Artificial intelligence in human resources management: Challenges and a path forward[1]

Prasanna Tambe,[2], Peter Cappelli, [2] and Valery Yakubovich[3]

Abstract

We consider the gap between the promise and reality of artificial intelligence in human resource management and suggest how progress might be made. We identify four challenges in using data science techniques in HR practices: 1) complexity of HR phenomena, 2) constraints imposed by small data sets, 3) ethical questions associated with fairness and legal constraints, and 4) employee reaction to management via data-based algorithms. We propose practical responses to these challenges and converge on three overlapping principles - causal reasoning, randomization, and process formalization—that could be both economically efficient and socially appropriate for using data analytics in the management of employees.

Introduction

The speed with which the business rhetoric in management moved from big data (BD) to machine learning (ML) to artificial intelligence (AI) is staggering. The match between the rhetoric and reality is a different matter, however. Most companies are struggling to make any progress building data analytics capabilities: 41% percent of CEOs report that they are not at all prepared to make use of new data analytic tools, and only 4 percent say that they are "to a large extent" prepared (IBM 2018).

"AI" conventionally refers to a broad class of technologies that allow a computer to perform tasks that normally require human cognition, including decision-making. Our discussion here is narrower, focusing on a sub-class of algorithms within AI that rely principally on the increased availability of data for prediction tasks. For certain, there have been major advances in the domains of pattern recognition and natural language processing (NLP) over the last several years. Deep learning using neural networks has

---

[1] In the spirit of randomization as a solution to fairness concerns, the order of the authors was assigned randomly. Suggested citation: Prasanna Tambe, Cappelli, Peter and Valery Yakubovich, Artificial Intelligence in Human Resources Management: Challenges and a Path Forward (November 30, 2018). Available at SSRN: https://ssrn.com/abstract=3263878 or http://dx.doi.org/10.2139/ssrn.3263878

[2] The Wharton School, University of Pennsylvania

[3] ESSEC Business School, France

become increasingly common in some data-rich contexts and has brought us closer to true AI, which represents the ability of machines to mimic adaptive human decision-making. Nevertheless, with respect to the management of employees, where the promise of more sophisticated decisions has been articulated loudly and often, few organizations have even entered the big data stage. Only 22 percent of firms say they have adopted analytics in human resources (LinkedIn 2018), and how sophisticated the analytics are in those firms is not at all clear.

The promise of data analytics, by contrast, is easier to see in fields like marketing. While there are many questions to be answered there, they tend to be distinguished by their relative clarity, such as, what predicts who will buy a product or how changes in its presentation affect its sales. Outcomes are easily measured, are often already collected electronically by the sales process, and the number of observations – sales of a particular item across the country over time, e.g. – is very large, making the application of big data techniques feasible. Although marketing is not without its ethical conundrums, the idea that companies should be trying to sell more of their products is well-accepted as is the idea that business will attempt to influence customers to buy more.

The effective application of AI to human resources problems presents very different challenges. They range from practical to conceptual, including the fact that the nature of data science analyses when applied to people has serious conflicts with criteria societies typically see as important for making consequential decisions about individuals. Consider the following:

- A first problem is the complexity of HR outcomes, such as what constitutes being a "good employee." There are many dimensions to that construct, and measuring it with precision for most jobs is quite difficult: performance appraisal scores, the most widely-used metric, have been roundly criticized for problems of validity and reliability as well as for bias, and many employers are giving them up altogether (Cappelli and Tavis 2017). Any reasonably complex job is interdependent with other jobs and therefore individual performance is hard to disentangle from group performance (Pfeffer and Sutton 2006).
- The data sets in human resources tend to be quite small by the standards of data science. The number of employees that even a large company may have is trivial

compared to the number of purchases their customers make, for example. Moreover, many outcomes of interest are rarely observed, such as employees fired for poor performance. Data science techniques perform poorly when predicting relatively rare outcomes.

- The outcomes of human resource decisions (such as who gets hired and fired) have such serious consequences for individuals and society that concerns about fairness – both procedural and distributive justice - are paramount. Elaborate legal frameworks constrain how employers must go about making those decisions. Central to those frameworks is the concern with causation, which is typically absent from algorithm-based analyses.
- Employment decisions are also subject to a range of complex socio-psychological concerns that exist among employees, such as personal worth and status, perceived fairness, and contractual and relational expectations, that affect organizational outcomes as well as individual ones. As a result, being able to explain and also to justify the practices one uses is much more important than in other fields.
- Finally, employees are capable of gaming or adversely reacting to algorithmic-based decisions. Their actions, in turn, affect organizational outcomes.

To illustrate these concerns, consider the use of an algorithm to predict who to hire. As is typical in problems like these, the application of machine learning techniques would create an algorithm based on the attributes of employees and their job performance in the current workforce. Even if we could demonstrate a causal relationship between sex and job performance, we might well not trust an algorithm that says hire more white men because job performance itself may be a biased indicator, the attributes of the current workforce may be distorted by how we hired in the past (e.g., we hired few women), and both the legal system and social norms would create substantial problems for us if we did act on it.

In 2018, Amazon discovered that its algorithm for hiring had exactly this problem for exactly this reason, and the company took it down as a result (Meyer, 2018). Even when the sex of applicants was not used as a criterion, attributes associated with women candidates, such as courses in “Women’s Studies” caused them to be ruled out.

If we instead build an algorithm on a more objective measure, such as who gets dismissed for poor performance, the number of such cases in a typical company is too small to construct an effective algorithm. Moreover, once applicants discover the content of our hiring algorithm, they are likely to respond differently in interviews and render the algorithm worthless. Most applicants already know, for example, to answer the question "what is your worst characteristic" with an attribute that is not negative, such as, "I work too hard."

Below, we address each of these challenges separately at each stage of what we call the AI Life Cycle: Operations – Data Generation – Machine Learning – Decision-Making. We rely on key ideas from Evidence-Based Management (EBMgmt) - a theory-driven causal analysis of "small data" (Barends and Rousseau 2018; Pfeffer and Sutton 2006; Rousseau 2014). We then suggest how, given these constraints, we might make progress in the application of machine learning tools to HR. Specifically, we focus on the role of causal models in machine learning (Pearl 2009, 2018). Establishing causation is central to concerns about fairness, which are fundamental to making decisions about employees, and machine learning-based algorithms typically struggle with that challenge.

