, but false or misleading in their effects; and real but hidden to users) result in significantly different distributions of moral responsibility for designers, users, and regulators (i.e., the primary moral agents involved). Importantly, relative to the existing literature, the ethical affordances framework I present provides a much clearer basis for analyzing and assigning moral responsibility to the major stakeholders. Finally, I outline some theoretical and practical implications associated with the ethical affordances framework, including some areas in need of future research.

[1]In this essay, the terms "morality" and "ethics" are primarily used interchangeably, following the headword "Ethics" of the Encyclopaedia Britannica (Singer, 1985).

Nevertheless, the author consciously distinguishes "moral" behaviours to highlight their personal and normative nature, while using "ethical" characteristics of algorithms to emphasize the standards of "good and bad" established within the business community or a given social context.

THE LOCUS OF MORAL RESPONSIBILITY: TWO COMPETING VIEWS

Moral Responsibility Should Be Attributed to Users According to Robotics Scientists

The current dominant perspective among AI and robotics scientists is that the moral responsibility for any action taken by an algorithm should be attributed to its owner or operator, e.g., the business user who purchased the software, or in the case of malfunctions, to its manufacturer. This view is similar to that involving accountability for conventional artifacts (Bryson, 2018; Bryson & Kime, 2011; Bryson & Theodorou, 2019) and stems from a classic example – a firearm, where the person who owns or operates it is responsible for ensuring that it is used in a safe and ethical manner. Relatedly, the manufacturer may have accountability, but only when the firearm malfunctions and causes harm. This approach to assigning moral responsibility is grounded in some crucial assumptions: (1) that the firearm (artifact) is performing the tasks the manufacturer intended, without malfunctions (Bryson, 2018); (2) that the designers share complete information with the operator of an artifact, and (3) the operators have full information and control over how the artifact is used (Bryson & Theodorou, 2019). An analogous analysis involving software could be undertaken, potentially leaving the purchaser accountable for any negative consequences that arise from its use (Bryson, 2018).

Moral Responsibility Should be Attributed to Designers According to Management Ethics

In comparison to the prevailing view among robotics scientists, management researchers and some in AI ethics argue that the developers of HRM algorithms are primarily responsible for the ethical considerations associated with their design. Their argument is that purchasers of HRM algorithmic software typically find themselves dealing with a pre-determined industrial product that offers little room for modification, let alone full control. For example, Martin (2019) states that there is a predetermined moral delegation associated with many technologies in performing a task, and that this delegation of moral responsibility by the designer has a direct effect on the moral behaviour of the others involved. Latour's (1992) study on doors suggests that designers face the decision to either delegate the task to human users by indicating to them exactly how a door should be opened or closed, or delegate the task to a non-human character, for example, an automatic door hinge. As such, from this perspective, *moral delegation* is essentially embedded in technology design and makes designers accountable for the moral implications on users. Even so, Johnson and Ruane (2017) argue that despite the ethical issues that may arise from algorithmic applications, the key to mitigation lies in the implementation of adequate algorithmic design processes. This view aligns with Latour's (1992) hypothetical example concerning mandatory seatbelt use before cars will start. Similarly, in amusement parks, the roller coaster's safety measures often include an automatic safeguard device that must be held in place for all passengers before the ride begins. In both cases, such safety design could potentially enable the designer to *redelegate* the moral obligation

such that it "alleviates the individual from having to take on that responsibility" (Latour, 1992, p. 152). That is, when a designer deliberately withholds a presumably immoral option from users (i.e., not fastening the seatbelt), the moral responsibility of the task can be shifted back from the users to the designers.

Importantly, to the extent that algorithms lack transparency for users, the designers have greater accountability (Martin, 2019); specifically, when algorithms appear opaque and inscrutable, it is difficult for users to take on moral responsibility for their actions. Thus, designers are left with greater responsibility to ensure their creations account for ethical considerations. This includes ensuring that users understand how the algorithm works, how the data are used, and the potential implications of the outputs. The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system developed by Northpointe is an example of an algorithm that lacks transparency. COMPAS is used by U.S. courts to assess the likelihood of a defendant reoffending. It has the potential for discriminatory recommendations in that those of Afro-Caribbean descent are consistently ranked as high-risk. Here, designers bear the responsibility for ensuring that their algorithm is designed in a transparent and ethical manner that avoids perpetuating biases and discriminatory practices. As will be discussed, the designer also likely has the capability to design a system that redelegates at least some of the ethical responsibility back to its users. In all, Martin (2019) puts the onus on the developer of the algorithm to take responsibility for both its overall ethical implications and on redelegating moral responsibilities in decision-making.

Concerns about the ability of users to gain meaningful control over algorithms have gained traction among experts from across disciplines. For example, the need

for a set of normative requirements that foster a legally, ethically, and societally acceptable form of human control of algorithms has been examined in the domains of automated driving systems (Calvert et al., 2020, 2018, 2021; Heikoop et al., 2019; Mecacci & Santoni de Sio, 2020; Santoni de Sio & van den Hoven, 2018) and medical automation (Braun et al., 2021; Ficuciello et al., 2019). These developments align with Martin (2019)'s claim that algorithms are often difficult for users to understand, making it difficult for them to assume ethical responsibility. Lacking the ability to comprehend and control algorithms, users may find themselves in a situation comparable to patients receiving prescriptions without being informed of the potential side effects. As a result, users may be excluded from culpability concerning the ethical implications associated with their recommendations.

Where Do We Go from Here: A Middle-of-the-road Approach?

Even given diverging opinions addressed above concerning who is responsible for the ethical implications associated with the design and use of algorithms, there is common ground. First, everyone acknowledges that algorithmic use typically has moral implications with the potential of unethical outcomes (Bryson & Kime, 2011; Bryson & Theodorou, 2019; Martin, 2019). Thus, both viewpoints emphasize the importance of stakeholders taking preventative measures to address these concerns. Second, all propose that accountability should be ascribed to specific actors, e.g., developers, owners, and/or operators (Bryson, 2018; Martin, 2019). Indeed, the assignment of responsibility is essential to accountability for unethical outcomes. Third, there is consensus that the responsibility for the ethical use of algorithms should not rest solely on users. Finally, all agree that transparency is a key factor in attributing

responsibility for the ethical use of algorithms (Bryson & Theodorou, 2019; Martin, 2019). This common ground suggests room for a middle-of-the-road approach to help address disagreements concerning accountability issues.

It is notable that a primary difference in perspectives lies in the relative emphasis placed on the development and execution phases respectively. Management ethics (Martin, 2019) focuses on the execution phase of the algorithm, highlighting the need for designers to redelegate responsibility and increase transparency. In comparison, AI scientists (Bryson, 2018) generally assume that the redelegation and transparency efforts have been perfectly realized as well-designed and transparent in the development phase, thereby leaving the responsibility for ethical decision making to the user during the execution phase.

While both arguments concerning the allocation of responsibility in using algorithmic artifacts are reasonably compelling, several questions remain unanswered. For example, can an algorithm that involves multiple stakeholders be truly well-designed and transparent? If an algorithm is poorly designed and/or opaque to its users, can we entirely absolve the users of accountability? Under these circumstances, how should accountability be distributed? What happens when an algorithm is used as intended, but causes unforeseen harm to a small subset of its stakeholders? How can we ensure that the operators of the artifact have the necessary information to make ethical decisions? These unresolved questions highlight the need for further discussion and research concerning the ethical implications associated with algorithmic artifacts.

The differing viewpoints among AI and management researchers highlights the existence of a responsibility gap (Matthias, 2004) in the development of intelligent systems, arising from the fact that systems possess the capability to learn from their

interactions with both agents and their environment. This makes it challenging, if not impossible, for humans to maintain complete control over the behaviour of the artifacts so as to accurately predict their actions (Santoni de Sio & Mecacci, 2021). Nevertheless, for humans to assume responsibility for these systems, it is necessary to possess knowledge and control over them.

AFFORDANCE THEORY

To foster a deeper understanding of the relationship between the ethical use of algorithms in management and the degree of transparency exhibited by them to users, I apply the theory of affordances (Gibson, 1977, 1986). The theory has received much attention in the information systems literature (Leonardi, 2013; Pozzi et al., 2014; Seidel et al., 2013; Volkoff & Strong, 2013; Yoo et al., 2012; Zammuto et al., 2007). The generic term, *affordance* was coined by Gibson (1977) as referring to all "action possibilities" that can be objectively measurable, independent of the user's ability to recognize them, but always dependent on their capabilities to be realized (Greeno, 1994; Hartson, 2003; Mcgrenere & Ho, 2000). In other words, the presence of "affordances" is not affected by whether users recognize them or not because they objectively exist. However, users' skills and capabilities determine whether they can make use of these affordances in practice. Thus, actors perceive and behave in an environment based on affordances or preconditions for an activity (Gibson, 1977, 1986). While affordances enable interactions among the actors and the environment, they do not necessarily imply that the specific activity will occur (Greeno, 1994). For instance, Microsoft Excel has the affordance to enable VBA macros[2], with or

[2]https://learn.microsoft.com/en-us/office/vba/library-reference/concepts/getting-started-with-vba-in-office

without the users' knowledge of such features and their coding ability. Even so, this affordance can only be meaningful when Excel users are aware of it and can code in the VBA programming language.

Gaver (1991) classified affordances into four quadrants based on two axes: information availability and action possibility. By distinguishing these two aspects of a design, the possibility of designing and analyzing properties of artifacts in their own terms is enabled (Gaver, 1991, p. 81).

Perceptible affordances refer to affordances that exist with available perceptual information that matches their intended use and is easy for users to operate. *False affordances* (or in Gaver's words, misinformation) refer to situations where information available to the user, either intentionally or unintentionally, suggests an affordance that does not exist; these may cause people to act without any real effects. A classic example of an intentionally designed false affordance is the inclusion of a "close door" button in elevators; they often have no real effect on the functioning of the elevator but may cause an actor to act on the misinformation by pressing the button. Actors are left to believe they influenced door closing, but the button often instructs nothing to the elevator. *Hidden affordances* refer to situations where no information is available concerning an existing affordance; examples include the Easter egg features that are part of many applications such as Google Search (e.g. typing "atari breakout" into the search bar of Google.com and clicking on "I'm feeling lucky" will allow the user to play the video game). Finally, *correct rejection* affordances refer to situations where no information is available to users and no affordance has been designed for the action (Gaver, 1991).

Researchers in artifact design have proposed additional terms to describe intentionally designed affordances that limit certain user behaviors (Lockton, 2005; Maier & Fadel, 2009; Norman, 1998); these are referred to as *intentionally limiting affordances*. Related terms include negative affordances (Maier & Fadel, 2009), anti-affordances (Norman, 1998, p.11) and disaffordances (Lockton, 2005) that have taken various forms such as greyed-out buttons on software or websites, compensation management systems, and computer configurations. For instance, the designer may leave buttons on the interface of software appear faded or inactive, to prevent certain users performing certain actions like adjusting an approved compensation plan, which may lead to discriminatory outcomes. They are departures from traditional notions of affordances as being solely supportive of user behaviours and have attracted significant interest from researchers seeking to better understand their implications for both users and for artifact design. Finally, as will be seen, *ethical affordances* are those that have implications for moral choices, and they directly influence the distribution of moral responsibility in AI contexts.

IDENTIFYING ACCOUNTABILITY VIA A THOUGHT EXPERIMENT

The Methodology of Thought Experiment

I now use thought experiment methodology to examine the nature and distribution of moral accountability tied to the design and use of HRM algorithms as a function of the type of ethical affordances in play. Thought experiments, referred to also as thought trials (Dietrich & Haider, 2015), imaginary illustrations (Lennox,

1991), or metatheorizing (Carr & Zanetti, 1999), are mental exercises that explore hypothetical scenarios and their potential outcomes (Gendler, 1998, 2011; Kornberger & Mantere, 2020). They involve imagining a situation and reasoning through the implications without actually carrying out the experiment in real life. Thought experiments are commonly used in philosophy, physics, and other sciences to explore complex concepts and theories; only recently have they been argued as suitable to organizational behaviour and related fields, including HRM, industrial and organizational psychology, entrepreneurship, and strategy (Aguinis et al., 2022) to explore scenarios that are impossible and/or ethical to test in real-life.

My thought experiment is a setting involving the design and application of an HRM algorithm for workforce management at a large corporation with multiple stakeholders. The experiment is appropriate here for several reasons in a general sense. First, it allows for the consideration of multiple perspectives and various potential consequences of a range of decisions and actions, without the ethical, legal, and practical barriers that likely would exist in reality, versus within the laboratory of the mind (Brown, 2010). Second, with the emerging nature of algorithm management, employers typically lack the motivation to share data with researchers especially when, as is likely in this context, it contains sensitive employee information that could expose the organization to legal and reputational risks. Third, there are practical limitations to collecting large scale data from employers. The sheer size and complexity of large businesses can make it challenging to gather comprehensive data, and ethical concerns may arise in obtaining and utilizing sensitive information. Finally, as with Schrödinger's cat thought experiment in quantum mechanics (Aguinis et al.,

2022; Schrödinger, 1935) [3], thought experiments can help researchers challenge existing assumptions and theories in emerging areas, such as the design and use of HRM algorithms, by exploring alternative explanations for phenomena, and by refining the theory-building process.

Specifically, I apply the decision tree framework proposed by Aguinis and colleagues (2022) to assess the feasibility of employing a thought experiment and to determine the appropriate type of thought experiment to apply. The sequential process of this decision tree is outlined as follows:

Decision Point #1: Is there a need to confirm or disconfirm theory?

There is a need to address the conflict between the two theoretical perspectives and determine which theory holds more validity or relevance under specific conditions. In particular, thought experiments provide a platform for defining and dissecting the core concepts and assumptions of each theory. This process helps in uncovering potential ambiguities or hidden premises that might contribute to the conflict. By enhancing conceptual clarity, thought experiments lay a solid foundation for robust evaluation. A thought experiment can provide a structured approach to clarify ambiguous concepts and assess the implications of conflicting theories.

Decision Point #2: Can an imagined scenario model the theory?

An imagined scenario can effectively model the clash between the two theories. By constructing scenarios that highlight the differences in predictions or outcomes

[3]Schrödinger's 1935 thought experiment demonstrates an unresolvable paradox of quantum superposition. In this experiment, a hypothetical cat inside a box is setup with a small amount of radioactive substance. If the substance decays, it sets off a Geiger counter, leading to the release of poison that could kill the cat. As the decay is a random probabilistic event, it is impossible to know without opening the box whether the cat is alive or not. Therefore, according to the principle of quantum superposition, the cat is simultaneously both alive and dead until an outside observer opens the box to check, "collapsing" it into a definite state. Schrödinger used the example to illustrate the absurdity of this existing view of quantum mechanics.

between the conflicting theories, especially when factoring diverse stakeholders and complexities in the scenario, it becomes possible to assign responsibility in more nuanced ways. The thought experiment, in this regard, can provide a tangible context for conducting such assessments.

