BerkeleyHaas
Haas School of Business
University of California Berkeley

Artificial Intelligence in Human Resources Management:

Challenges and a Path Forward

California Management Review
1–28

DOI: 10.1177/0008125619867910
journals.sagepub.com/home/cmr
SAGE

Prasanna Tambe[1], Peter Cappelli[1,2], and Valery Yakubovich[1,3]

SUMMARY

There is a substantial gap between the promise and reality of artificial intelligence in human resource (HR) management. This article identifies four challenges in using data science techniques for HR tasks: complexity of HR phenomena, constraints imposed by small data sets, accountability questions associated with fairness and other ethical and legal constraints, and possible adverse employee reactions to management decisions via data-based algorithms. It then proposes practical responses to these challenges based on three overlapping principles—causal reasoning, randomization and experiments, and employee contribution—that would be both economically efficient and socially appropriate for using data science in the management of employees.

KEYWORDS: data analysis, human capital, human resource ethics, hiring and recruitment, information systems, decision-making tools

The speed with which the business rhetoric in management moved from big data to machine learning 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% of CEOs report that they are not at all prepared to make use of new data analytic tools, and only 4% say that they are "to a large extent" prepared.[1]

[1]Wharton School, Philadelphia, PA, USA
[2]The National Bureau of Economic Research, Cambridge, MA, USA
[3]ESSEC Business School, Cergy, France

AI conventionally refers to a broad class of technologies that allow a computer to perform tasks that normally require human cognition, including adaptive decision making. Our discussion here is narrower, focusing on a subclass of algorithms within AI in use now that rely principally on the increased availability of data for prediction tasks. There have been major advances in some AI applications, such as pattern recognition and language translation, and deep learning using neural networks in some data-rich contexts that has brought us closer to true AI. 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% of firms say they have adopted analytics in human resources (HR),[2] and how sophisticated the analytics are in those firms is not at all clear.

The effective application of AI to HR problems presents different challenges than it does in other areas. They range from practical to conceptual, including the fact that data science analyses—when applied to decisions about people—can create serious conflicts with what society typically sees as important for making consequential decisions about individuals.

To illustrate some of 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 the relationship between those attributes and job performance. If we found a causal relationship between an attribute (such as sex and job performance), we might 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 and of our data 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. It had been built on historical job performance data, when white men had been the best performers (indeed white men were most of the employees), and the algorithm gave higher scores to white male applicants as a result. When the sex of applicants was not included as a measure, attributes associated with women candidates, such as courses in "Women's Studies," caused them to be ruled out. The company soon stopped using the system as there was no simple way to fix it.[3]

When we build an algorithm on a more objective measure, such as who steals from the company, the number of such cases in a typical company is likely to be too small to construct an effective algorithm. With a task such as hiring, once applicants discover the content of our hiring algorithm, they are likely to adjust their behavior to it 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 judged as negative, such as, "I work too hard."

We address challenges like these as they play out in what we call the AI Life Cycle: Operations, Data Generation, Machine Learning, and Decision Making. AI algorithms can respond to those challenges using the approaches of contemporary data science as an alternative to managerial judgment. We bring together key ideas from Evidence-Based Management (EBMgmt)—a theory-driven analysis of "small data"[4] and out-of-the-mainstream approaches to machine learning[5]—in order to position causation as central to all four challenges we identified. We also suggest that randomization can be a useful component of an AI-augmented decision process, given that it is already present in many managers' decisions,[6] it is often perceived as fair,[7] and algorithms may otherwise struggle to make fair and valid decisions.[8]