We also suggest that randomization can be useful as a decision process (Denrell, Fang, and Liu 2015; Liu and Denrell 2018), given its perceived fairness and the difficulty that analytics may otherwise have in making fair and valid decisions. We base our arguments on knowledge of contemporary practice as well as on interactions with practitioners, and in particular, a 2018 workshop that brought data science faculty together with the heads of the workforce analytics function from 20 major US corporations.

The AI Life Cycle

Figure 1 depicts a conventional AI Life Cycle: Operations, Data Generation, Machine Learning, and Decision-Making.

FIGURE 1 HERE

"**Operations**" constitute the phenomenon of interest, such as how an organization hires employees. One of the reasons for the interest in applying data science tools to human resources is because HR performs so many operations and so much money is involved in them. In the US economy as a whole, roughly 60 percent of all spending is on

labor. In service industries, the figure is much higher (MLR 2017). Below are the most common operations in human resources with corresponding prediction tasks for workforce analytics:

HR operation	**Prediction task**
Recruiting – identifying possible candidates and persuading them to apply	Are we securing good candidates?
Selection – choosing which candidate should receive job offers	Are we offering jobs to those who will be the best employees?
On-boarding - bringing an employee into an organization	Which practices cause new hires to become useful faster?
Training	What interventions make sense for which individuals, and do they improve performance?
Performance management – identifying good and bad performance	Do our practices improve job performance?
Advancement – determining who gets promoted	Can we predict who will perform best in new roles?
Retention	Can we predict who is likely to leave and manage the level of retention?
Employee benefits	Can we identify which benefits matter most to employees to know what to give them and what to recommend when there are choices, and what are the effects of those benefits (e.g., do they improve recruiting and retention)?

Each of these operations involves administrative tasks, each affects the performance of the organization in important ways, and each includes specific offices, job roles, written instructions and guidelines to execute as well as the actual activities and interactions of all parties. These operations produce volumes of data, in the form of texts, recordings, and other artifacts. As operations move to the virtual space, many of these

outputs are in the form of "digital exhaust," which is trace data on digital activities (e.g. online job applications, skills assessment) that may be used to build recruiting algorithms.

Human resource information systems, applicant tracking systems, digital exhaust, and other markers are all critical inputs for the "**data generation"** stage. Typically, this input has to be extracted from multiple databases, converted to a common format, and joined together before analysis can take place.

By "**machine learning**" (ML) we refer to a broad set of techniques that can adapt and learn from data to create algorithms that perform better and better at a task, typically prediction. Within business contexts, the most common application of machine learning technologies has been "supervised" applications, in which a data scientist creates a machine learning algorithm, determines the most appropriate metric to assess its accuracy, and trains the algorithm using the training sample. Some of the most commonly used prediction algorithms, such as logistic regression and random forest infer the outcome variable of interest from statistical correlations among observed variables. The accuracy of preliminary models is assessed on the development sample until it stabilizes at some acceptable level. The final model is run on the test sample; the accuracy of the predictions on this sample is the ultimate indicator of the model's quality.

For hiring, for example, we might see which applicant characteristics have been associated with better job performance and use that to select candidates in the future. "Algorithmic management," the practice of using algorithms to guide incentives and other tools for "nudging" platform workers and contractors in the direction of the contractee (Lee et al 2015), is applied to regular employees (e.g., Netessine and Yakubovich 2012). At present, this is principally the case in making recommendations. IBM, for example, uses algorithms to advise employees on what training make sense for them to take, based on the experiences of similar employees; the vendor Quine uses the career progression of previous employees to make recommendations to client's employees about which career moves make sense for them. Vendors such as Benefitfocus develop customized recommendations for employee benefits, much in the same way that Netflix recommends content based on consumer preferences or Amazon recommends products based on purchasing or browsing behavior.

These algorithms differ in some important ways from traditional approaches used in HR. In industrial psychology, the field that historically focused the most attention on human resource decisions, research on hiring, say, would test separate explanatory hypotheses about the relationship between individual predictors and job performance. The researcher picks the hypothesis to examine and the variables with which to examine it. This process produces lessons for hiring, one test at a time, e.g., the relationship between personality test scores and job performance, then in another exercise, the relationship between education and job performance, and so forth.

Machine learning, in contrast, generates one algorithm that makes use of many variables. The variables may not be in the cannon of the theoretical literature associated with the topic, and the researcher is not hypothesizing or indeed even examining the relationship between any one variable and the outcome being predicted. Indeed, one of the attractions of ML is its investigation of non-traditional factors because the goal is to build a better prediction rather than advancing the theory of the field in which the researcher is based by providing evidence on particular hypotheses.

"**Decision-making**," the final stage, deals with the way in which we use insights from the machine learning model in everyday operations. In the area of human resource decisions, individual managers may have more discretion now in how they use empirical evidence from data science and other models than they did in the heyday of the great corporations when hiring and other practices were standardized across an entire company. Managers today typically have the option of ignoring evidence about predictions, using it as they see fit, and generating their own data about actions like hiring in the form of interviews they structure themselves.