Decision Point #3: Is the theory well-developed?

Since both theories are in conflict, it is safe to assume that they have been thoroughly developed and have gained recognition for their differing viewpoints, based on the number of citations of each paper in their recognized field. This, in turn, provides an opportunity for conducting rigorous comparative analyses through the application of thought experimentation to enable an in-depth exploration of their respective metis and limitations in a systematic, comprehensive manner.

Decision Point #4: Is the purpose theory confirmation or disconfirmation?

Addressing the conflict between these theories extends beyond a mere determination of relative superiority. It also serves as a platform for potentially revealing novel insights. This process involves not only examining which theory aligns more consistently with intricate scenarios, but also probing into the areas where the theories diverge, seeking to discover new pathways or paradigms. In essence, the thought experiment not only aims to confirm or disconfirm existing theories but also acts as a "crucible" for generating fresh perspectives and refining our understanding, ultimately enriching the associated discourse.

I argue that addressing theoretical conflict through a thought experiment aligns well with the original four decision points. In following this decision tree, a thought experiment becomes a suitable approach for navigating theory-based conflict using a

Type IV experiment. I apply the Type IV thought experiment to conduct a comparative assessment between two competing theories within the confines of a scenario. In all, a Type IV thought experiment, which involves a comparative assessment of conflicting theories, is the most suitable choice for addressing the theoretical conflict at hand.

The Setting Underlying the Thought Experiment

ACME, a large company located in Country C, has purchased TalentAnalytica, a management algorithmic solution, from Byte, a supplier of software. ACME is currently using TalentAnalytica to evaluate the performance potential of both new recruits and their existing employees. Respectively, the new recruits are evaluated based on their job application and interview performance, while existing employees are assessed based on their performance history. The software has a "fairness index" feature intended to promote ethical management practices. The index is calculated using a machine-learning algorithm based on the prior performance data of all ACME employees in their HR department. Users can view the index to compare the treatment of minority and majority groups and make any necessary fairness-related adjustments to their HR decisions. This information can be shared with the union/employees should management choose to do so. Though the HR managers do not fully understand how the fairness index is calculated, they provided Byte with prior employee data on which the machine learning model was built. ACME's use of TalentAnalytica is subject to the regulatory environment of Country C.

Analysis of Accountability

To fully comprehend the accountability issues in the hypothetical scenario, it is crucial to identify the stakeholders and their respective moral roles. Moral agents and moral patients are two concepts useful in describing entities involved, in terms of, for example, the capability to act morally and/or be the object of moral considerations (Bryson, 2018). *Moral agents* are individuals or entities that have both the autonomy and underlying knowledge to make moral decisions; as such they can be held responsible for their actions (cf. Duranti (2005); Gray & Wegner (2009); Kant (1785); Karlsson (2002)). Moral agents typically include managers, corporations, and governments. In the thought experiment, Byte the algorithm designer, ACME the algorithm user, and Country C, the HRM regulator are the primary moral agents. On the other hand, *moral patients* lack moral autonomy but are still deserving of moral consideration since they are the ones affected by the ethical decisions made by *moral agents*. Therefore, ethical decision-making requires that the well-being of these *moral patients* is taken into consideration by *moral agents*. (cf. Duranti, 2005; Gray & Wegner, 2009; Kant, 1785; Karlsson, 2002). Moral patients include ACME's employees, job applicants, and the public.

Having identified the moral agents involved, the relative distribution of moral responsibility can be examined. As I will document, the nature of the distribution varies significantly depending on the type of ethical affordances involved. As a way of fully characterizing moral responsibility in AI contexts, the Santoni de Sio and Mecacci (2021) framework will be used; it consists of four forms of responsibility

concerning AI: culpability, fair communication[4], public accountability, and active responsibility.

Culpability refers to both moral and legal responsibility for wrongdoing or harm caused by an individual or organization (Calo, 2015; Matthias, 2004; Pagallo, 2013; Sparrow, 2007). It has been highlighted, for example, in discussions relating to the use of autonomous weapon systems (Heyns & UN. Human Rights Council. Special Rapporteur on Extrajudicial, 2013; Meloni, 2016) and in the broader debate concerning the degree to which algorithms and AI are explainable to stakeholders (Doran et al., 2017; Mittelstadt, 2016; Pasquale, 2015). Importantly, culpability involves all three moral agents in our thought experiment, Byte the designer, ACME the user, and the Country C regulators. Specifically, as the software developer, Byte may be culpable if TalentAnalytica causes harm or wrongdoing. That is, although ACME provided the prior HR data to build the machine learning model, Byte has a responsibility to ensure that the algorithm is both transparent and does not unduly discriminate against legally protected groups. If the fairness index is calculated in a way that results in unfair treatment or harm to protected classes, Byte may be culpable. Moreover, ACME, purchaser of TalentAnalytica has a responsibility to use the software in an ethical manner. Specifically, since The HR managers at ACME have access to the fairness index, they may have the capability to adjust their decisions to ensure the use of TalentAnalytica does not cause harm. Alternatively, if ACME fails to use the software in a fair and ethical manner, it may be culpable for the harm caused. Finally, the regulatory arm of Country C has a responsibility to oversee the fair use of technology in its jurisdiction. Accordingly, lack of awareness

[4]The authors describe this type of responsibility as "moral accountability," but define it as "duty of human persons to explain one's reasons and actions to others". Therefore, for the sake of clarity, I use the phras "fair communication" to convey the original meaning of this terminology.

for discrimination caused by TalentAnalytica could result in culpability for Country C.

Fair communication refers to the duty of involved persons to explain their actions and reasoning to others as appropriate (McKenna, 2012; Wolf, 1993). Medical doctors for example, as moral agents, have an ethical obligation to explain the reasons for their diagnoses to patients. This type of responsibility is regarded as a fundamental aspect of being a reflective member of society (Gardner, 2007). It is especially relevant when the use of algorithmic solutions directly affects individual applicants and/or employees. As with culpability, all three moral agents in the thought experiment have fair communication responsibilities. Byte, the developer of TalentAnalytica, has a moral duty to explain to both ACME and the Country C regulators how the fairness index is calculated and the basis for the predictions the algorithm makes. Relatedly, Byte must take responsibility for adjusting the algorithmic design and/or the user interface to ensure their explanations are explicit and accessible. ACME, as the user of TalentAnalytica, has a moral duty to explain to their employees how the fairness index is used to assess performance and as an aid in HRM decisions. ACME must also take responsibility for decisions based on TalentAnalytica in relation to its dealings with their employees, job applicants and regulators. Country C has a moral duty to ensure the legal framework and the associated regulations are clear to the major stakeholders including the designers and users. Clear guidance and support must be provided to all parties should an employee file a formal complaint concerning a TalentAnalytica-based employment decision.

Public accountability refers to the duty of moral agents to explain their actions

in public forums as needed (Bovens, 2007). Relatedly, the effective design and implementation of accountability mechanisms serves not only to improve the efficacy of complex public decision-making processes, but also helps ensure the outcomes are aligned with the principles of liberal democracy (Bovens, 1998, 2007). Indeed, concerns have been raised regarding the legitimacy and desirability of AI-based automated decision-making in public administration contexts (European Commission, 2019; Hildebrandt, 2019; Noto La Diega, 2018). Difficulties associated with explaining and ensuring a full understanding of algorithmic decision-making processes (the "black box" problem); Castelvecchi (2016) are often highlighted in these debates. Moreover, challenges to clear explanations may arise not only because of *technical* black box matters, but also due to *organizational* and *legal* black boxes created or amplified by using AI in public administration (Noto La Diega, 2018).

Public accountability issues are likely to be especially challenging if the developers and/or users of HRM algorithms are publicly listed corporations, as they would then likely be subjected to a wider range of shareholders and interests, relative to closely held entities. As with the fair communication requirement, Country C regulators as public agents have a responsibility to establish and enforce a legal framework applicable to the design and use of HRM algorithms. It should ensure that designers and users of algorithmic solutions are transparent about their use, as reflected by auditability and bias free decisions. Regulators should also provide a public forum where concerns about the use of algorithmic solutions in general can be raised and investigated.

Active responsibility refers to the duty to promote and achieve societally shared goals and values (Bovens, 1998). As such, it is future-oriented and involves the

objectives, principles, and legal standards that professionals, including programmers, are expected to uphold, in addition to outcomes they should work to prevent. In the thought experiment, both Byte and ACME have active accountability responsibilities tied to the use of TalentAnalytica. Byte must ensure that their algorithmic solution promotes ethical practices, including fairness and non-discrimination; their algorithms should be designed and thoroughly tested with these considerations in mind. If the use of TalentAnalytica perpetuated unfair discrimination, Byte would be failing in their active responsibility to promote societal goals and values. ACME, as the user, also has active accountability as they must share their goals and values in a clear manner with Byte to proactively work to minimize the potential for harm in the design and use of the software. This includes ensuring that the fairness index is both legally compliant and consistent with the internal values of ACME. If ACME fails to take these steps, and the system perpetuates the bias or discrimination of the designers, they would be failing in their active responsibility to promote societal goals and values. Given the moral agents involved and the four areas in which they potentially have ethical responsibility have been identified, I now address the manner in which various types of ethical affordances are implicated in the relative distribution of moral responsibility.

THE ETHICAL AFFORDANCES OF ALGORITHMS AND MORAL DISTRIBUTION

Based on Gibson(1977; 1986)'s concept of affordance, I use *ethical affordance* to refer to the affordances that enable the moral agents associated with the design and use of HRM algorithms to fulfill their responsibilities. Ethical affordances are

objectively measurable moral action possibilities (which may or may not be recognized by the user) that depend on user capabilities for moral responsibilities to be effectively fulfilled. They can be thought of as control knobs associated with an algorithm that characterize the distribution of culpability, fair communication, public accountability, as well as the responsibility for actively promoting societal goals and values. As will be seen, these knobs are typically, but not always, perceptible to users, via a button, menu item, or toggle of the software. As noted earlier, affordances can also be (un)intentionally designed to be false or hidden (Gaver, 1991; Gibson, 1977, 1986).

Figure 4 summarizes the categorization of the Gaver (1991) affordances quadrants based on whether the information related to the affordance is available and whether it points to a moral action.

Insert Figure 4 about here.

Importantly, designers are an especially crucial moral agent as they determine how ethical affordances are presented to users. That is, the knobs associated with ethical affordances are a product of the deliberations and choices made by designers during the development phase. These involve both the use of mathematical models and user-provided data. Typically, HRM algorithms involve heuristic statistical models that efficiently use analytical resources to provide immediate recommendations for decision-making (Cheng & Hackett, 2021). These heuristics are implemented not only through major decisions such as the inclusion or exclusion of variables, the causal assumptions made, and the selection of the desired model outcomes, but also through relatively minor decisions such as data sifting (Sambasivan et al., 2021). Moreover, although the mathematical choices made during algorithm design may not appear

to have ethical consequences, ultimately they are often implicated (Passi & Sengers, 2020).

Consider a scenario in which TalentAnalytica is used to predict the performance of a job applicant based on their employment history, but there are gaps in the history. To address missing data for such applicants, the Byte designers must make a mathematical decision concerning how to replace the missing data as part of assessing applicants' potential. They may, for example, select an imputation technique to estimate the missing data based on the patterns in the available data (Enders, 2017). Importantly though, imputation techniques may differ in their impacts on diversity, equity and inclusion (DEI) and on the degree to which there is capability to counteract the implications associated with the historical oppression of marginalized groups (Woods et al., 2023). A potential issue, for example, is that applicants with missing history may have different outcomes than those who are included, resulting in biased conclusions concerning selection system effectiveness, as the outcomes of the missing data imputation may be more favourable in one group than the other. For example, some historically marginalized groups have had limited access to certain training opportunities or have faced discrimination affecting their past performance evaluations. When imputing missing data for these groups, the algorithm might generate higher performance predictions than what would have been the case if the true historical data were available. In this scenario, although the mathematical imputation technique may appear not to have ethical implications, it may yield biased predictions and unfair treatment.

Crucial to the current discussion is that the designer has *choices* to make concerning both the selection of the imputation technique *and* whether there will be a

visible menu item for the users to make a change (a perceptible ethical affordance) or, to keep the interface simple, the imputation choice will be hidden (a hidden ethical affordance). Both choices have moral consequences and result in different responsibility distributions. Accordingly, as shown in Figure 4, I contend that in relation to Gaver's (1991) affordances quadrants, that ethical affordances should also be categorized according to whether the affordance is perceptible, as well as to whether a moral action is indicated. The availability of ethical affordance and the perceptibility information associated with the affordance are both crucial to the determination of the responsibility distribution among the moral agents associated with the design and use of the algorithm. For instance, if Byte both provides ACME with a toggle to control how the missing data are calculated along with a clear explanation of the potential ethical implications associated with each type of mathematical decision, a perceptible ethical affordance results; importantly, responsibility for this component of the algorithm is then effectively redelegated from the designer Byte to the user ACME.

More broadly, as I will now illustrate, the framework in Figure 4 can be used to clearly examine the accountability of the various moral agents in both the development and execution phases associated with HRM algorithms. As will be seen, application of the framework also lends clarity to the primary differences between the management and computer science disciplines regarding the distribution of ethical responsibilities involving the design and use of these algorithms.

Perceptible and Real Ethical Affordances and Moral Distribution

As described earlier, the concept of perceptible ethical affordances refers to aspects of the user interface that explicitly present information that points to moral actions that could artifactually enable or constrain the ethical behavior of users. The focus of perceptible ethical affordances rests both on the perceptible nature of the design feature to users, and that the option is real, yielding a means to achieve ethical objectives. In the thought experiment, the fairness index designed by Byte as a component of TalentAnalytica would be characterized as a perceptible ethical affordance only if its features were both perceptible to users and if they enabled ACME to achieve their intended DEI outcomes. More broadly, perceptible ethical affordances via visible interfaces, warnings, or grey-out buttons, directly impact the nature of the distribution of culpability and active accountability associated with moral decisions. As such, it is important to explicitly examine their role in both the development and execution of HRM algorithms.