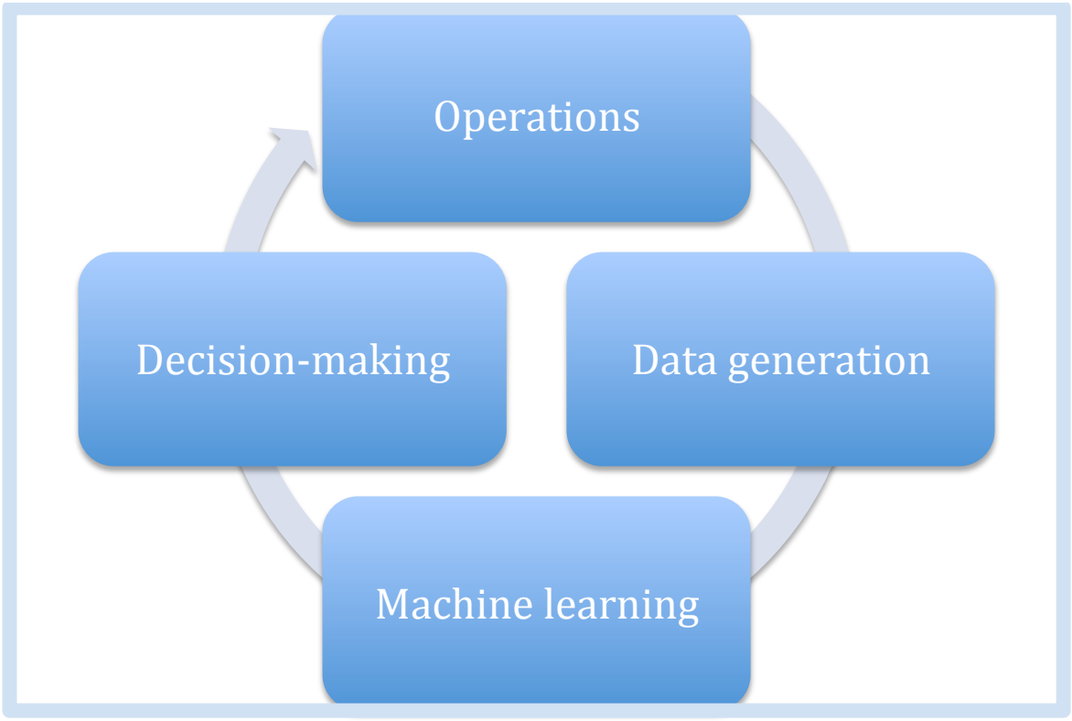


Figure 1. The life cycle of an AI-supported HR practice

Addressing AI Challenges: One Stage at a Time

In this section, we explore in detail the four general challenges to AI outlined in the Introduction: complexity of HR phenomena, small data, ethical and legal constraints, and employee reactions to AI-management. To make these challenges tractable, we discuss them in the context of the particular stages of the AI Life Cycle in which they are most relevant.

Data Generation stage

The *complexity* inherent in many HR phenomena manifests itself at the Data Generation stage. The most important source of complexity may be the fact that it is not easy to measure what constitutes a "good employee," given that job requirements are broad, monitoring of work outcomes is poor, and biases associated with assessing individual performance are legion. Moreover, complex jobs are interdependent with one another and thus one employee's performance is often inextricable from the performance of the group (Pfeffer and Sutton 2006). Without a clear definition of what it means to be a good employee, a great many HR operations face considerable difficulty in measuring performance, which is the outcome driving many HR decisions.

In terms of the data, not all attributes of HR actions are measured; not all details of operations leave digital traces that could be captured, and not all traces left can be

extracted and converted to a usable format at a reasonable cost. For example, employers may not track the channels through which applicants come to them – from referrals vs. visiting our website vs. job boards, and so forth. Most employers collect a limited amount of data on applicants, and they do not retain it for those applicants that they screen out. These choices limit the types of analyses that can be performed and the conclusions that can be drawn.

There is no list of "standard" variables that employers choose to gather and to retain through their HR operations as there might be in fields like accounting. That reduces the extent to which best practices in analytics can be transferred across organizations. Behavioral measures from attitudinal surveys, for example, vary considerably across organizations, measures of job performance differ, differences in cost accounting mean that that the detail that employers have on the costs of different operations differs enormously (e.g., are training costs tracked, and if so, are they aggregated in ways that limit the ability to examine them?), and so forth.

When tackling the challenge of data generation, employers can benefit from the lessons drawn from fields like performance management:

- Do not expect perfect measures of performance as they do not exist. It is better to choose reasonable measures and stick with them to see patterns and changes in results than to keep tinkering with systems to find the perfect measure.
- Aggregate information from multiple perspectives and over time. Digital HR tools allow for rapid real-time evaluations among colleagues using mobile devices, for example.
- Objective measures of performance outcomes based on ex ante determined goals and Key Performance Indicators are best, but they are never complete. Complement them with measures to capture less tangible outcomes, such as whether the employee fits into the company's culture, even if those measures are subjective, to prevent a situation where employees optimize on the few objective measures at the expense of everything else.
- Integrate HR data with the company's business and financial data to analyze the effects of HR practices and outcomes on business unit performance.

The complexity of HR phenomena creates another problem in the form of specialized vendors who address only one task. It is very common for an employer to have a system from one vendor to track employee performance scores, from another for applicant tracking software, from a third for compensation and payroll data, and so forth. Arguably the biggest practical challenge in using data in human resources is simply database management, aggregating existing data so that it can be examined because the systems are rarely compatible. It is no surprise that such database challenges were one of the biggest challenges reported by the HR analytics practitioners in our workshop (see Figure 2). In addition to technical barriers, our respondents reported the resistance of other functions to sharing their data with HR Departments.

To illustrate how rudimentary most of the existing database management efforts still are with HR operations, the vast majority of our practitioners reported that the software they most often used to organize and manage their data was Excel. Very few used more purpose-built tools such as Tableau that are common in data analytics. Software for bridging datasets and "data lakes" that can archive and access to different data sets clearly represent a way forward, but they can be difficult to integrate, can be viewed as confining, and face their own limitations, so they remain under-used in the HR world. To demonstrate its commitment to digital transformation as well as to benefit from it, companies' top management has to make data sharing a priority in the short-run and invest in data standardization and platform integration in the long-run.

Given these database concerns, it can be costly to analyze a question in HR for the first time. Data analytics managers, therefore, have to be careful about where to "place bets" in terms of assembling data for analysis, let alone when collecting new data. How should managers decide which HR questions to investigate, especially when so few have been examined before?

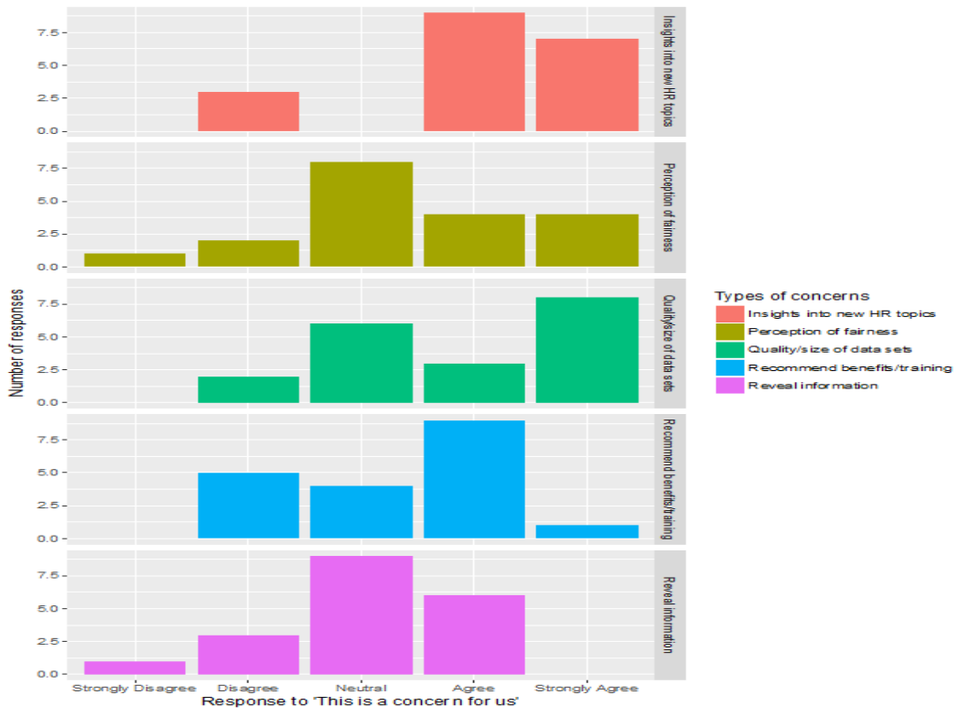
This challenge was the most important concern expressed by our practitioners (see Figure 2). Beyond the obvious criteria of cost is the likelihood of generating useable results. Our practitioners said that in this context, they relied on induction to make the choice: they ask people in HR operations what they have seen and what they think the important relationships are. Some go to senior management and solicit answers to the question of what types of problems prevent the managers from "sleeping at night." Such

experience-driven heuristics are a typical approach under uncertainty. The practitioners also indicated that another factor shaping where they placed their bets is whether anyone was willing to act on results they found.

A more systematic response would include examining the research literature in order to establish what we already know about different research questions, as the evidence-based management has long advocated (Barends and Rousseau 2018). The fact that this approach appears not to be used very often reflects the disconnect between the data science community, which understands analytics but not HR, and the HR community, which understands HR but not analytics. Many leading IT companies, such as Amazon, Google, Facebook, and Microsoft, hire as many PhDs in social sciences as in data sciences to help close this disconnect. In the longer run, AI may be able to parse the research literature itself to identify the questions that can be asked and the models that can be tested given the available data.

The last step in the process of deciding what to analyze is with an audit of what data are necessary to answer the research question and how difficult it is to assemble. For example, if the employer wants to use a machine-learning algorithm in hiring, it needs to have historical data on job candidates who were not hired, something that many employers do not retain. It may not be possible to answer questions that are important and that data science is well-suited to answer because the data are not available.

FIGURE 2 HERE



Small Data is a fundamental concern for human resource analytics. Most employers do not hire many workers, nor do they do enough performance appraisals or collect enough other data points for their current workforce to use machine learning techniques because they do not have that many employees. The machine learning literature has shown that access to larger data has substantial advantages in terms of predictive accuracy (Fortuny, Martens, and Provost 2014).

At the same time, even if data sets are not big enough for machine learning exercises, small data are often sufficient for identifying relationships: we may not be able to build a machine learning algorithm for hiring, but we probably do have enough data to answer questions about specific hiring criteria, such as whether recruiting from the CEO's Alma Mater really produces better hires. The management literature has an important advantage over data science in articulating causal relationships, as opposed to prediction from correlations among observed variables in machine learning. Only recently, some powerful voices in the computer science community have articulated the problem of causation as critical for the future of AI in human affairs (Pearl 2018). We consider the issue of causality in more detail below.

The less data we have, the less we can learn from data analytics, and the more we need from theory and prior research to identify causal predictors of the outcome of interest. AI-management requires that managers put their assumptions on the table, though, and persuade the other stakeholders in their accuracy, ultimately by using data

and empirical analysis. The formulation of such assumptions often turns into a contest among stakeholders. This is a place where process formalization that presumes contributions from stakeholders is required.

Where a formal process reveals large disagreements as to causal factors, a way forward might include generating additional data from randomized experiments in order to test causal assumptions. Google became known for running experiments for all kinds of HR phenomena, from the optimal number of interviews per job candidate to the optimal size of the dinner plate in the cafeteria (Bock 2015). If discussions, experiments, and leadership's persuasion do not lead to a reasonable consensus on the causal model that generates the outcome of interest, AI-analyses are likely to be counterproductive and thus should be avoided until more or better data can be collected.

One attraction of using vendors is their ability to combine data from many employers to generate their algorithms. Such approaches have long been used with standard paper-and-pencil selection tests, or as they are sometimes known now, pre-employment tests, such as those for sales roles. For instance, the company ADP, which handles outsourced payroll operations for thousands of companies, has been able to harness this scale to build predictive models of compensation and churn. Client companies are willing to make their data available for this exercise in return for access to the predictive models and benchmarked comparisons.

The complication for individual employers is knowing to what extent their context is distinct enough that an algorithm built on data from elsewhere will make effective predictions in their own organization. As is discussed further below, such evidence is essential to address legal concerns.

Employee reactions to data collection efforts. The fact that employees can bias their responses and the data depending on how they think the data will be used is an important concern for all HR analytic efforts. Because of this, there is demand for alternative sources of data that might be viewed as more authentic.

A great many employers now make use of social media information precisely because they believe employees are being more authentic in it. That data gets used in hiring (e.g., looking for evidence of bad behavior, looking for evidence of fit) and to assess "flight risk" or retention problems (e.g., identifying updated LinkedIn profiles).

Banks have tighter regulations requiring oversight of employees and have long analyzed email data for evidence of fraudulent behavior. They are now using it as well to identify other problems. For example, the appearance of terms like "harassment" in email traffic may well trigger an internal investigation to spot problems in the workplace.

The vendor Vibe, for example, uses natural language processing tools to gauge the tone of comments that employees post on internal chat boards, thereby helping to predict employee flight risk. Applications such as these can face some key challenges when introduced into the workplace. For instance, when employees realize their posts are being used to derive these types of measures, it can influence what and how they choose to post. Then, there are the issues that may arise around whether employees consider such use of the data to infringe upon their privacy.

Several of the companies at our workshop reported that they built models on predicting flight risk and that the best predictors did not come from traditional psychology-based findings but from data sources like social media. Many employers felt that there was an ethical problem with their own use of social media; others felt that data was ok to use but that tracking sentiment on email messages using natural language algorithms was out of bounds; still others thought that any employee-related data was appropriate to use as long as it was anonymized.

Many of these and similar considerations fall under the purview of privacy, which acquires new dimensions in the digital age: data persistence, data repurposing, and data spillovers (Tucker 2017). Data can persist well beyond the time it was generated and employers might use them for purposes unanticipated by the creator, e.g., the words from an email exchange with a colleague might be used to predict flight risk. Data of one person may also inadvertently affect other people, for example, the creators' friends tagged in posts and photos. Here employers have to account for governments' regulations of privacy issues, such as "the right to be forgotten" or the EU's General Data Protection Regulation (GDPR). The former states that business has to satisfy individuals' demands to delete their digital traces after some period of time; the latter is a comprehensive treatment of all the aspects of data privacy in the digital age (www.eugdpr.org).