Development Phase: As addressed in the earlier imputation technique example, designers, during the HRM algorithm development phase, have the option of incorporating perceptible ethical affordances that redelegate culpability and active accountability. These affordances have the potential to align with legal and moral requirements, and support shared goals and societal values. For instance, during the Covid-19 pandemic, AI systems could incorporate interfaces that allow users to make accommodations for employees in protected classes including sex, disability, or caregiver status (Cheng et al., 2022). By including these as selectable options, ACME could adjust the outputs to allow for ethical decisions that prioritize inclusivity and

reflect the regulatory framework in Country C. Relatedly, ethical affordances could be designed to restrict ethically risky behaviour by including perceptible warnings or grey-out buttons that cannot be modified by users. For example, TalentAnalytica may display a warning message or a grey-out button to ACME staffers who attempt to deny accommodations legally required in Country C. Moreover, the design may entail that these restrictions could be overridden only if certain criteria are met, for example, if ACME moved its operations to Country D, where different accommodations were required (Kozuka, 2019). By providing these perceptible ethical affordances, TalentAnalytica could not only guide ACME users towards behaviour in compliance with legal and moral requirements, but also restrain them from unethical decisions. For both Byte and ACME, culpability would be reduced, and active responsibility promoted.

Regarding the goal of fair communication, the perceptible affordances must both support users' moral actions and provide clear, understandable instructions and explanations concerning moral objectives. For example, when TalentAnalytica recommends an employee promotion, a clear explanation for the recommendation must be included that addresses how the decision aligns with both the ethical requirements outlined by ACME and the Country C regulatory framework. Instructions via user manuals or help functions should also be provided so that the recommendation parameters can be adjusted by ACME. By providing explanations and instructions, Byte promotes ethical decision-making through TalentAnalytica and enables ACME to better understand and utilize the algorithm so that users make informed decisions and can take responsibility for their actions.

Public accountability can be fostered using perceptible ethical affordances by

helping to ensure transparency and traceability of actions taken within the algorithm. An option can be designed into TalentAnalytica that records and tracks the actions taken during the recruitment and selection process. The record would include critical information such as job requirements, applicant qualifications, and the reasons for selecting or rejecting each candidate, with the capability to identify and record potential biases in the process. This transparent traceability enables the evaluation of algorithmic functioning, fostering public accountability (Mittelstadt, 2016).

In all, the creation of perceptible ethical affordances for users during the algorithm development phase has the potential to shift some moral responsibility from the designer to the users of the software. While the incorporation of these affordances does not absolve the fundamental responsibility of the designer to provide an algorithm that is fair, unbiased, and aligned with legal requirements, it does enable users to act in ways consistent with their moral guidelines, be aligned with regulatory requirements, and live up to societal values including fairness and justice (Köchling & Wehner, 2023), to mitigate ethical concerns. Fair communication and public accountability are also promoted.

Execution Phase: Designers also have the option of incorporating perceptible ethical affordances into the execution phase of HRM algorithms. Mechanisms that identify and address biases in the algorithm during its application can be added that allow ACME to act proactively to prevent biased decision-making. The execution phase also presents an opportunity for designers to incorporate channels for user feedback to foster fair communication between users and moral patients that serve to identify areas in need of improvement. For example, if Byte provides a help desk

or ticket system for ACME to provide feedback concerning the functioning of TalentAnalytica, execution can be revised as required to limit unintended outcomes in a timely manner. The nature of the adjustments can then be shared with stakeholders including job applicants. TalentAnalytica should also be designed in the execution stage to yield the data required by ACME to comply with Country C legal requirements. Relatedly, Country C has a moral responsibility to design and implement regulatory apparatus to effectively monitor ACME and its use of TalentAnalytica for legal compliance. Establishing mechanisms for external auditing and evaluation of algorithmic outcomes provides an additional layer of public accountability.

Finally, the range of issues raised above suggests that the distribution of moral responsibility among the designers, users, and regulators during the execution stage is notably murky due to the complexity of algorithmic systems and the potential for unintended consequences. For example, users may lack the technical expertise required to identify and correct algorithmic biases in the algorithm without ongoing support from the designers. Designers may also need to be involved if the incorporation of user feedback requires substantial structural changes to the algorithm. Finally, regulators may not be in a strong position to offer legal guidance should they lack the institutional expertise to evaluate algorithmic systems. In all, given the current state of technological development and societal structure, accountability at the application stage is unclear. As such, all the major stakeholders have an active role in promoting ethical decision-making and ensuring that the algorithmic system operates in a fair and transparent manner.

Perceptible But False Ethical Affordances and Moral Distribution

As described earlier, some ethical affordances are false in that they are perceivable and actionable, even though the associated action may be invalid or misleading. False ethical affordances (comparable to the "close door" button on many elevators) do not have a real ethical effect and can result in undesirable consequences, as the recommended transparent action has no measurable effect and/or simply misleads users into a false sense of ethical decision-making.

An example of a false ethical affordance is an HRM analytical tool that produces recommendations based on a spurious as opposed to a causal model. Byte, for example, may have included resume errors in its screening algorithm for the assessment of job applicants at ACME, based on patterns picked up by its machine learning algorithm. Even so, Cheng et al. (2018) states that selection models that use correlated rather than causal variables, such as resume errors, may result in candidate pools that are potentially discriminatory. Furthermore, when decisions, such as candidate selection, rely on spurious associations between variables, the resulting outcomes may become unreliable, without significant improvement from random selection (Cheng et al., 2018). In such scenarios, the user is led to believe their employees have, for example, been selected based on conscientiousness but the reality is that no meaningful prediction has been provided. Byte has culpability in such contexts for failing to ensure that an actionable affordance in TalentAnalytica fulfills its intended non-discriminatory ethical objectives.

Related to the above example, there is also the potential for perceptible ethical affordances to yield misleading instructions and/or explanations that compromise fair

communication. For example, as part of Byte's automated resume screening function, the explanation for a denied application may be inaccurate or unclear. ACME may be left to believe that the hiring denial of a protected class was reasonable and fair when in fact it was random (Cheng et al., 2018) and/or inadvertently discriminatory.

False ethical affordances can also occur because of incomplete or unavailable history in the algorithm system, where user actions are not fully tracked, resulting in biases in the algorithm of public concern. For example, if TalentAnalytica does not save a record for the initial parameters and/or fluctuations of the fairness index it would not be able to provide recounts that fulfill public accountability requirements.

Many of the examples above align with the Bryson (2018) categorization of situations in which the artifact fails to perform as intended, suggesting that the designer should be held accountable. Indeed, during the execution phase of an HRM algorithm, poor implementation of the mechanisms that exist to identify and address biases (e.g., internal identification, feedback incorporation, and external auditing) can pose a significant problem. If these mechanisms are found to be invalid, the algorithm will not operate as intended. TalentAnalytica, for example, may have mechanisms allowing users to both flag potential biases and suggest corrective measures, but if Byte failed to incorporate these findings into TalentAnalytica 2.0, the biases may be perpetuated, resulting in unfair hiring practices. In any case Byte, the designer, has active responsibility to ensure such mechanisms are valid and function well to support both ACME and societal values.

In addition to misleading users, invalid mechanisms mislead Country C regulators. For instance, if when audited, the ethical check button yields "green check" in the face of invalid mechanisms, job applicants are harmed by denied opportunities due to

biases. Public accountability suffers because Country C regulators have been misled.

As implied earlier, designers are notably responsible for false ethical affordances, regardless of the phase in which the issues appear, because they created misleading features. Relatedly, designers are in the best position to understand the limitations and potential risks of their algorithmic systems. They have access to the: (1) information concerning the design process, (2) data used to train the algorithm, and (3) decision-making rules that the algorithm follows. Thus, it is the designer's responsibility to recognize and address false ethical affordances in the system. This may involve modifying the algorithm to remove false affordances, enhancing the transparency and accountability of the system, and/or seeking input from users and other stakeholders to identify areas for improvement. Users, in comparison, likely lack the technical ability or knowledge to recognize false ethical affordances and/or understand their implications. Similarly, given the number of algorithmic systems in use (Cheng & Hackett, 2021), regulators likely lack the resources and/or expertise to fully monitor and evaluate them. In all, Martin's (2019) perspective is especially applicable to the case of false ethical affordances, where designers must take responsibility for the ethical and social implications of the algorithmic systems they create and implement. Users and regulators are likely to lack the means to identify false and/or misleading ethical affordances in the application phase.

Real But Hidden Ethical Affordances and Moral Distribution

As referred to earlier, some affordances are hidden, often intentionally, by designers and remain obscured until a user takes action to reveal them (Gaver, 1991); the

"swipe left to delete" function found in many native mobile apps is an example. Hidden ethical affordances potentially allow for user control and ethical decision-making, but they are not immediately apparent because designers opt to lessen the overall visual complexity of the algorithm and streamline workflow. In TalentAnalytica for example, affordances that boost productivity and profit margins may be perceptible, while many ethical affordances are hidden.

An area in which ethical affordances could easily be hidden involves the mathematical model used to establish fairness criteria. There are several alternatives here, each with a potentially different impact on the false positives and false negatives associated with various applicant groups (Chouldechova & Roth, 2018). The designers at Byte could elect to default to a single fairness criterion without giving HR at ACME a knob or a pull-down menu displaying the choices available to adjust the initial parameters, or to access the available capabilities to retrain the model. The impetus can be to promote the software as foolproof and/or to provide a more streamlined interface. Importantly, ACME can adjust the fairness index to suit their needs once they know that it is possible, but this typically requires a high level of software proficiency as hidden ethical affordances may or may not be explicitly documented in user manuals.

The use of hidden ethical affordances for the purposes of marketing and ease of use is a design choice notably at the crux of the accountability debate. Specifically, computer scientists tend to believe that software programs are designed as clear boxes, in which all the processes and outcomes are transparent and easily understandable (Bryson, 2018). In comparison, management researchers tend to criticize the black box nature of the software, in which the processes and outcomes are opaque and difficult to understand (Johnson & Ruane, 2017; Martin, 2019). The choice to hide

certain ethical affordances for the sake of marketing and a streamlined interface likely contributes to the black box perceptions and potentially creates ethical issues when users are unaware and/or lack access to all of the ethical implications embedded in the HRM algorithm.

The responsibility of designers concerning hidden ethical affordances is threefold. First, they must ensure that hidden ethical affordances align with legal and moral guidelines, as well as societal values, in a manner comparable to the requirements for perceptible ethical affordances. Second, the range of ethical affordances should be prioritized so that the most important ones are made perceptible to users. Third, even hidden ethical affordances should be extensively documented as part of a comprehensive user manual. Relatedly, user training should address the existence of hidden ethical affordances and how to utilize them. In this way, designers can fulfill their responsibility to foster ethical decision making among users.

Users, in addition to designers, bear some responsibility concerning hidden ethical affordances. First, users should adopt an outcome-driven perspective to re-examine the algorithm decisions from an ethical perspective (De Cremer & Kasparov, 2022) and not rely solely on the designers. Second, users should aspire to actively learn about hidden ethical affordances and master them to take maximum control of the ethical decision-making process and ensure the algorithm is as fully aligned with their ethical values as possible. Lastly, users should highlight to designers the hidden ethical affordances they regard as most critical, and request that they be made perceptible in algorithm updates, as part of ongoing feedback to designers concerning their product offering. By working collaboratively with designers, users can create an environment where ethical considerations are at the forefront of program development and usage.

Finally, the Country C regulator could consider playing a role in bridging the gap between the skill level of users (HR staff at ACME) and the complexity of algorithms with hidden ethical affordances, by for example, establishing licensing requirement (analogous to that required for the operation of motor vehicles) before an organization can purchase and use HRM algorithms. Certification would require that operators understand the ethical implications of the algorithm, including the ability to appropriately use hidden ethical affordances. Of course, as raised in connection with the potential role of regulators in other areas of algorithmic development and use, the viability of a certification process would depend on them having the resources required for its design, implementation, and enforcement.

In all, regarding hidden ethical affordances, the major stakeholders all have moral responsibilities. Designers must prioritize ethical affordances and make the crucial ones perceptible, while documenting and training users on the hidden options as documented in the user manual. Users have the responsibility to have an outcome-driven perspective that does not rely solely on the designers. This involves learning and mastering the hidden ethical affordances as well as requesting software updates as appropriate. Regulators can design and implement licensing requirements to help ensure the skill level of users is up to the task of ethically and effectively using HRM algorithms.

The ethical affordances of HRM algorithms, their definitions, examples, and suggested attribution of accountability are described in Table 5.

Insert Table 5 about here.

DISCUSSION AND FUTURE RESEARCH DIRECTIONS

The theoretical contribution of this paper lies in its innovative introduction of a design-centred perspective to help evaluate and understand the ethical implications associated with the design and use of HRM algorithms. I highlight the conflicts between two views, one rooted in computer science and the other in management, regarding who should bear moral responsibility in various contexts. Various factors that impact the distribution of moral responsibility among multiple stakeholders including designers, users, and regulators are examined using the ethical affordances construct, a pivotal concept that clarifies the nature of the distributions of the involved. A thought experiment is used to illustrate how the application of the ethical affordances framework provides a clearer basis for analyzing and assigning moral responsibility among major stakeholders. Four forms of moral responsibility: culpability, fair communication, public accountability, and active responsibility are examined. The application of the framework demonstrates how different types of ethical affordances result in varying distributions of moral responsibility among stakeholders. I now suggest several areas for future research.

First, there is a need to examine how the developers of algorithms can incorporate ethical affordances into their products to improve transparency, accountability, and user acceptance. Despite having the means and knowledge to contribute as well as some preliminary discussions (Ettlinger, 2018; Shin & Park, 2019), as a practical matter, management researchers have rarely been involved in the development and/or theorizing of affordances. Among many examples, HRM algorithms can explicitly include the measurement of constructs concerning trust, satisfaction, and fairness, and

such affordances with ethical implications can increase the confidence users have in the output. For instance, trust argumentation conceptualizes the trust relationships between users and recommenders (friends or strangers) to filter out unreliable or biased recommendations to improve recommendations to the users (Bedi & Vashisth, 2014). HRM Researchers could also contribute to the customization of these ethical affordances to meet the specific needs and values of various stakeholders, including employees, customers, and regulators, particularly in the HRM context. For instance, if the female group of employees perceive the HR system as less fair with high barriers to voice, an affordance of surveys and further voice channels can be added to increase the perceptions of justice and job satisfaction. The overall goal here is to improve the design and effective adoption of HRM algorithms in organizations.

Second, I warn against the marginalization of HR researchers and practitioners in the crucial regulatory processes governing the emerging field of management algorithms. With the increasing adoption of algorithms in various organizational settings, it is becoming essential to establish frameworks and guidelines that comprehensively assess the ethical and legal implications of these technologies (DeVos et al., 2022), involving professionals with domain expertise. Responding to the calls from the computer science discipline that everyday users should be empowered to contribute to the auditing of complex algorithms (Shen et al., 2021), HR and management researchers could help develop a comprehensive regulatory framework to fully evaluate the impact of algorithms on all the major stakeholders, and to help ensure compliance with legal and ethical standards. Moreover, cross-disciplinary researchers can collaborate with policymakers, industry associations, and governments to develop standards and best practices for algorithmic governance that balance innovation, efficiency, and social

responsibility. Such collaboration can facilitate a more comprehensive understanding of the challenges and opportunities associated with algorithmic decision-making to foster greater transparency and accountability, to ultimately build public trust in these technologies.