In terms of technological solutions to the issue of data privacy, computer scientists are actively working on privacy-preserving data analytic methods that rely on

the notion of differential privacy in building algorithms. Here, data is randomized during the collection process, which leads to "learning nothing about an individual while learning useful information about the population" (Roth 2014: 5). Analysts do not know whose data are used in the analysis and whose data are replaced by noise, but they do know the noise generating procedure and thus can estimate the model anyway.

The practical problem with using "authentic" data, such as that in email traffic or on social media, is that it is not clear how "authentic" it really is. It is certainly true that individuals are not necessarily shaping their social media entries with the goal of influencing employers, but few people would believe that those entries are necessarily authentic. They are typically designed to create an image of the individual that is different from reality: entries about vacation cruises far outnumber entries about doing the laundry even though most of us spend far more time on the latter than the former.

The nature of such data will change quickly as soon as individuals recognize that employers are monitoring those entries: expect far more entries about self-improvement, achievements at work, and so forth. Efforts to use computer games to assess candidates is yet another effort to obtain authentic data where the employees do not necessarily know how to spin their responses. But they are already getting help from businesses like the JobTestPrep company that helps potential candidates for jobs at Lloyds Bank figure out how to score well on Lloyds' selection game.[4] Getting authentic data on applicants will remain a challenge because of the ability of candidates to game such efforts.

Machine Learning stage

An ML algorithm for predicting which candidates to hire may well perform better than anything an employer has used before. Indeed, a reasonable complaint is that prior research in human resources has not done much to help employers: the fact that most of the predictors advocated in that research, such as personality and IQ scores, predict so little of job performance (a typical validity coefficient of .30, for example, translates to explaining nine percent of the variance in performance) creates an enormous opportunity

[4] See, e.g. https://www.jobtestprep.co.uk/lloydsbank

for data analytics to do better. It will because its goal is just to predict, and it is not limited to a small number of one-at-a-time results, such as a personality test.

As noted above, finding good data with which to build an algorithm can be challenging. Because clients rarely have data on employee performance in which they feel confident, a common approach in the vendor community is to build an algorithm based on the attributes of a client firm's "best performers," which are easier to identify. Then applicants are assessed against that algorithm. Consider, for example, the vendor HireVue that helps clients conduct video interviews. Part of its offering now includes controversial algorithms based on facial expressions captured on those videos. The algorithms are trained on data from top performers at the client firm, and job candidates are assessed based on how similar their expressions are to those of the algorithm.

Is it possible that facial expressions actually predict job performance? Social scientists may find examples like this absurd because there is no reason to expect such a relationship. The machine learning models and the data scientists behind them, of course, do not care whether we know what the reason might be for such a relationship or whether it corresponds with what we know from research on humans. They only care if there is such a relationship.

Examples like this algorithm raise many concerns, though, even for the basic goal of producing a good algorithm. First, they "select on the dependent variable" by examining only those who are successful. The algorithm may well capture attributes of good performers accurately, but it is not identifying whether those attributes are truly distinct from those of other performers. Good performers and bad performers may have the same expressions in response to situations, but we will never know without examining both groups.

The use of an algorithm or indeed any decision rule in hiring is a challenge for the "learning" aspect of machine learning because of the sample selection it generates: Once we rule out hiring candidates who are not chosen by the algorithm, the opportunity to see whether other attributes might lead to better performers diminishes and may end – say if job requirements change or if new attributes appear among candidates. In other words, the opportunity for the machine learning algorithm to keep learning disappears if we use only that algorithm to drive hiring decisions. The only way to avoid this problem is to on

occasion turn of the algorithm, to not use it to hire, in order to see whether candidates that do not fit its criteria continue to perform worse or perhaps perform better.

This problem that selection based on the hiring criteria prevents learning about that criteria holds for any criterion. With the more standard hiring practice of using only a few selection criteria, it is possible to turn them off one-at-a-time to see the effect, for example, of recruiting from a different set of schools. An algorithm generated by machine learning operates as one entity rolling many variables together into an overall model. As a result, it is much more difficult to turn off just one criterion.

Selection can also induce a type of spurious relationship among workers' characteristics called the collider effect in epidemiology and again in data science (Pearl 2018). It occurs when samples are selected in ways that restrict the range of the variables, sometimes known as "range restriction" in psychology. An employer who selects new hires based on college grades and conscientiousness tests might well find that candidates who have neither good grades nor good scores on conscientious tests are not hired. When the employer looks for a relationship between college grades and conscientiousness among its employees, it finds the relationship is negative, even though in the broader population the relationship is positive.

More generally, this selection process can reduce the range on variables of interest making it more difficult to find true effects. For example, if we only hire candidates with good college grades, it may be difficult to identify a true, positive relationship between grades and job performance because the variance of grades in the sample is too limited to identify that relationship. Range restriction also happens when applicants self-select into a firm's pool of applicants, the first step in the well-known "attraction-selection-attrition" framework (Schneider 1987). Algorithms that are based solely on data from the current workforce create this problem.

Several aspects of the modeling process per se can also be challenging. For instance, there is more than one measure of "fit" with the data. A well-known case of this problem concerned the use of a machine learning algorithm by judges in Broward County, Florida to determine whether a person charged with a crime should be released on bail. The algorithm was trained based on data about whether parolees violated the terms of their parole. The challenge in the data is that the majority of the individuals in

the dataset were white, and so the algorithm was driven largely by information about whites. The algorithm predicted the rate of recidivism correctly at an equal rate for whites and blacks, but when it did not predict accurately, it was far more likely to over predict for blacks than for whites (Spielkamp 2017). The problem is that the algorithm cannot optimize on more than one measure of fit.

Decision-Making stage

There are three main challenges when decision makers try to apply the predictions produced by machine learning. The first concerns fairness and legal issues, the second relates to a lack of explainability of the algorithm, and the third to the question of how employees will react to algorithmic decisions.