Third, while drafting this dissertation, I noticed the emergence of new tools often categorized as "AI (Artificial Intelligence)," such as ChatGPT, which could potential impact HRM (Andrieux et al., 2024). Early scholarly work in management, information system, and human resources management have described such technologies as either a research frontier, a system, or capabilities that encompass but are not limited to those explained by other algorithms (Berente et al., 2021; Jia et al., 2024; Varma et al., 2023). However, it is crucial to recognize that such systems often lack the true "intelligence" inherent in humans. They execute tasks without comprehension, communicate without emotions, and make recommendations without moral reflections. Additionally, I argue that the term "Artificial Intelligence" may overstate the capabilities of AI systems, leading to unrealistic expectations or even fears about their potential impact on society. Therefore, although it may be challenging to forego employing the term "AI," I propose adopting more accurate and nuanced language when addressing this subject in the future, such as "machine learning," "neural networks," or "GPT" to better reflect the *actual* capabilities and limitations of those systems. Moreover, I also call for a thorough examination of those capabilities and limitations through the lens of affordances, particularly those of ethical and moral implications. Such in-depth understanding can also serve to centralize HR researchers and practitioners in the governance of algorithmic management.

Lastly, given the increasing complexity of management algorithms, there is a

growing demand for specialists proficient in analyzing and evaluating intricate algorithmic outcomes and their interaction with HR systems. As a result, the urgency for specialized business education and certification is a pressing concern. The intricacies of these algorithms require both technical knowledge and a deep understanding of the role of HRM in organizations, including the promotion of ethical decision making. This blend of knowledge is required especially to, for example, identify and prioritize ethical affordances that are perceptible. Relatedly, HRM specialists well-versed in both complex algorithm analysis and HR system dynamics are crucial to enhancing software interpretation, execution and adaptation. The creation of targeted programs and certifications aids in educating HRM algorithm participants, ensuring their adept and ethical utilization.

CONCLUSION

My exploration of the contentious issues surrounding accountability in the design and use of HRM algorithms, reveals blurred boundaries of responsibility among the major stakeholders (i.e., designers, users, affected applicants and employees, and regulators). By applying affordance theory from an artifact-centric perspective, I highlighted various types of ethical affordances that each raise different ethical considerations. In the process, I developed a framework that can be used to assign various levels of moral responsibility more clearly to the various moral agents involved in the design, use, and regulation of management algorithms. Among other considerations, the framework highlights differences among the ethical affordances in the level of perceptible information offered to users. Areas where stakeholders can usefully collaborate to address accountability issues were also identified. In all, the framework

provided yields important insights and recommendations to guide the design and use of HRM algorithms in an ethical, socially responsible manner.

My framework contributes to both the theoretical and applied realms. From a theoretical standpoint, I lend much needed clarity to ongoing debate concerning how to attribute moral responsibility to various actors involved in the design and use of management algorithms. The related insights provide the foundation for theory development by offering new perspectives concerning the complex accountability issues associated with the design and use of HRM algorithms. From an applied perspective, I offer many practical suggestions for designers, users, and regulators, especially highlighting the importance of designing ethical affordances that are perceptible to users and to other stakeholders, as well as the need for comprehensive user documentation and training. Finally, I also call for regulatory institutions to establish a comprehensive legal framework, and to advocate for a certification system for users of HRM algorithms to build confidence in their design and use among the range of major stakeholders.

Chapter 5 OVERALL CONCLUSION AND FUTURE RESEARCH AGENDA

This thesis consists of a review paper and two conceptual papers focusing on HRM algorithms, with the overarching goal of making research concerning HRM algorithms more relevant to the major stakeholders. Significant gaps between HRM research and practice can be addressed using my conceptual frameworks to enhance our understanding of algorithmic fairness and the delegation of moral accountability.

My review paper critically examines the use of algorithms in HRM and highlights the cautious outlook in the existing scholarship. It emphasizes the need for high-quality research to bridge the gap between academics and management practice, shedding light on the definition of HRM algorithms and exploring specific topics and pressing research questions. By considering both academic research and trade journal articles, I provide a comprehensive view of the field.

The two conceptual papers address key fairness and accountability issues associated with HRM algorithms. The first conceptual piece presents a conceptual framework to enhance our understanding of algorithmic fairness in HR management. By examining perspectives from both computer science and management, I broaden the definition of algorithmic fairness by incorporating causal, legal, and moral components. The framework highlights the importance of traceability in HR algorithms, offering researchers and practitioners an opportunity to review and identify potential systemic discrimination in management systems.

The second conceptual paper adopts an artifact-centric approach to analyze the properties of HRM algorithms and explores the ethical implications in various

decision-making scenarios. It applies the concept of ethical affordances to the development and use of HRM algorithms thereby yielding a framework that provides a basis for clearly assigning moral responsibility to various stakeholders during both the development and execution stages.

Based on the findings and frameworks presented in this thesis, there are three primary streams of future research that can advance the field:

Theoretical Advancements: Given the limited body of theoretical research concerning HRM algorithms, my objective is to extend the affordance-oriented theoretical perspective regarding their use. This extension aims to contribute valuable insights for the design science and software development of such algorithms, lessening the gap between the HRM field and information systems, computer science/AI, and AI ethics. In a related domain, Karahanna et al. (2018) present a theoretical perspective on social media use, emphasizing needs-affordance-features based on self-determination and psychological ownership theory. They highlight the affordances offered by social media applications to fulfill psychological needs. However, HRM algorithms are inherently more intricate than the psychological needs of a homogeneous body of "users" of similar interests. As I highlight, HRM algorithms encompass a blend of statistical, legal, and moral considerations across various stakeholders. Further exploration from a design science standpoint can articulate the specific affordances tailored for different stakeholders beyond a moral perspective. Additionally, design science research can leverage my proposed recommendations to offer guidelines for developing effective features in HRM algorithms. Understanding the effects of synergies (complementary) or conflicts (substitutive) among different affordances designed to meet the needs of diverse stakeholders is crucial. These aspects present intriguing avenues for future

design science research, offering potentially fruitful paths for HRM algorithm studies to discern how to effectively balance the power dynamics within the HRM system.

Empirical Application of AI Instruments: Following the rapid evolution of AI instruments, my empirical research focuses on harnessing the capabilities of AI in OB/HRM research methodology. This trajectory includes but is not limited to: a. *Integration of AI algorithms and simulations in experiment design:* This involves the incorporation of AI algorithms and simulations to attain near-perfect control over experimental conditions. b. *Automatic pattern recognition and hypotheses generation:* Utilizing AI for automated pattern recognition to dissect complex research trends and generate hypotheses based on identified patterns. c. *Text analysis based on natural language processing and generative AI models for qualitative analysis:* Leveraging natural language processing and generative AI models for comprehensive text analysis, particularly in qualitative research, to extract meaningful insights. d. *Chatbots and advanced programming for survey administration:* Employing chatbots and sophisticated programming techniques to enhance the process of survey administration, ensuring efficiency and participant engagement. The integration of these AI-driven methodologies holds the potential to revolutionize organizational behaviour and HRM research by providing more robust experimental control, efficient pattern recognition, deeper qualitative insights, and streamlined survey administration. Embracing these advancements is pivotal for staying at the forefront of research practice.

Investigating Bias and Discrimination: A key aspect of my future research agenda involves actively investigating bias and discrimination issues tied to HRM algorithms. I will focus on identifying and mitigating algorithmic biases, examining fairness and equity of decision-making processes, and investigating potential discriminatory effects

on various demographic groups. Furthermore, I will explore strategies to reduce bias in data collection, preprocessing, and algorithmic models while considering the ethical and legal implications of algorithmic biases in HRM practices. Following the spirit of affordance-oriented theoretical contribution and harnessing the computational power of AI instruments, my research aims to contribute to the development of fair, unbiased HRM algorithms. By advancing theoretical understanding, conducting empirical studies, and investigating bias and discrimination, future research in the field can contribute to the development of responsible and effective HRM algorithmic practices. This multidimensional approach will help bridge the gap between theoretical frameworks and practical applications, leading to more informed decision-making and improved HRM outcomes.

Chapter 6 AFTERWORD

To my fellow researchers, particularly PhD students and those embarking on the path of pursuing an artifact-centric research perspective from a management view, I want to acknowledge the challenges and excitement that lie ahead. This stream of work can be daunting, given its interdisciplinary nature and the rapid development of the practice. With an overwhelming number of new articles emerging from various disciplines such as computer science, AI robotics, AI ethics, information systems, operational management, and general management, along with the continuous emergence of disruptive tools like generative AI, it can be intimidating to ponder the question, "What can *I* do?"

However, I want to assure you that you are not alone in this journey. We are part of a vibrant community of researchers who are collectively pushing the boundaries of knowledge in this field. Despite the ever-evolving landscape, we have the opportunity to contribute meaningfully to the understanding of artifact-centric research and its implications for management. The key is to keep trying, remain curious, and embrace the interdisciplinary nature of our work. By collaborating and exchanging ideas, and engaging in ongoing dialogues with scholars across various disciplines, we can, hopefully, collectively answer the question, "What can *we* do?"

It's important to remember that we have a supportive network of researchers who are willing to help and collaborate. Don't hesitate to reach out to me or others in the field whenever you need guidance, advice, or simply someone to bounce ideas off of. Together, we can navigate the intricacies of this research stream and make valuable contributions to our field.

References

Adams, J. S. & Freedman, S. (1976). Equity Theory Revisited: Comments and Annotated Bibliography. In Berkowitz, L. & Walster, E., editors, ***Advances in Experimental Social Psychology***, volume 9, pages 43–90. Academic Press.

Aerts, D. & Czachor, M. (2004). Quantum aspects of semantic analysis and symbolic artificial intelligence. ***Journal of Physics A: Mathematical and General***, 37(12):L123–L132.

Aguinis, H., Beltran, J. R., Archibold, E. E., Jean, E. L., & Rice, D. B. (2022). Thought experiments: Review and recommendations. ***Journal of Organizational Behavior***, 1:1–17.

Alexander, F. (2016). Watson Analytics Use Case for HR: Retaining valuable employees.

Andrieux, P., Johnson, R. D., Sarabadani, J., & Van Slyke, C. (2024). Ethical considerations of generative AI-enabled human resource management. ***Organizational Dynamics***, 53(1):101032.

Angrave, D., Charlwood, A., Kirkpatrick, I., Lawrence, M., & Stuart, M. (2016). HR and analytics: Why HR is set to fail the big data challenge. ***Human Resource Management Journal***, 26(1):1–11.

Angrist, J. D. & Krueger, A. B. (1999). Empirical strategies in labor economics. In Ashenfelter, O. C. & Card, D., editors, ***Handbook of Labor Economics***, volume 3, pages 1277–1366. Elsevier.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. ***The Leadership Quarterly***, 21(6):1086–1120.

Aouadni, I. & Rebai, A. (2017). Decision support system based on genetic algorithm and multi-criteria satisfaction analysis (MUSA) method for measuring job satisfaction. ***Annals of Operations Research***, 256(1):3–20.

Bansal, P., Bertels, S., Ewart, T., MacConnachie, P., & O'Brien, J. (2012). Bridging the Research–Practice Gap. ***Academy of Management Perspectives***, 26(1):73–92.

Barends, E., Rousseau, D., & Briner, R. (2014). Evidence-based Management: The Basic Principles. Technical report, Center for Evidence-based Management, Amsterdam.

Barney, J. B. & Wright, P. M. (1998). On becoming a strategic partner: The role of human resources in gaining competitive advantage. ***Human Resource Management***, 37(1):31–46.

Bartunek, J. M. & Rynes, S. L. (2010). The Construction and Contributions of "Implications for Practice": What's in Them and What Might They Offer? ***Academy of Management Learning & Education***, 9(1):100–117.

Basukie, J., Wang, Y., & Li, S. (2020). Big data governance and algorithmic management in sharing economy platforms: A case of ridesharing in emerging markets. ***Technological Forecasting and Social Change***, 161:120310.

Becker, J.-m. & Ismail, I. R. (2016). Accounting for sampling weights in PLS path modeling: Simulations and empirical examples. ***European Management Journal***, 34(6):606–617.

Bedi, P. & Vashisth, P. (2014). Empowering recommender systems using trust and argumentation. ***Information Sciences***, 279:569–586.

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing Artificial Intelligence. ***MIS Quarterly***, 45:1433–1450.

Bersin, J. (2012). How BigData Tools Helps HR Understand You. ***Forbes***, Feb 29.

Bersin, J. (2013a). Are applicant tracking systems now a commodity?

Bersin, J. (2013b). Big Data in Human Resources: Talent Analytics (People Analytics) Comes of Age. https://www.forbes.com/sites/joshbersin/2013/02/17/bigdata-in-human-resources-talent-analytics-comes-of-age/.

Bersin, J., Mariani, J., & Monahan, K. (2016). Will IoT technology bring us the quantified employee? The Internet of Things in human resources. [Web log article]., Deloitte.

Birgillito, G. & Birgillito, M. (2018). Algorithms and ratings: Tools to manage labour relations. Proposals to renegotiate labour conditions for platform drivers. ***Labour & Law Issues***, 4(2):C. 25–50.

Boltanski, L. & Thévenot, L. (1983). Finding one's way in social space: A study based on games. ***Social Science Information***, 22(4-5):631–680.

Boolos, G. S., Burgess, J. P., & Jeffrey, R. C. (2002). ***Computability and Logic***. Cambridge University Press, Cambridge ; New York, 4th edition edition.

Boudreau, J. W. & Berger, C. J. (1985). Decision-Theoretic Utility Analysis Applied to Employee Separations and Acquisitions. ***Journal of Applied Psychology***, 70(3):581–612.

Boulton, C. (2017). Think twice before you hire a chief AI officer. ***CIO***.

Bovens, M. (1998). ***The Quest for Responsibility: Accountability and Citizenship in Complex Organisations***. Cambridge University Press, Cambridge, England ; New York.

Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework. ***European Law Journal***, 13(4):447–468.

Brantingham, P. J., Valasik, M., & Mohler, G. O. (2018). Does Predictive Policing Lead to Biased Arrests? Results From a Randomized Controlled Trial. ***Statistics and Public Policy***, 5(1):1–6.

Braun, M., Hummel, P., Beck, S., & Dabrock, P. (2021). Primer on an ethics of AI-based decision support systems in the clinic. ***Journal of Medical Ethics***, 47(12):e3–e3.

Brown, J. R. (2010). ***The Laboratory of the Mind: Thought Experiments in the Natural Sciences***. Routledge, New York, 2 edition edition.

Bryson, J. J. (2018). Patiency is not a virtue: The design of intelligent systems and systems of ethics. ***Ethics and Information Technology***, 20(1):15–26.

Bryson, J. J. (2020). The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation. In Dubber, M. D., Pasquale, F., & Das, S., editors, ***The Oxford Handbook of Ethics of AI***, page 0. Oxford University Press.

Bryson, J. J. & Kime, P. P. (2011). Just an artifact: Why machines are perceived as moral agents. In ***Twenty-Second International Joint Conference on Artificial Intelligence***, page 6.

Bryson, J. J. & Theodorou, A. (2019). How Society Can Maintain Human-Centric Artificial Intelligence. In ***Human-Centered Digitalization and Services***, pages 305–323. Springer, Singapore.

Bucher, E. L., Schou, P. K., & Waldkirch, M. (2021). Pacifying the algorithm – Anticipatory compliance in the face of algorithmic management in the gig economy. ***Organization***, 28(1):44–67.

Buolamwini, J. & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In ***Proceedings of the 1st Conference on Fairness, Accountability and Transparency***, pages 77–91. PMLR.

Caldwell, R. (2008). HR business partner competency models: Re-contextualising effectiveness. ***Human Resource Management Journal***, 18(3):275–294.

Calo, R. (2015). Robotics and the Lessons of Cyberlaw. ***California Law Review***, 103:513.

Calvert, S. C., Heikoop, D. D., Mecacci, G., & van Arem, B. (2020). A human centric framework for the analysis of automated driving systems based on meaningful human control. ***Theoretical Issues in Ergonomics Science***, 21(4):478–506.

Calvert, S. C., Mecacci, G., Heikoop, D. D., & de Sio, F. S. (2018). Full platoon control in Truck Platooning: A Meaningful Human Control perspective. In ***2018 21st International Conference on Intelligent Transportation Systems (ITSC)***, pages 3320–3326.

Calvert, S. C., van Arem, B., Heikoop, D. D., Hagenzieker, M., Mecacci, G., & de Sio, F. S. (2021). Gaps in the Control of Automated Vehicles on Roads. ***IEEE Intelligent Transportation Systems Magazine***, 13(4):146–153.

Canós-Darós, L. (2013). An algorithm to identify the most motivated employees. ***Management Decision***, 51(4):813–823.

Cardoen, B., Demeulemeester, E., & Beliën, J. (2010). Operating room planning and scheduling: A literature review. ***European Journal of Operational Research***, 201(3):921–932.

Carr, ADRIAN. & Zanetti, L. A. (1999). Metatheorizing the Dialectic of Self and Other: The Psychodynamics in Work Organizations. ***American Behavioral Scientist***, 43(2):324–345.

Casey, L. (2023). Dealing With Long Covid at Work. ***Wall Street Journal***.

Castelvecchi, D. (2016). Can we open the black box of AI? ***Nature News***, 538(7623):20.

Chaffin, D., Heidl, R., Hollenbeck, J. R., Howe, M., Yu, A., Voorhees, C., & Calantone, R. (2017). The Promise and Perils of Wearable Sensors in Organizational Research. ***Organizational Research Methods***, 20(1):3–31.

Cheng, M. (2017). Causal Modeling in HR Analytics: A Practical Guide to Models, Pitfalls, and Suggestions. ***Academy of Management Proceedings***, 2017(1):17632.

Cheng, M. M. & Hackett, R. D. (2021). A critical review of algorithms in HRM: Definition, theory, and practice. ***Human Resource Management Review***, 31(1):100698.

Cheng, M. M., Li, C., & Hackett, R. D. (2018). Simulation and big data: In search of causality in big data-related managial decision making. Working Paper 2018-02, Michael Lee-Chin & Family Institute for Strategic Business Studies.

Cheng, X., Zhang, X., Yang, B., & Fu, Y. (2022). An investigation on trust in AI-enabled collaboration: Application of AI-Driven chatbot in accommodation-based sharing economy. ***Electronic Commerce Research and Applications***, 54:101164.

Chiappa, S. (2019). Path-Specific Counterfactual Fairness. ***Proceedings of the AAAI Conference on Artificial Intelligence***, 33:7801–7808.

Chouldechova, A. & Roth, A. (2018). The Frontiers of Fairness in Machine Learning. ***arXiv:1810.08810 [cs, stat]***.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). ***Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences***. Routledge, New York, 3rd edition.

Cohen-Charash, Y. & Spector, P. E. (2001). The Role of Justice in Organizations: A Meta-Analysis. ***Organizational Behavior and Human Decision Processes***, 86(2):278–321.

Colomo-Palacios, R., González-Carrasco, I., López-Cuadrado, J. L., Trigo, A., & Varajao, J. E. (2014). I-Competere: Using applied intelligence in search of competency gaps in software project managers. ***Information Systems Frontiers***, 16(4):607–625.

Cook, K. S. & Hegtvedt, K. A. (1983). Distributive Justice, Equity, and Equality. ***Annual Review of Sociology***, 9:217–241.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). ***Introduction to Algorithms***. The MIT Press, Cambridge, Mass, third edition edition.

Cowgill, B. (2019). Bias and Productivity in Humans and Machines. SSRN Scholarly Paper ID 3433737, Social Science Research Network, Rochester, NY.

Cowgill, B. & Tucker, C. E. (2020). Algorithmic Fairness and Economics. SSRN Scholarly Paper ID 3361280, Social Science Research Network, Rochester, NY.

Crilly, D., Hansen, M., & Zollo, M. (2016). The Grammar of Decoupling: A Cognitive-Linguistic Perspective on Firms' Sustainability Claims and Stakeholders' Interpretation. ***Academy of Management Journal***, 59(2):705–729.

Danaher, J. (2016). The Threat of Algocracy: Reality, Resistance and Accommodation. ***Philosophy & Technology***, 29(3):245–268.

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. ***Reuters***.

Datta, A., Tschantz, M., & Datta, A. (2015). Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. ***Proceedings on privacy enhancing technologies***, 2015(1):92–112.

Davenport, T. H. (2014). ***Big Data at Work: Dispelling the Myths, Uncovering the Opportunities***. Harvard Business Review Press, Boston Massachusetts, illustrated edition edition.

De Cremer, D. & Kasparov, G. (2022). The ethical AI—paradox: Why better technology needs more and not less human responsibility. ***AI and Ethics***, 2(1):1–4.

Deadrick, D. L. & Gibson, P. A. (2009). Revisiting the research–practice gap in HR: A longitudinal analysis. ***Human Resource Management Review***, 19(2):144–153.

DeVos, A., Dhabalia, A., Shen, H., Holstein, K., & Eslami, M. (2022). Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In ***CHI Conference on Human Factors in Computing Systems***, pages 1–19, New Orleans LA USA. ACM.

Dhieb, N., Ghazzai, H., Besbes, H., & Massoud, Y. (2020). A Secure AI-Driven Architecture for Automated Insurance Systems: Fraud Detection and Risk Measurement. ***IEEE Access***, 8:58546–58558.

Dietrich, A. & Haider, H. (2015). Human creativity, evolutionary algorithms, and predictive representations: The mechanics of thought trials. ***Psychonomic Bulletin & Review***, 22(4):897–915.

Dignan, L. (2018). LinkedIn launches Talent Insights for HR analytics, talent planning.

Doran, D., Schulz, S., & Besold, T. R. (2017). What Does Explainable AI Really Mean? A New Conceptualization of Perspectives.

Dubin, R. (1978). ***Theory Building***. Free Press, New York, 2nd edition edition.

Duranti, A., editor (2005). ***A Companion to Linguistic Anthropology***. Wiley-Blackwell, Malden, Mass., 1st edition edition.

Eapen, B. R., Baba, V., & Sohail, M. (2023). Evidence-based management: A design theory, template, and technology for a knowledge delivery platform. ***Metamorphosis***, In press.

Edwards, D. J. & Holt, G. D. (2009). Construction plant and equipment management research: Thematic review. ***Journal of Engineering, Design and Technology***, 7(2):186–206.

Enders, C. K. (2017). Multiple imputation as a flexible tool for missing data handling in clinical research. ***Behaviour Research and Therapy***, 98:4–18.

Ernst, A. T., Jiang, H., Krishnamoorthy, M., & Sier, D. (2004). Staff scheduling and rostering: A review of applications, methods and models. ***European Journal of Operational Research***, 153(1):3–27.

Erts, N. (2018). Top 10 HR Trends To Watch Out For In 2018. ***HR Strategy and Planning Excellence Essentials***, May.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. ***Nature***, 542(7639):115–118.

Ettlinger, N. (2018). Algorithmic affordances for productive resistance. ***Big Data & Society***, 5(1):2053951718771399.

European Commission, E. (2019). Ethics Guidelines for Trustworthy AI. Technical Report, High-level Expert Group on Artificial Intelligence, European Commision.

Evans, L. & Kitchin, R. (2018). A smart place to work? Big data systems, labour, control and modern retail stores. ***New Technology, Work and Employment***, 33(1):44–57.

Fabi, B., Raymond, L., & Lacoursière, R. (2009). Strategic alignment of HRM practices in manufacturing SMEs: A Gestalts perspective. ***Journal of Small Business and Enterprise Development***, 16(1):7–25.

Faraj, S., Pachidi, S., & Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. ***Information and Organization***, 28(1):62–70.

Feelders, A. (2002). Data mining in economic science. ***Dealing with the Data Flood: Mining Data, Text and Multimedia***, pages 166–175.

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.

Findley, D. F. & Parzen, E. (1998). A Conversation with Hirotugu Akaike. In Parzen, E., Tanabe, K., & Kitagawa, G., editors, ***Selected Papers of Hirotugu Akaike***, Springer Series in Statistics, pages 3–16. Springer, New York, NY.

Floridi, L., Fresco, N., & Primiero, G. (2015). On malfunctioning software. ***Synthese***, 192(4):1199–1220.

Freedman, D. A. (2009). ***Statistical Models: Theory and Practice***. Cambridge University Press, Cambridge ; New York, 2 edition edition.

Friedman, J. H. (1997). On Bias, Variance, 0/1—Loss, and the Curse-of-Dimensionality. ***Data Mining and Knowledge Discovery***, 1(1):55–77.

Frutos-Pascual, M. & Zapirain, B. G. (2017). Review of the Use of AI Techniques in Serious Games: Decision Making and Machine Learning. ***IEEE Transactions on Computational Intelligence and AI in Games***, 9(2):133–152.

Gal, U., Jensen, T. B., & Stein, M.-K. (2020). Breaking the vicious cycle of algorithmic management: A virtue ethics approach to people analytics. ***Information and Organization***, 30(2):100301.

Galière, S. (2020). When food-delivery platform workers consent to algorithmic management: A Foucauldian perspective. ***New Technology, Work and Employment***, 35(3):357–370.

Gallagher, D. (2018). Social Networks Are Losing Friends. ***Wall Street Journal***, page 10.

Gardner, J. (2007). The Mark of Responsibility. In Gardner, J., editor, ***Offences***

and Defences: Selected Essays in the Philosophy of Criminal Law, page 0. Oxford University Press.

Gaver, W. W. (1991). Technology affordances. In ***Proceedings of the SIGCHI Conference on Human Factors in Computing Systems***, CHI '91, pages 79–84, New York, NY, USA. Association for Computing Machinery.

Geisser, S. (2017). ***Predictive Inference***. Chapman and Hall/CRC, 1st edition edition.

Gendler, T. S. (1998). Galileo and the Indispensability of Scientific Thought Experiment. ***The British Journal for the Philosophy of Science***, 49(3):397–424.

Gendler, T. S. (2011). ***Intuition, Imagination, and Philosophical Methodology***. Oxford University Press, Oxford.

George, G., Haas, M. R., & Pentland, A. (2014). Big Data and Management. ***Academy of Management Journal***, 57(2):321–326.

Gerchak, Y., Greenstein, E., & Weissman, I. (2004). Estimating Arbitrator's Hidden Judgement in Final Offer Arbitration. ***Group Decision and Negotiation***, 13(3):291–298.

Ghoshal, S. (2005). Bad Management Theories Are Destroying Good Management Practices. ***Academy of Management Learning & Education***, 4(1):75–91.

Gibson, J. (1977). Theory of Affordances. In R., S. & J, B., editors, ***Perceiving, Acting and Knowing***, pages 67–82. Lawrence Erlbaum Associates, Hillsdale, N.J.

Gibson, J. (1986). ***The Ecological Approach to Visual Perception***. Lawrence Erlbaum Associates.

Gigerenzer, G. & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. ***Psychological Review***, 103(4):650–669.

Gigerenzer, G. & Selten, R., editors (2002). ***Bounded Rationality: The Adaptive Toolbox***. MIT Press.

Gigerenzer, G. & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In ***Simple Heuristics That Make Us Smart***, Evolution and Cognition, pages 3–34. Oxford University Press, New York, NY, US.

Glikson, E. & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. ***Academy of Management Annals***, 14(2):627–660.

Gorton, G. B. & Pennacchi, G. G. (1995). Banks and loan sales Marketing nonmarketable assets. ***Journal of Monetary Economics***, 35(3):389–411.

Gray, K. & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. ***Journal of Personality and Social Psychology***, 96(3):505–520.

Greeno, J. G. (1994). Gibson's affordances. ***Psychological Review***, 101(2):336.

Grilli, L. & Rampichini, C. (2007). Multilevel Factor Models for Ordinal Variables. ***Structural Equation Modeling: A Multidisciplinary Journal***, 14(1):1–25.

Gupta, N. & Shaw, J. D. (2014). Employee compensation: The neglected area of HRM research. ***Human Resource Management Review***, 24(1):1–4.

Gutjahr, W. J. (2011). Optimal dynamic portfolio selection for projects under a competence development model. ***OR Spectrum***, 33(1):173–206.

Hall, P., Racine, J., & Li, Q. (2004). Cross-Validation and the Estimation of Conditional Probability Densities. ***Journal of the American Statistical Association***, 99(468):1015–1026.

Hambrick, D. C. (1994). What if the Academy actually mattered? ***Academy of Management. The Academy of Management Review***, 19(1):11.

Hao, K., Cutter, C., & Morenne, B. (2023). From CEOs to Coders, Employees Experiment With New AI Programs. ***Wall Street Journal***.

Harrison, D. A., Virick, M., & William, S. (1996). Working without a net: Time, performance, and turnover under maximally contingent rewards. ***Journal of Applied Psychology***, 81(4):331–345.

Hartson, R. (2003). Cognitive, physical, sensory, and functional affordances in interaction design. ***Behaviour & Information Technology***, 22(5):315–338.

Harzing, A.-W. (2018). 63rd edition of the Journal Quality List on-line. https://harzing.com/blog/2018/07/63rd-edition-of-the-journal-quality-list-on-line.