Fairness

Within the HR context, there are numerous questions related to fairness. One of the most obvious of these is the recognition that any algorithm is likely to be backward looking. The presence of past discrimination in the data used to build a hiring algorithm, for example, is likely to lead to a model that may disproportionately select on white males. Actions using those algorithms risk reproducing the demographic diversity – or lack thereof - that exists in the historical data. The biased outcomes of the Amazon hiring algorithm noted above was caused by exactly this common problem: because fewer women were hired in the past and because men had higher performance scores, the algorithm was selecting out women – even when sex is not in the candidate dataset, selecting out attributes of women, such as taking "women's studies" courses.

In the HR context, there is a wide-spread belief that evaluations of candidates and employees are shaped heavily by the biases of the evaluator, most commonly as related to demographics. Algorithms can reduce that bias by standardizing the application of criteria to outcomes and by removing information that is irrelevant to performance but that might influence hiring manager decisions, such as the race and sex of candidates (Cowgill 2018). Factors that may seem inappropriate may nonetheless improve the predictive power of the algorithms, such as the social status of one's alma mater. How we balance the trade-off between appropriateness and predictive power is not clear.

The fact that employment decisions are so important to individual candidates/employees and to broader society has led to an extensive legal framework designed to guide those decisions. The vast majority of individuals in the US labor force – everyone other than white men under age 40 who do not have disabilities or relevant medical conditions – are protected against discrimination in any employment decision. Other countries have similar rules. Discrimination means adverse actions taken based on one's demographic attributes, and in practice that is measured by "adverse impact," evidence that any employer's decisions have a lower incidence of good outcomes (e.g., hires and promotions) and/or a higher incidence of bad outcomes (e.g., dismissals) than the base rate we would expect from their distribution in the relevant population (see Walsh 2013 Part II for details on the relevant US legal requirements).

With respect to the actions that could be based on algorithms, in other words, those that attempt to predict future outcomes, the only defense against evidence of adverse impact is first to show that the decisions taken actually do predict the desired outcomes and second to show that no other process for making decisions would produce at least as accurate predictions with less adverse impact.

These legal constraints raise considerable challenges for algorithmic based employment decisions. The first is simply that in order to assess whether they have an adverse impact, we have to identify the relationships within the algorithm between any of the attributes of protected groups and the relevant outcomes: Does it give women a lower score, for example, or does it give lower scores to attributes disproportionately associated with women? This is a considerable analytic task for most algorithms.

Letting supervisors make employment decisions without guidance may well lead to far more bias and possibly more adverse impact than the algorithms generate. But that bias is much harder to hold accountable because it is unsystematic and specific to each hiring manager. Algorithms used across the entire organization may have less bias than relying on disparate supervisors, but bias that does result is easier to identify and affects entire classes of individuals. All of this makes it much easier to challenge hiring decisions based on algorithms. Will employers find it worthwhile to take on greater legal risk in order to reduce total bias?_ How will the courts consider evidence concerning algorithms in these decisions? So far, we have no record on these issues.

If we return to the parole violation example above, it would seem that a better approach to building an algorithm to predict parole violations would be to generate a separate one for blacks and for whites. In the context of HR decisions, that might seem appealing as well, to generate separate hiring algorithms, for example, for men and women. While there may be challenges in using such algorithms (e.g., how do we compare the scores of these two different models), the legal frameworks will not allow us to treat these demographic groups differently.

These examples raise the more general concern about fundamental tradeoffs between accuracy and fairness that must be confronted in any HR machine learning implementation (Loftus et. al. 2018). Consider how the role of context changes our judgments. Most of the participants at our workshop, for example, found it perfectly acceptable to use algorithms to make decisions that essentially reward employees – who to promote, who to hire in the first place. But what about the inevitable use of algorithms to punish employees? An algorithm that predicts future contributions will most certainly be introduced at some point to make layoff decisions. How about one that predicts who will steal from the company or commit a crime?

Here we face a dilemma. The Utilitarian notion of advancing the collective good might well argue for using predictive algorithms to weed out problems and costly employees. When the goal is simply optimizing the success of a business, making decisions about employees based on the probability of their actions seems sensible. The Kantian deontological position, on the other hand, suggests that individuals should be judged based on their own actions. Western societies and their legal systems all value this approach. It can be highly objectionable using this framework to make decisions that reward or punish individuals based on attributes that are only associated with desired behaviors especially if those attributes are simply probabilistic forecasts of future behavior. Firing an employee because they have attributes associated with those who have embezzled in the past, for example, would generally be seen as objectionable.

We see two approaches that can make progress on at least some of the above issues. The first and arguably most comprehensive approach is causal discovery, that is, identifying in the data those variables that cause the outcome of interest, such as good job performance. This is a fundamental distinction between data science as it is most often

applied to generating algorithms that are valued principally for their predictive accuracy and conventional statistics.

Consider the question as to whether the social status of an applicant's alma mater predicts their job performance if they were hired. From the perspective of generating algorithms, it is enough if the social status measure contributes to the overall accuracy of an algorithm predicting job performance. Traditional statistics, on the other hand, might ask whether the relationship between social status and job performance is true on its own – not just as part of a more complex algorithm – and whether it was causal. Establishing causation is a much more difficult exercise.

Demonstrably causal algorithms are more defendable in the court of law and thus address at least some legal constraints discussed above. They are fairer due to the explicit specification of causal paths from socio-demographic characteristics to performance, which allows individuals to be acknowledged for their performance enhancing characteristics (e.g., grit or intrinsic motivation) independently of group membership (e.g., the alma mater status) and to intervene in order to compensate for their socio-demographic disadvantages (e.g., to create a strong support network that graduates from top schools get by default). As a result, employees "minimize or eliminate the causal dependence on factors outside an individual's control, such as their perceived race or where they were born" (Loftus et. al. 2018: 7) and thus are treated as individuals rather than group members. Individual fairness, in this case, replaces group fairness.

Computer algorithms can assist in causal discovery by searching for causal diagrams that fit the available data. Such algorithms are being actively developed; their interpretation does not require advanced training but does require data about possible causes and their confounders (Malinsky and Dansk 2017). When data are incomplete, one can test for the causality of specific factors with randomized field experiments. This is one reason why randomization is our second approach to addressing fairness and other challenges to AI in HR management.

Instead of boosting the low predictive power of many HR algorithms with non-causal covariates, which exacerbate unfairness, we propose to accept that HR outcomes are often random (Denrell, Fang, and Liu 2015; Liu and Denrell 2018). Cowgill (2018) shows that noise and inconsistency in human decision-making regarding HR creates

quasi-experimental variation, which is complementary with machine learning in the sense that can be used to de-bias algorithms, if good outcome measures are present.