Hatfield, J. W. & Milgrom, P. R. (2005). Matching with Contracts. ***The American Economic Review***, 95(4):913–935.

Hecht, J. (2018). Lidar for Self-Driving Cars. ***Optics and Photonics News***, 29(1):26–33.

Heikoop, D. 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.

HelloWallet (2012). HelloWallet Launches Financial Wellness Workforce Assessment For Employers: Previews Online Diagnostic Tool At White House Conference. ***PR Newswire***.

Herington, J. (2020). Measuring Fairness in an Unfair World. In ***Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society***, pages 286–292. Association for Computing Machinery, New York, NY, USA.

Heyns, C. & UN. Human Rights Council. Special Rapporteur on Extrajudicial, S. o. A. E. (2013). Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, Christof Heyns :: Addendum.

Hildebrandt, M. (2019). Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning. ***Theoretical Inquiries in Law***, 20(1):83–121.

Holbeche, L. (2009). ***Aligning Human Resources and Business Strategy***. Routledge, Amsterdam London, 2 edition edition.

Hsiao, J. P.-h., Jaw, C., Huan, T.-C., & Woodside, A. G. (2015). Applying complexity theory to solve hospitality contrarian case conundrums. ***International Journal of Contemporary Hospitality Management***, 27(4):608–647.

Hult, C. (2005). Organizational Commitment and Person-Environment Fit in Six Western Countries. ***Organization Studies***, 26(2):249–270.

Jago, A. S. (2019). Algorithms and Authenticity. ***Academy of Management Discoveries***, 5(1):38–56.

Jahan, A., Ismail, M. Y., Mustapha, F., & Sapuan, S. M. (2010). Material selection based on ordinal data. ***Materials & Design***, 31(7):3180–3187.

Jenkins Jr., G. D., Mitra, A., Gupta, N., & Shaw, J. D. (1998). Are financial incentives related to performance? A meta-analytic review of empirical research. ***Journal of Applied Psychology***, 83(5):777–787.

Jia, N., Luo, X., Fang, Z., & Liao, C. (2024). When and How Artificial Intelligence Augments Employee Creativity. ***Academy of Management Journal***, 67(1):5–32.

John-Mathews, J.-M., Cardon, D., & Balagué, C. (2022). From Reality to World. A Critical Perspective on AI Fairness. ***Journal of Business Ethics***, 178(4):945–959.

Johnson, S. & Ruane, J. (2017). Jobs in the age of artificial intelligence. ***Project Syndicate***.

Jones, D. A. & Skarlicki, D. P. (2013). How perceptions of fairness can change: A dynamic model of organizational justice. ***Organizational Psychology Review***, 3(2):138–160.

Kahneman, D., Slovic, P., & Tversky, A., editors (1982). ***Judgment under Uncertainty: Heuristics and Biases***. Cambridge University Press, Cambridge ; New York, first soft cover edition edition.

Kang, S. K., DeCelles, K. A., Tilcsik, A., & Jun, S. (2016). Whitened Résumés: Race and Self-Presentation in the Labor Market. ***Administrative Science Quarterly***, 61(3):469–502.

Kant, I. (1785). ***Grundlegung Zur Metaphysik Der Sitten***. Hartknoch, Leipzig.

Karahanna, E., Xu, S. X., Xu, Y., & Zhang, N. A. (2018). The Needs–Affordances–Features Perspective for the Use of Social Media. ***MIS Quarterly***, 42(3):737–A23.

Karlsson, M. M. (2002). Agency and Patiency: Back to Nature? ***Philosophical Explorations***, 5(1):59–81.

Kean, S. (2018). Getting Smarter All the Time. ***Wall Street Journal***, page 2018 15.

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.

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding Discrimination through Causal Reasoning. In ***Advances in Neural Information Processing Systems***, volume 30, pages 656–666. Curran Associates, Inc.

Kim, S., Wang, Y., & Boon, C. (2021). Sixty years of research on technology and human resource management: Looking back and looking forward. ***Human Resource Management***, 60(1):229–247.

Knuth, D. (1997). ***Art of Computer Programming, The: Fundamental Algorithms, Volume 1***. Addison-Wesley Professional, Reading, Mass, 3rd edition edition.

Koch, J. L. & Rhodes, S. R. (1981). Predictors of turnover of female factory workers. ***Journal of Vocational Behavior***, 18(2):145–161.

Köchling, A. & Wehner, M. C. (2023). Better explaining the benefits why AI? Analyzing the impact of explaining the benefits of AI-supported selection on applicant responses. ***International Journal of Selection and Assessment***, 31(1):45–62.

Kornberger, M. & Mantere, S. (2020). Thought Experiments and Philosophy in Organizational Research. ***Organization Theory***, 1(3):2631787720942524.

Kozuka, S. (2019). Japan's Regulatory Response to Digital Platforms:. ***Zeitschrift für Japanisches Recht***, 24(48):95–110.

Kroeck, K. G., Barrett, G. V., & Alexander, R. A. (1983). Imposed Quotas and Personnel Selection A Computer Simulation Study. ***Journal of Applied Psychology***, 68(1):123–136.

Kuron, L. K. J., Schweitzer, L., Lyons, S., & Ng, E. S. W. (2016). Career profiles in the "new career": Evidence of their prevalence and correlates. ***Career Development International***, 21(4):355–377.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual Fairness. In ***Advances in Neural Information Processing Systems***, volume 30, pages 4066–4076. Curran Associates, Inc.

Lado, A. A. & Wilson, M. C. (1994). Human resource systems and sustained competitive advantage: A competency-based perspective. ***The Academy of Management Review***, 19(4):699–727.

Lamberton, C., Brigo, D., & Hoy, D. (2017). Impact of Robotics, RPA and AI on the insurance industry: Challenges and opportunities. ***Journal of Financial Perspectives***, 4(1):8–20.

Laney, D. (2001). 3D data management: Controlling data volume, velocity, and variety. Techreport 949, META Group.

Latour, B. (1992). Where are the missing masses? The sociology of a few mundane artifacts. In Bijker, W. & Law, J., editors, ***Shaping Technology/Building Society: Studies in Sociotechnical Change***, pages 225–258. MIT Press, Cambridge, MA.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. ***Big Data & Society***, 5(1):2053951718756684.

Lee, W.-P., Liu, C.-H., & Lu, C.-C. (2002). Intelligent agent-based systems for personalized recommendations in Internet commerce. ***Expert Systems with Applications***, 22(4):275–284.

Lemmergaard, J. (2009). From administrative expert to strategic partner. ***Employee Relations***, 31(2):182–196.

Lennox, J. (1991). Darwinian Thought Experiments: A Function for Just-so Stories.

In ***Thought Experiments in Science and Philosophy. G. Massey, T. Horowitz Eds***, pages 223–245. Rowman & Littlefield.

Leonardi, P. (2013). When does technology use enable network change in organizations? A comparative study of feature use and shared affordances. ***MIS Quarterly***, 37(3):749–775.

Liang, T.-P., Yang, Y.-F., Chen, D.-N., & Ku, Y.-C. (2008). A semantic-expansion approach to personalized knowledge recommendation. ***Decision Support Systems***, 45(3):401–412.

Lin, C. & Hsu, M.-l. (2010). Holistic decision system for human resource capability identification. ***Industrial Management & Data Systems***, 110(2):230–248.

Lind, E. A., Kanfer, R., & Earley, P. C. (1990). Voice, control, and procedural justice: Instrumental and noninstrumental concerns in fairness judgments. ***Journal of Personality and Social Psychology***, 59:952–959.

Liu, P., Chen, J., Lu, Z., & Song, X. (2015). Transformation Structural Equation Models With Highly Nonnormal and Incomplete Data. ***Structural Equation Modeling: A Multidisciplinary Journal***, 22(3):401–415.

Lockton, D. (2005). Architectures of Control in Consumer Product Design.

Ludwig, S., de Ruyter, K., Friedman, M., Brüggen, E. C., Wetzels, M., & Pfann, G. (2013). More than Words: The Influence of Affective Content and Linguistic Style Matches in Online Reviews on Conversion Rates. ***Journal of Marketing***, 77(1):87–103.

Maier, J. R. A. & Fadel, G. M. (2009). Affordance based design: A relational theory for design. ***Research in Engineering Design***, 20(1):13–27.

Malik, A., Budhwar, P., Patel, C., & Srikanth, N. R. (2022). May the bots be with you! Delivering HR cost-effectiveness and individualised employee experiences in an MNE. ***The International Journal of Human Resource Management***, 33(6):1148–1178.

Marler, J. H. & Boudreau, J. W. (2017). An evidence-based review of HR Analytics. ***The International Journal of Human Resource Management***, 28(1):3–26.

Martin, H., Wright, R., & Cowan, A. (2014). Human Resources Trends and Metrics: HR Measurement Benchmarking, Third Edition. Technical report, The Conference Board of Canada, Ottawa.

Martin, K. (2019). Ethical Implications and Accountability of Algorithms. ***Journal of Business Ethics***, 160(4):835–850.

Martínez-Plumed, F., Ferri, C., Nieves, D., & Hernández-Orallo, J. (2019). Fairness and Missing Values. ***arXiv:1905.12728 [cs, stat]***.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. ***Ethics and Information Technology***, 6(3):175–183.

Mcgrenere, J. & Ho, W. (2000). Affordances: Clarifying and Evolving a Concep. In ***Proceedings of Graphics Interface 2000***, pages 179–186, Montreal.

McKenna, M. (2012). ***Conversation and Responsibility***. Oxford University Press, Oxford, New York.

Mecacci, G. & Santoni de Sio, F. (2020). Meaningful human control as reason-responsiveness: The case of dual-mode vehicles. ***Ethics and Information Technology***, 22(2):103–115.

Medvedeva, M., Wieling, M., & Vols, M. (2020). The Danger of Reverse-Engineering of Automated Judicial Decision-Making Systems. ***arXiv:2012.10301 [cs]***.

Meijerink, J. & Bondarouk, T. (2023). The duality of algorithmic management: Toward a research agenda on HRM algorithms, autonomy and value creation. ***Human Resource Management Review***, 33(1):100876.

Meijerink, J., Boons, M., Keegan, A., & Marler, J. (2021). Algorithmic human resource management: Synthesizing developments and cross-disciplinary insights on digital HRM. ***The International Journal of Human Resource Management***, 32(12):2545–2562.

Meister, J. (2017). The Future Of Work: The Intersection Of Artificial Intelligence And Human Resources. https://www.forbes.com/sites/jeannemeister/2017/03/01/the-future-of-work-the-intersection-of-artificial-intelligence-and-human-resources/.

Mellit, A., Kalogirou, S. A., Hontoria, L., & Shaari, S. (2009). Artificial intelligence techniques for sizing photovoltaic systems: A review. ***Renewable and Sustainable Energy Reviews***, 13(2):406–419.

Meloni, C. (2016). State and Individual Responsibility for Targeted Killings by Drones. In ***Drones and Responsibility***. Routledge.

Mittelstadt, B. (2016). Auditing for Transparency in Content Personalization Systems. ***International Journal of Communication***, 10:12.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. ***Big Data & Society***, 3(2):2053951716679679.

Monrouxe, L. V., Rees, C. E., Endacott, R., & Ternan, E. (2014). 'Even now it makes me angry': Health care students' professionalism dilemma narratives. ***Medical Education***, 48(5):502–517.

Moore, P. V., Upchurch, M., & Whittaker, X. (2018). Humans and Machines at Work: Monitoring, Surveillance and Automation in Contemporary Capitalism. In Moore, P. V., Upchurch, M., & Whittaker, X., editors, ***Humans and Machines at Work: Monitoring, Surveillance and Automation in Contemporary Capitalism***, Dynamics of Virtual Work, pages 1–16. Springer International Publishing, Cham.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. ***Proceedings of the National Academy of Sciences***, 109(41):16474–16479.

Mowday, R. T. (1997). 1996 Presidential Address: Reaffirming Our Scholarly Values. ***The Academy of Management Review***, 22(2):335–345.

Nadkarni, S. & Chen, J. (2014). Bridging Yesterday, Today, and Tomorrow: CEO Temporal Focus, Environmental Dynamism, and Rate of New Product Introduction. ***Academy of Management Journal***, 57(6):1810–1833.

Neale, M. (2023). How Hiring Managers Can Avoid Dangerous Misuses of Generative AI.

Newman, D. T., Fast, N. J., & Harmon, D. J. (2020). When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. ***Organizational Behavior and Human Decision Processes***, 160:149–167.

Noe, R., Hollenbeck, J., Gerhart, B., Wright, P., & Eligh, L. (2016). ***Strategic Human Resource Management: Gaining a Competitive Advantage***. McGraw Hill Ryerson, Toronto, Canada, 2 edition edition.

Nof, S. Y. & Grant, F. H. (1991). Adaptive/predictive scheduling: Review and a general framework. ***Production Planning & Control***, 2(4):298–312.

Norman, E. (1998). The Nature of Technology for Design. ***International Journal of Technology and Design Education***, 8(1):67–87.

Noto La Diega, G. (2018). Against the Dehumanisation of Decision-Making – Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information. ***JIPITEC***, 9(1).

Orlikowski, W. J. & Iacono, C. S. (2001). Research Commentary: Desperately Seeking the "IT" in IT Research—A Call to Theorizing the IT Artifact. ***Information Systems Research***, 12(2):121–134.

Pagallo, U. (2013). ***The Laws of Robots: Crimes, Contracts, and Torts***. Springer Netherlands, Dordrecht.

Parent-Rocheleau, X. & Parker, S. K. (2022). Algorithms as work designers: How

algorithmic management influences the design of jobs. ***Human Resource Management Review***, 32(3):100838.

Pasquale, F. (2015). ***The Black Box Society: The Secret Algorithms That Control Money and Information***. Harvard University Press.

Passi, S. & Sengers, P. (2020). Making data science systems work. ***Big Data & Society***, 7(2):2053951720939605.

Pearce, J. L. (2004). What Do We Know and How Do We Really Know It? ***Academy of Management Review***, 29(2):175–179.

Pearl, J. (1984). ***Heuristics: Intelligent Search Strategies for Computer Problem Solving***. Addison-Wesley, Reading, Mass, first edition edition.

Pearl, J. (2009). ***Causality***. Cambridge University Press, Cambridge, second edition.

Pearl, J. (2010). An Introduction to Causal Inference. ***The International Journal of Biostatistics***, 6(2).

Pearl, J. & Mackenzie, D. (2018). ***The Book of Why: The New Science of Cause and Effect***. Basic Books, New York, 1st edition edition.

Peña, A., Serna, I., Morales, A., & Fierrez, J. (2020). FairCVtest Demo: Understanding Bias in Multimodal Learning with a Testbed in Fair Automatic Recruitment. In ***Proceedings of the 2020 International Conference on Multimodal Interaction***, pages 760–761. Association for Computing Machinery, New York, NY, USA.