Moreover, research shows that employees perceive random processes as fair in determining complex and thus uncertain outcomes (Lind and Van den Bos 2002). "Flipping a coin" has a long history as a device for settling disputes, from ties in election outcomes to allocating fishing rights (see Stone 2011). Introducing andomization is especially attractive where there are "losers" in the outcomes and where they remain in the organization or relationship, such as employees who are not selected for promotion. Telling them that the decision literally was made on a coin toss is much easier to bear than either telling them it was a close choice (you were almost as good, on the one hand, but something small could have changed the outcome) or that it was not close (you were not almost as good, but there is nothing you could have done that would have mattered).

It might also be helpful to introduce something less than complete randomness to the process to help with its acceptability. For example, when predictive scores are not tied but are merely close, we might introduce a weighted random aspect where the candidate with the higher score gets a proportionately greater chance.

Explainability

Closely related to the notion of fairness is explainability, in this case the extent to which employees understand the criteria used for data analytic-based decisions. A simple seniority decision rule – more senior workers get preference over less senior ones – is easy to understand and feels objective even if we do not always like its implications. A machine learning algorithm based on a weighted combination of 10 performance-related factors is much more difficult to understand, especially when employees make inevitable comparisons with each other and cannot see the basis of different outcomes. (Professors who have to explain to students why their grade is different than that of their friend who they believe wrote a similar answer are familiar with this problem.) Algorithms get more accurate the more complicated they are, but they also become more difficult to understand and explain.

A well-known example of the importance of explainability to users comes from the Oncology application of IBM Watson. This application met considerable resistance

from oncologists because it was difficult to understand how the system was arriving at its decisions. When the application disagreed with the doctor's assessment, this lack of transparency made it difficult for medical experts to accept and act upon the recommendations that the system produced (Bloomberg 2018). Especially in "high stakes" contexts, such as those that affect people's lives—or their careers--explainability is likely to become imperative for the successful use of machine learning technologies. We expect major progress in this area in the coming years, due to a wave of investment from the commercial and government sectors geared towards explainable AI. For instance, the US Defense Advanced Research Projects Agency (DARPA), known for its successful funding of path-breaking research in IT, has just launched a major initiative on explainable artificial intelligence (XAI) with deliverables, software toolkits and computational models, expected by 2021 (https://www.darpa.mil/program/explainable-artificial-intelligence).

Employee reactions to algorithmic decisions

Changes in formal decision-making of the kind associated with the introduction of algorithms unavoidably affect employees' experiences and behavior. In this regard, we can learn a great deal from Scientific Management's efforts to develop optimal decision rules. Employment practices and decisions about work organization were based on a priori engineering principles and human experiments. Although they may have been much more efficient than previous practices, they were bitterly resented by workers, leading to a generation of strife and conflict between workers and management. From the perspective of front-line workers and their supervisors, the situation may have looked very similar to the AI model we outline here: decisions would been handed down from another department in the organization, the justification for them would be that they were the most efficient that science could provide, understanding the basis of the decision is extremely difficult, and trying to alter them would simply be a mistake.

To illustrate, it is widely believed that the relationship with one's supervisor is crucial to the performance of their subordinates and that the quality of that relationship depends on social exchange: "I as supervisor look after you, and you as subordinate perform your job well." Even when employees have little commitment to their employer

as an organization, they may feel commitment to their supervisor. How is this exchange affected when decisions that had been made by the supervisor are now made by or even largely informed by an algorithm rather than a supervisor?

If my supervisor assigns me to work another weekend this month, something I very much do not want to do, I might do it without complaint if I think my supervisor has otherwise been fair to me. I might even empathize with the bind my supervisor is in when having to fill the weekend shift. If not, I might well go complain to her and expect some better treatment in the future. When my work schedule is generated by software, on the other hand, I have no good will built up with that program, and I cannot empathize with it. Nor can I complain to it, and I may well feel that I will not catch a break in scheduling in the future. We know, for example, that people respond very differently to decisions that are made by algorithms than decisions made by people (Dietvorst, Simmons, and Massey 2016). If there is good news to give me, such as a bonus, it builds a relationship with my supervisor if she appears to have at least been involved in the decision, something that does not happen if that decision is generated by an algorithm.

Yet, there may be occasions where decisions are easier to accept when made by an algorithm than when made by a human, especially when those decisions have negative consequences for us. Uber riders, for example, respond negatively to surge pricing increases when they perceive that they are set by a human (trying to exploit them) as opposed to by an algorithm. Experimental evidence suggests that willingness to accept and use algorithms depends in part on how they update to deal with mistakes (Dietvorst et al forthcoming).

Related to these issues is the engagement in decisions that individuals have that is otherwise lost with the shift to algorithms. Research increasingly shows that algorithms perform better than human judgment when used to predict repetitive outcomes, such as reading x-rays and predicting outcomes about employees or job candidates (Cowgill 2018). But if algorithms take over hiring and supervisors play no role in the process, will they be as committed to the new hires as if they had made the hiring decisions?

Discussion and Conclusions

While general-purpose AI is still a long shot in any domain of human activity, the speed of progress towards specialized AI systems in health care, automobile industry, social media, advertising and marketing is considerable. Far less progress has been made in issues around the management of employees even on the first step of the AI path, which is decisions guided by algorithms. We identify four reasons why: complexity of HR phenomena, data challenges from HR operations, fairness and legal constraints, and employee reactions to AI-management.

Causal reasoning is the first principle relevant to addressing these challenges across the stages of the AI Life Cycle. Because the creation of algorithms relies on association rather than causation, an absence of notions of causation makes it much more difficult to create the datasets needed for analysis: we need more data because we do not know what to choose. Causal reasoning also helps greatly with issues of fairness and explainability. The benefits of causal reasoning do come with costs. Employers must first accept the greater costs (based on the need for more data) and lower predictive power from algorithms where we do not have causal models, and they must work to develop consensus about causal assumptions in advance of modeling. These challenges explain why the data science community is quite skeptical about causally reasoning AI systems.

Randomization is a second principle that can help with algorithmic-based decisions. First, randomizing the inputs into an algorithm is akin to experimentation and can help to establish causality. Second, randomly choosing an HR outcome with the probability predicted by an algorithm where we cannot predict outcomes with much accuracy acknowledges the inherently stochastic nature of HR outcomes and unavoidable inaccuracy of algorithms. Employees may perceive such randomization—such as flipping a coin—to produce fairer outcomes under uncertainty.

Formalizing processes is also necessary to build reasonable algorithms. It ensures that the parties are aware of the assumptions built into any algorithms, the costs of building them, and the likely challenges from employees who are adversely affected by them. In the process, formalization can be enabling rather than coercive (Adler and Borys 1996).

To what extent the changes we suggest require a restructuring of the HR function is an important question. Certainly, HR leaders need to understand and facilitate the Data Generation and Machine Learning stages of the AI Life Cycle. The integration of HR data with business and financial data should allow an HR Department to quantify in monetary terms its contribution to the company's bottom-line.

Line managers will have to refresh their skill set as well. For them, AI should imply "augmented intelligence," an informed use of workforce analytics' insights in decision-making. The literature on evidence-based management proposes a Bayesian approach to systematically updating managerial beliefs with new information (Barends and Rousseau 2018). We consider it a helpful departure point for AI-management as well.

The tension between the logic of efficiency and of appropriateness affects most organizational action (March and Simon 1993). In the case of HR, the drive for efficiency and concerns about fairness do not always align. We hope that the conceptual and practical insights in this paper will move AI-management in HR forward on both counts, those of efficiency and appropriateness.

Bibliography

Acktar, Reese, Dave Winsborough , Uri Ort , Abigail Johnson , Tomas Chamorro-Premuzic. 2018. Detercting the Dark Side of Personality Using Social Media. Personality and Individual Differences, 132:90-97.

Adler, Paul and Bryan Boris. 1996. "Two Types of Bureaucracy: Enabling and Coercive." *Administrative Science Quarterly* 41: 61-89.

Barends, Eric and Denise M. Rousseau. 2018. *Evidence-Based Management: How to Use Evidence to Make Better Organizational Decisions.* Kogan Page.

Bloomberg, J. 2018. Don't Trust Artificial Intelligence? Time to Open the AI Black Box. *Forbes*. Last accessed at https://www.forbes.com/sites/jasonbloomberg/2018/09/16/dont-trust-artificial-intelligence-time-to-open-the-ai-black-box/#577a14153b4a on Nov 27, 2018.

Bock, Laslo. 2015. *Work Rules! Insights from Inside Google That Will Transform How You Live and Lead.* Hachette Book Group.

Cappelli, Peter. 2017. “There’s No Such Thing as Big Data in HR.” *Harvard Business Review*. June.

Cappelli, Peter and AnnaTavis. 2017. The Performance Management Revolution. *Harvard Business Review,* November.

Carrillo-Tudela, C., Hobijin, B., Perkowski, P., and Visschers, L. 2015.

Denrell, Jerker, Christina Fang, Chengwei Liu. 2015. “Change Explanations in Management Science.” *Organization Science* 26(3): 923-940.

IBM 2018.

Cowgill, Bo (2017) The Labor Market Effects of Hiring through Machine Learning Working Paperowgill, Bo. 2018. “Bias and Productivity in Humans and Algorithms. Theory and Evidence from Résumé Screening.” Working paper.

Dietvorst, Berkeley, Simmons, Joseph P. ,and Massey, Cade, Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err (July 6, 2014). Forthcoming in *Journal of Experimental Psychology: General.*

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2016). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155-1170.

Junqué de Fortuny, E., Martens, D., & Provost, F. (2013). “Predictive modeling with big data: is bigger really better?” *Big Data*, *1*(4), 215-226.

Lee, M. K., Kusbit, D., Metsky, E., & Dabbish, L. 2015. *Working with machines: The impact of algorithmic, data-driven management on human workers*. Proceedings of the 33rd Annual ACM SIGCHI Conference: 1603-1612. Begole, B., Kim, J., Inkpen, K & Wood, W (Eds.), New York, NY: ACM Press.

Lind, E. Allan and Kees Van den Bos. 2002. “When Fairness Works: Toward a General Theory of Uncertainty Management.” *Research in Organizational Behavior* 24: 181-223.

LinkedIn. 2018. The Rise of HR Analytics.

Liu, Chengwei and Jerker Denrell. 2018. “Performance Persistence Through the Lens of Chance Models: When Strong Effects of Regression to the Mean Lead to Non-Monotonic Performance Associations.” Working paper.

Loftus, Joshua R., Chris Russel, Matt J. Kusner, and Ricardo Silva. "Causal Reasoning for Algorithmic Fareness." **arXiv:1805.05859**

Malinsky, Daniel and David Danks. 2017. "Causal Discovery Algorithms: a Practical Guide." *Philosophy Compass* https://doi.org/10.1111/phc3.12470.

Meyer, David. 2018. *Amazon Reportedly Killed an AI Recruitment System Because It Couldn't Stop the Tool from Discriminating Against Women.* Fortune. October 10th. http://fortune.com/2018/10/10/amazon-ai-recruitment-bias-women-sexist/

March, James and Herbert Simon. 1993. *Organizations.* Oxford: Blackwell.

Monthly Labor Review. 2017. Estimating the U.S. Labor Share. Bureau of Labor Statistics, February. https://www.bls.gov/opub/mlr/2017/article/estimating-the-us-labor-share.htm

Netessine, Serguei and Valery Yakubovich. 2012. "The Darwinian Workplace." *Harvard Business Review,* 90(5): 25-28.

Pearl, Judea. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Pfeffer, Jeffrey and Robert I. Sutton. 2006. *Hard Facts, Dangerous Half-Truths and Total Nonsense: Profiting from Evidence-Based Management.* Harvard Business Review Press.

Rousseau, Denise (Editor). 2014. *The Oxford Handbook of Evidence-Based Management.* Oxford University Press.

Schneider, Benjamin. 1987. *The People Make the Place.* Personnel Psychology. 40: 437-453.

Spielkamp, Michael. 2017. Inspecting Algorithms for Bias. MIT Technology Review. June 12. https://www.technologyreview.com/s/607955/inspecting-algorithms-for-bias/

Srivastava, Sameer and Amir Goldberg. 2017. "Language as a Window into Culture." *California Management Review* 60(1): 56-69.

Stone, Peter. 2011 *The Luck of the Draw: The Role of Lotteries in Decision Making.* Oxford: Oxford University Press.

Tucker, Catherine. 2017. "Privacy, Algorithms, and Artificial Intelligence." In *The Economics of Artificial Intelligence: An Agenda.* Edited by Ajay K. Agrawal,

Joshua Gans, and Avi Goldfarb. University of Chicago Press. Forthcoming. http://www.nber.org/chapters/c14011.

Walsh, David. 2017. *Employment Law for Human Resource Practice.* Mason, Ohio: South-Western.