Pessach, D. & Shmueli, E. (2020). Algorithmic Fairness. ***arXiv:2001.09784 [cs, stat]***.

Phan, P., Wright, M., & Lee, S.-H. (2017). Of Robots, Artificial Intelligence, and Work. ***Academy of Management Perspectives***, 31(4):253–255.

Pozzi, G., Pigni, F., & Vitari, C. (2014). Affordance Theory in the IS Discipline: A Review and Synthesis of the Literature. In ***AMCIS 2014 PROCEEDINGS***, page 14, Savannah.

Rasmussen, T. & Ulrich, D. (2015). Learning from practice: How HR analytics avoids being a management fad. ***Organizational Dynamics***, 44(3):236–242.

Rawls, J. (1999). ***A Theory of Justice: Revised Edition***. Belknap Press, Cambridge, MA.

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***, 35(5-6):545–575.

Rogers, H. (1967). ***Theory of Recursive Functions and Effective Computability***. The MIT Press, Cambridge, Mass.

Rousseau, D. M. (2006). Is there Such a thing as "Evidence-Based Management"? ***Academy of Management Review***, 31(2):256–269.

Rousseau, D. M., Manning, J., & Denyer, D. (2008). 11 Evidence in Management and Organizational Science: Assembling the Field's Full Weight of Scientific Knowledge Through Syntheses. ***Academy of Management Annals***, 2(1):475–515.

Rynes, S. L., Colbert, A. E., & Brown, K. G. (2002). HR Professionals' beliefs about effective human resource practices: Correspondence between research and practice. ***Human Resource Management***, 41(2):149–174.

Rynes, S. L., Giluk, T. L., & Brown, K. G. (2007). The Very Separate Worlds of Academic and Practitioner Periodicals in Human Resource Management: Implications for Evidence-Based Management. ***Academy of Management Journal***, 50(5):987–1008.

Sachan, S., Yang, J.-B., Xu, D.-L., Benavides, D. E., & Li, Y. (2020). An explainable AI decision-support-system to automate loan underwriting. ***Expert Systems with Applications***, 144:113100.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In ***Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems***, pages 1–15, Yokohama Japan. ACM.

Santoni de Sio, F. & Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. ***Philosophy & Technology***, 34(4):1057–1084.

Santoni de Sio, F. & van den Hoven, J. (2018). Meaningful Human Control over Autonomous Systems: A Philosophical Account. ***Frontiers in Robotics and AI***, 5.

Schildt, H. (2017). Big data and organizational design – the brave new world of

algorithmic management and computer augmented transparency. ***Innovation***, 19(1):23–30.

Schrödinger, E. (1935). Die gegenwärtige Situation in der Quantenmechanik. ***Naturwissenschaften***, 23(48):807–812.

Schuler, R. S. & Jackson, S. E. (1987). Linking Competitive Strategies with Human Resource Management Practices. ***Academy of Management Perspectives***, 1(3):207–219.

Schuler, R. S. & MacMillan, I. C. (1984). Gaining competitive advantage through human resource management practices. ***Human Resource Management***, 23(3):241–255.

Seale, D. A. & Rapoport, A. (1997). Sequential Decision Making with Relative Ranks: An Experimental Investigation of the "Secretary Problem">. ***Organizational Behavior and Human Decision Processes***, 69(3):221–236.

Seidel, S., Recker, J., & Brocke, J. (2013). Sensemaking and sustainable practicing: Functional affordances of information systems in green transformations. ***MIS Quarterly***, 37(4):1275–1299.

Shapiro, A. (2017). Reform predictive policing. ***Nature News***, 541(7638):458.

Shen, H., DeVos, A., Eslami, M., & Holstein, K. (2021). Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors. ***Proceedings of the ACM on Human-Computer Interaction***, 5(CSCW2):433.

Shin, D. & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. ***Computers in Human Behavior***, 98:277–284.

Shmueli, G. (2010). To Explain or to Predict? ***Statistical Science***, 25(3):289–310.

Shmueli, G. & Koppius, O. R. (2011). Predictive Analytics in Information Systems Research. ***MIS Quarterly***, 35(3):553–572.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. ***Nature***, 529(7587):484–489.

Silverman, R. E. & Gellman, L. (2015). Women in the Workplace (A Special Report) — Apps to Battle Job Bias: Software takes on hiring and workplace practices. ***Wall Street Journal***, page R.7.

Simonite, T. (2015). Probing the Dark Side of Google's Ad-Targeting System. ***MIT Technology Review***, July 6.

Singer, P. (1985). Ethics. In ***The New Encyclopaedia Britannica, 15th Edition***, volume 4, pages 627–648. Encyclopaedia Britannica, Chicago, 15th edition.

Somers, M. J. & Casal, J. C. (2009). Using Artificial Neural Networks to Model Nonlinearity. ***Organizational Research Methods***, 12(3):403–417.

Sparrow, R. (2007). Killer Robots. ***Journal of Applied Philosophy***, 24(1):62–77.

Spellman, B. A. & Kincannon, A. (2001). The Relation between Counterfactual ("But for") and Causal Reasoning: Experimental Findings and Implications for Jurors' Decisions. ***Law and Contemporary Problems***, 64(4):241–264.

Strub, M. Z., Lapinsky, M., & Abrahamson, A. (1994). A synergistic approach to enhancing your process. ***The Journal for Quality and Participation***, 17(4):60–63.

Swartz, T. (2003). Bayesian Modeling and Computations in Final-Offer Arbitration. ***Journal of Business & Economic Statistics***, 21(1):74–79.

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.

Tankard, C. (2016). What the GDPR means for businesses. ***Network Security***, 2016(6):5–8.

Tausczik, Y. R. & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. ***Journal of Language and Social Psychology***, 29(1):24–54.

Times, F. (2000). Economists make a difference on pay. ***Financial Times***, page 20.

Todd, P. M. & Gigerenzer, G. (2003). Bounding rationality to the world. ***Journal of Economic Psychology***, 24(2):143–165.

Tonidandel, S., King, E. B., & Cortina, J. M. (2018). Big Data Methods: Leveraging Modern Data Analytic Techniques to Build Organizational Science. ***Organizational Research Methods***, 21(3):525–547.

Vaishya, R., Javaid, M., Khan, I. H., & Haleem, A. (2020). Artificial Intelligence (AI) applications for COVID-19 pandemic. ***Diabetes & Metabolic Syndrome: Clinical Research & Reviews***, 14(4):337–339.

Varma, A., Dawkins, C., & Chaudhuri, K. (2023). Artificial intelligence and people management: A critical assessment through the ethical lens. ***Human Resource Management Review***, 33(1):100923.

Vassilopoulou, J., Kyriakidou, O., Özbilgin, M. F., & Groutsis, D. (2023). Scientism as illusio in HR algorithms: Towards a framework for algorithmic hygiene for bias proofing. ***Human Resource Management Journal***, In press.

Vencat, E. F. (2006). The Power of We; Companies are using YouTube-like technology to tap the ideas and energy of employees. ***Newsweek International***.

Viswesvaran, C. & Ones, D. S. (2002). Examining the Construct of Organizational Justice: A Meta-Analytic Evaluation of Relations with Work Attitudes and Behaviors. ***Journal of Business Ethics***, 38(3):193–203.

Volkoff, O. & Strong, D. (2013). Critical realism and affordances: Theorizing it-associated organizational change processes. ***MIS Quarterly***, 37(3):819–834.

Wachter, S. (2018). Normative challenges of identification in the Internet of Things: Privacy, profiling, discrimination, and the GDPR. ***Computer Law & Security Review***, 34(3):436–449.

Walker, J. (2012). Meet the New Boss: Big Data. ***Wall Street Journal***.

Walsh, D. J. (2023). ***Employment Law for Human Resource Practice***. Cengage Learning, 7th edition edition.

Walsh, J. P., Tushman, M. L., Kimberly, J. R., Starbuck, B., & Ashford, S. (2007). On the Relationship Between Research and Practice: Debate and Reflections. ***Journal of Management Inquiry***, 16(2):128–154.

Walter, D. (2018). Microsoft workplace analytics: How your business can benefit. ***Business News Daily***.

Wang, X., Wang, L., Zhang, L., Xu, X., Zhang, W., & Xu, Y. (2017). Developing an employee turnover risk evaluation model using case-based reasoning. ***Information Systems Frontiers***, 19(3):569–576.

Wang, Y. & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. ***Journal of Personality and Social Psychology***, 114(2):246–257.

Weidinger, L., McKee, K. R., Everett, R., Huang, S., Zhu, T. O., Chadwick, M. J., Summerfield, C., & Gabriel, I. (2023). Using the Veil of Ignorance to align AI systems with principles of justice. ***Proceedings of the National Academy of Sciences***, 120(18):e2213709120.

Whetten, D. A. (1989). What Constitutes a Theoretical Contribution? ***Academy of Management Review***, 14(4):490–495.

Wilson, H. & Daugherty, P. (2017). Will AI create as many jobs as it eliminates? ***MIT Sloan Blogs***, 23.

Wilson, H., Daugherty, P., & Bianzino, N. (2017). The jobs that artificial intelligence will create. ***MIT Sloan Management Review***.

Wolf, S. (1993). ***Freedom within Reason***. Oxford University Press, Oxford, New York.

Woods, A. D., Gerasimova, D., Van Dusen, B., Nissen, J., Bainter, S., Uzdavines, A., Davis-Kean, P. E., Halvorson, M., King, K. M., Logan, J. A. R., Xu, M., Vasilev, M. R., Clay, J. M., Moreau, D., Joyal-Desmarais, K., Cruz, R. A., Brown, D. M. Y., Schmidt, K., & Elsherif, M. M. (2023). Best practices for addressing missing data through multiple imputation. ***Infant and Child Development***, in press:e2407.

Wright, P. M., McMahan, G. C., & McWilliams, A. (1994). Human resources and sustained competitive advantage: A resource-based perspective. ***The International Journal of Human Resource Management***, 5(2):301–326.

Yoo, Y., Boland, R. J., Lyytinen, K., & Majchrzak, A. (2012). Organizing for Innovation in the Digitized World. ***Organization Science***, 23(5):1398–1408.

Zammuto, R. F., Griffith, T. L., Majchrzak, A., Dougherty, D. J., & Faraj, S. (2007). Information Technology and the Changing Fabric of Organization. ***Organization Science***, 18(5):749–762.

Zhou, C.-C., Yin, G.-F., & Hu, X.-B. (2009). Multi-objective optimization of material selection for sustainable products: Artificial neural networks and genetic algorithm approach. ***Materials & Design***, 30(4):1209–1215.

ZionMarketResearch (2021). Workforce Analytics Market By Deployment Mode (On Premises and Cloud), By Verticals (BFSI, Government, Retail, Healthcare, Manufacturing and Others) - Global Industry Analysis, Size, Share, Growth, Trends and Forecast 2021 – 2028. Technical Report ZMR-943, Zion Market Research.

Figure 1: Published articles on HRM algorithms over time

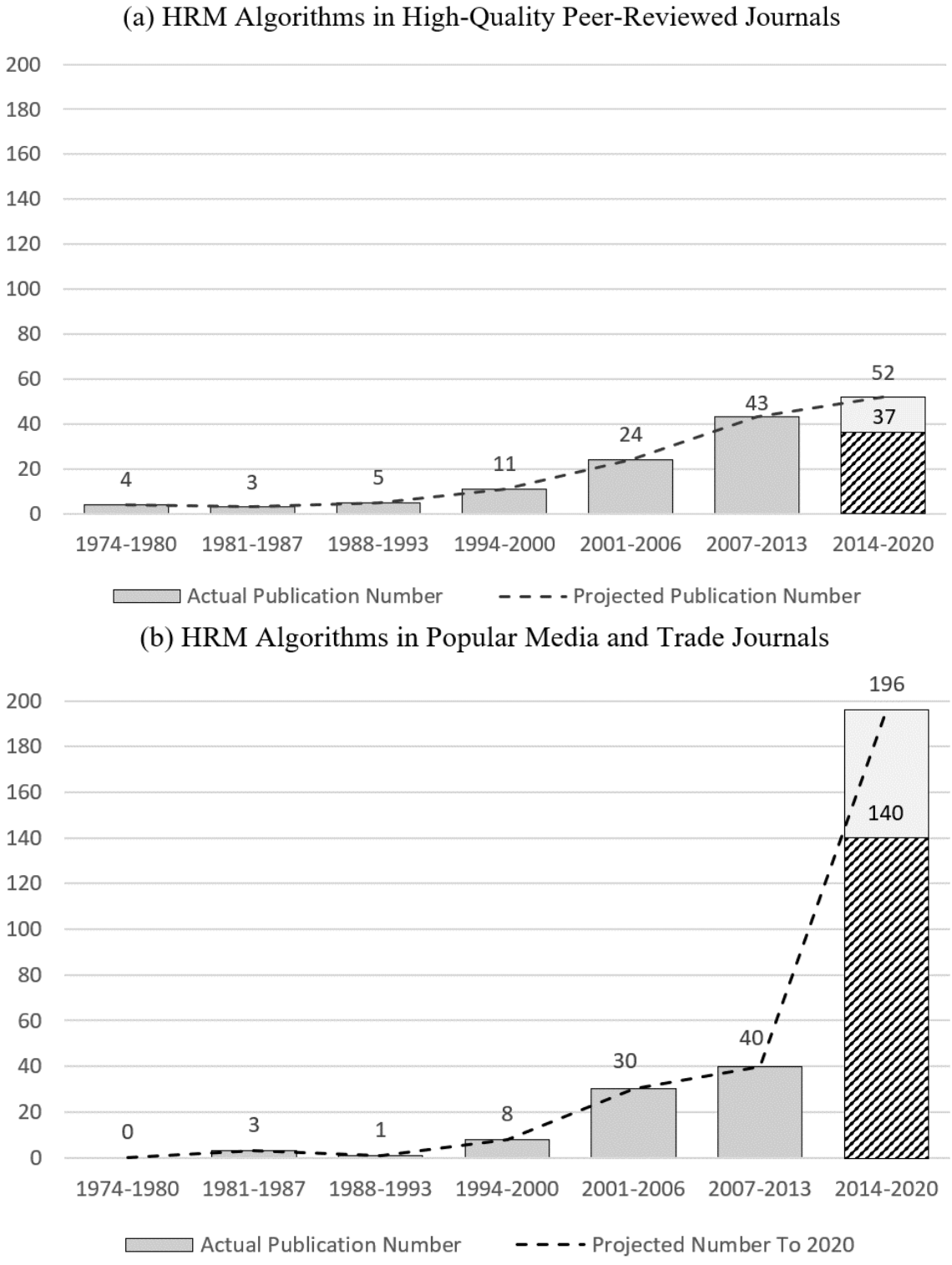

Figure 2: Comparison of Percentages of Publications in Popular Media and Trade Journals

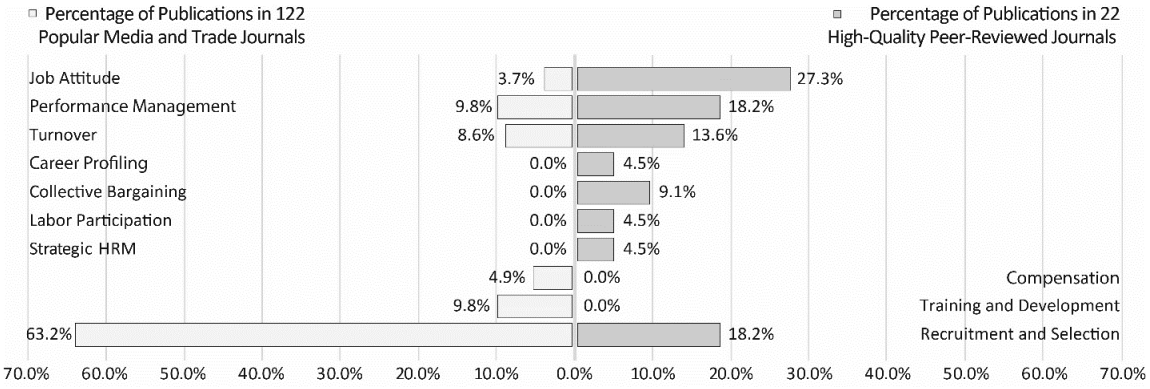

Figure 3: A Process Model of Building Algorithmic Fairness in AI Management

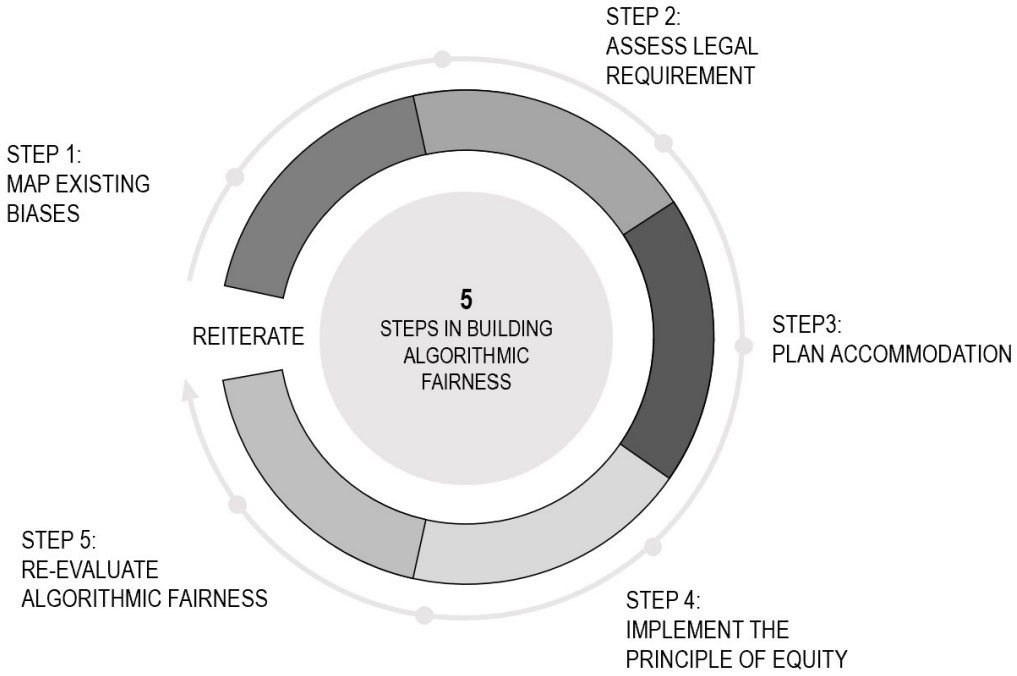

Figure 4: Categorization of Affordances Based on the Availability of Information and Affordances (Modified from Gaver (1991))

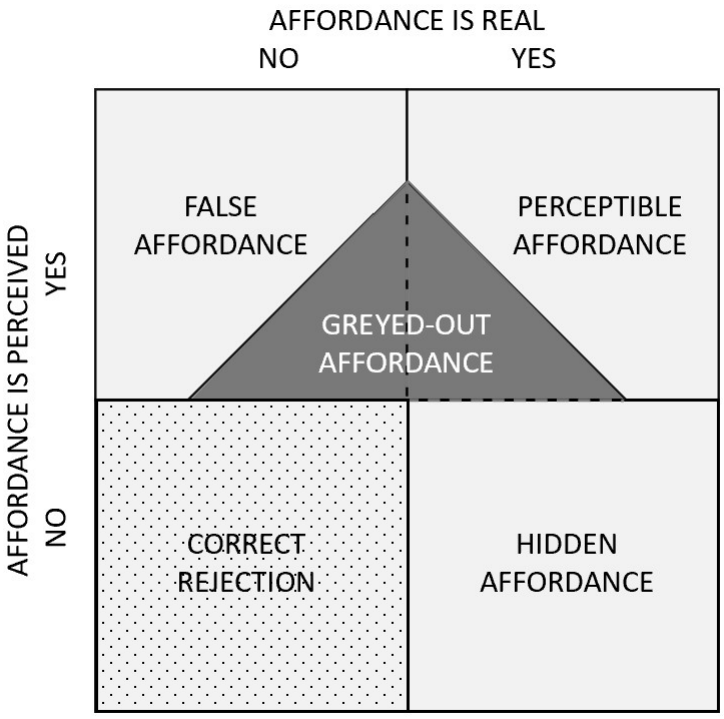

Table 1: Peer-reviewed Journal Articles on HR Topics

Study Authors	Journal	Subject Area*	Purpose of the Algorithm in HRM
Koch and Rhodes (1981)	Journal of Vocational Behavior	Organization Behavior/Studies, Human Resource Management, Industrial Relations	To predict the turnover of female workers
Kroeck, Barrett, and Alexander (1983)	Journal of Applied Psychology	Psychology	To predict recruitment and performance outcomes under imposed quota for minorities
Boudreau and Berger (1985)	Journal of Applied Psychology	Psychology	To describe how ownership separations and acquisitions influence human resources
Strub, Lapinsky, and Abrahamson (1994)	The Journal for Quality and Participation	Operations Research, Management Science, Production & Operations Management	To reduce HR overhead
Swartz (2003)	Journal of Business and Economic Statistics	Economics	To analyze bidding behaviour in final-offer arbitration
Hall, Racine, and Li (2004)	Journal of the American Statistical Association	Economics	To predict female labor force participation
Gerchak, Greenstein, and Weissman (2004)	Group Decision and Negotiation	General & Strategy	To estimate the arbitrator's hidden judgments in final offer arbitration
Hatfield and Milgrom (2005)	The American Economic Review	Economics	To match the bilateral contracts between individuals and organizations

Continued on next page

Table 1 – continued from previous page

Study Authors	Journal	Subject Area*	Purpose of the Algorithm in HRM
Gilbert and Strauss (2007)	Technometrics	Operations Research, Management Science, Production & Operations Management	Using constant social interaction data (phone calls, e-mails, co-authorships, scholarly references or citations) to predict the same variables in the future
Grilli and Rampichini (2007)	Structural Equation Modeling: A Multidisciplinary Journal	Marketing	To estimate the job satisfaction of graduates
Fabi, Raymond, and Siemsen (2007)	Journal of Management Information Systems	Management Information Systems, Knowledge Management	To forecast the most recent in a sequence of time series observations
Lutz and Spinnewijn (2008)	Journal of Labor Economics	Economics	To understand the trade-offs between profit sharing and fixed wages
McCarthy, Benjamin, and Rivard (2008)	Journal of Management Information Systems	Management Information Systems, Knowledge Management	To predict trust development in initial virtual team meetings
Barnes, Clark, and Pashby (2009)	Journal of Business and Economic Statistics	Economics	To describe the seasonal patterns of the Canadian labor market
McGrath-Champ, Firestone, and Grigsby (2009)	Journal of Environmental Economics and Management	Economics	To identify peer relationships

Continued on next page

Table 1 – continued from previous page

Study Authors	Journal	Subject Area*	Purpose of the Algorithm in HRM
Voncken et al. (2009)	European Journal of Operational Research	Operations Research, Management Science, Production & Operations Management	To analyze call centre workforce management
Eriksson and Lagerström (2009)	Structural Equation Modeling: A Multidisciplinary Journal	Marketing	To investigate the validity of measures for market orientation

Table 2: Econometric Methods to Infer Causality (Adapted from Antonakis et al., 2010)

Econometrics Methods		Brief description
Statistical adjustment	Finding all causes	Measure and control for all causes of y (impractical and not recommended)
	Propensity score analysis	Compare individuals who were selected to treatment to statistically similar controls using a matching algorithm
Quasi-experiment design	Simultaneous-equation models	Using "instruments" (exogenous sources of variance that do not correlate with the error term) to purge the endogenous x variable from bias.
	Regression discontinuity	Select individuals to treatment using a modeled cut-off.
	Difference-in-differences models	Compare a group who received an exogenous treatment to a similar control group over time.
	Heckman selection models	Predict selection to treatment (where treatment is endogenous) and then control for unmodeled selection to treatment in predicting y.

Table 3: Sources of Unfairness in Management Algorithms Objectives

Categories	Sources of Unfairness	Details	Examples
Data input	Non-uniform noises in the dataset	Biased measurement	Using "number of publications" without considering discipline or quality factors to measure performance.
		Historically biased human decisions (internal to the organization)	Using "past promotions" as a performance classifier when very few women had been promoted to leadership positions.
		Historically biased human decisions (external to the organization)	Using "sales revenue" as a performance measurement when salespeople of color are hired to work in a neighborhood with strong historical racial discrimination.
		Erroneous reports	Biases related to self-reported information.
	Missing data	Missing values	Career break due to pregnancy or disability.
		Missing sample	The company never recruited certain employees of the minority group.
		Selection bias	Higher turnover rates in black employees due to high tolerance of systemic racial discrimination.

Continued on the next page

Table 3 – continued from previous page

Categories	Sources of Unfairness	Details	Examples
Algorithms	Proxy variables	Using variables that are "proxies" for prohibited grounds	Using resume errors as a variable when it could be a proxy for country of origin/language.
	Algorithmic pathways	Feedback loop	"Self-fulfilling prophecy": Judges using COMPAS may keep defendants in detention longer based on labels of "high risk," which could increase the chances of those defendants being re-arrested.
Objectives	Conflicting objectives between overall interests and minority interests	Minimizing overall aggregated prediction errors	Using minimizing overall aggregated prediction errors as an objective for optimization.
	Conflicting objectives between the organization and its employees	Maximizing overall profit	Using profit maximization as an objective for optimization.

Table 4: Theoretical comparison between counterfactual fairness, equity theory, and justice theory

Theoretical Perspectives	Counterfactual fairness	Equity Theory	Organizational Justice Theory		
			Distributive Justice	Procedural Justice	Interactional Justice
Definition	Objective justice of tangible outcomes	Perceived ratios of contributions and benefits of one another	Perceived justice of tangible outcomes	Perceived justice of the process or procedures used in rectifying the service failure	Perceived fairness of the ways being treated
Objectivity	Relatively Objective	Subjective	Subjective	Subjective	Subjective
Universality	Aiming at building a universal standard	Based on individual perceptions	Based on individual perceptions	Based on individual perceptions	Based on individual perceptions
Focus of Comparison	Outcomes	Ratios of contributions and benefits (outcomes)	Outcomes	Procedures	Interaction/Communication
Moral Principle	Principle of Equality	Principle of Equity	Principle of Equity / Justice	Principle of Equity / Justice	Principle of Equity / Justice
Example	sameness	accommodation	sameness and accommodation where applicable	involvement	dignity
Moral Agents	programmer + manager	manager	manager	manager + employees	manager + employees

Table 5: Ethical Affordance of Management Algorithms and Moral Distribution

Type of Ethical Affordance	Development Cycle	Type of Responsibility			Attribution of Accountability
		Culpability	Active responsibility	Fair communication	Public accountability	
Perceptible Ethical Affordance	Development Phase	apparent interface that users can perceive, control, and make an ethical decision consistent with legal and moral requirements	apparent interface that users can perceive, control, and make an ethical decision based on prioritizing societal values such as fairness, justice, and human dignity	clear and understandable explanations for the objectives and the instructions of perceptible affordances	complete history that tracks and records all actions of the users, as well as any potential biases in the algorithm itself that may cause public concern	User
		apparent warnings or grey-out buttons that users can perceive, but not able to modify based on regulations. Only users who meet certain criteria can override such restrictions	apparent warnings or grey-out buttons that users can perceive, but not able to modify based on prioritizing societal values such as fairness, justice, and human dignity			
	Execution Phase		mechanism that identifies and addresses biases in the algorithm during its application	channels that incorporates feedback and complaints from users and stakeholders to modify the algorithm	mechanisms for external auditing and evaluation of the algorithmic system's outcomes and ethical implications	Shared
False Ethical Affordance	Development Phase	apparent interface that users can perceive, control, and make an ethical decision, yet is invalid or inconsistent with legal and moral requirements	apparent interface that users can perceive, control, and make an ethical decision based on prioritizing societal values such as fairness, justice, and human dignity, yet is invalid or inconsistent with the claim	misleading explanations for the objectives and the instructions of perceptible affordances	incomplete or unavailable history that fails to tracks and records all actions of the users, as well as any potential biases in the algorithm itself that may cause public concern	Designer
	Execution Phase		mechanism that identifies and addresses biases in the algorithm during its application, yet is invalid or incorrect	channels that exist but do not incorporate feedback and complaints from users and stakeholders to modify the algorithm	mechanisms that mislead external auditing and evaluation of the algorithmic system's outcomes and ethical implications	
Hidden Ethical Affordance	Development Phase	hidden function that users cannot perceive, yet can control, and make an ethical decision consistent with legal and moral requirements	hidden function that users cannot perceive, yet can control, and make an ethical decision based on prioritizing societal values such as fairness, justice, and human dignity	hidden explanations for the objectives and the instructions of perceptible affordances	hidden history that tracks and records all actions of the users, as well as any potential biases in the algorithm itself that may cause public concern	Shared
	Execution Phase		hidden mechanism that identifies and addresses biases in the algorithm during its application	hidden channels that incorporates feedback and complaints from users and stakeholders to modify the algorithm	hidden mechanisms for external auditing and evaluation of the algorithmic system's outcomes and ethical implications	