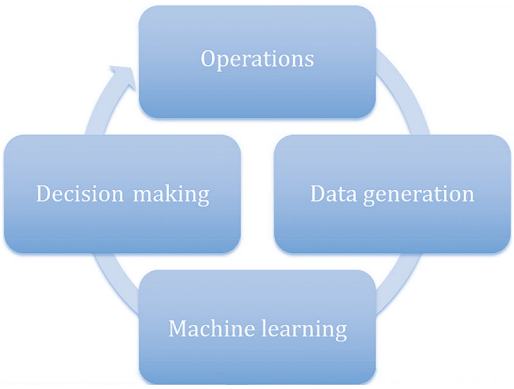
To bridge the state-of-the-art in data science with the needs of HR practice, we brought data science faculty together with the heads of the workforce analytics function from 20 major U.S. corporations (known for their sophisticated management systems) for a one-day workshop in the fall of 2018. Prior to the workshop, we circulated a short survey with open-ended questions about their corporations' ongoing initiatives regarding analytics and algorithmic decision making, barriers they face, and breakthroughs they expect. The workshop itself consisted of four sessions on the topics of data management, social media as a source of HR data, fairness and ethics of HR decisions, and employee recommendations. Each session included a presentation by a data scientist followed by an open discussion. Our practitioners' examples and comments from the survey and workshop may not be representative of business at large. Nevertheless, they were helpful for informing our thinking and for articulating the four challenges.

HR Challenges and the AI Life Cycle

There are several issues in HR that differentiate it from many other areas where AI techniques have been applied. The first is the complexity of HR outcomes. Consider, for example, what it means to be 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,[9] and many employers are giving them up altogether.[10] Any reasonably complex job is interdependent with other jobs, and therefore individual performance is hard to disentangle from group performance. A vast literature documents numerous problems with existing performance systems as well as our field's failure to establish a clear link between individual, team, and organizational performance.[11] Given the uncertain quality of performance evaluations by humans, can we use them for training AI algorithms? Doing so might well mean scaling up arbitrary or outright discriminatory human decisions.

A second problem for data science is that many of the important outcomes in HR, such as dismissals, are relatively rare events, especially in smaller organizations. Machine learning and other data science techniques require large numbers

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



Note: AI = artificial intelligence; HR = human resource.

of observations.[12] Data science techniques perform poorly when predicting relatively rare outcomes.[13]

A third issue is that the outcomes of HR decisions, such as who gets hired and fired, have serious consequences for individuals and society with regard to ethics as well as to both procedural and distributive justice fairness. Elaborate legal frameworks also hold employers accountable for making those decisions in a fair manner. Central to those frameworks is the concern with "explainability," knowing what attributes are driving the decision. This is something that is typically absent from methods underlying many state-of-the-art prediction algorithms.

Employment actions are 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. These affect organizational outcomes as well as individual ones. When employees do not understand or accept how decisions are made, they are capable of gaming the system or disrupting it in ways that affect organizational outcomes. While a human decision maker can monitor adversarial behavior and adjust his or her decisions accordingly, even state-of-the-art algorithms find this to be a challenging problem. Dealing with manipulation of this type is the focus of a machine learning technique known as "adversarial machine learning."

Keeping these challenges in mind, we turn now to a taxonomy for organizing the separate tasks that are involved in making use of data science. Figure 1 depicts a conventional AI Life Cycle: operations, data generation, machine learning, and decision making.

Operations are the tasks of HR, such as how an organization hires employees. HR performs a great many tasks involving considerable amount of money, which makes it an attractive target for improvement in processes. In the U.S.

TABLE 1. The HR Life Cycle.

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)?

Note: HR = human resource.

economy as a whole, roughly 60% of all spending is on labor. In service industries, the figure is much higher.[14] Table 1 lists the most common tasks in HR with the corresponding prediction tasks they raise for workforce analytics. They correspond to the "Human Resources Life Cycle," which is commonly used to organize HR tasks.[15]

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. These operations produce volumes of data in the form of texts, recordings, and other artifacts. As operations move to the virtual space, some of those data are in the form of "digital exhaust," or trace data on digital activities, such as how job applicants navigate an employer's website.

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

that these database management tasks are a fundamental challenge in analyzing HR practices and outcomes.

Machine learning refers to a broad set of techniques that learn from data to create algorithms, typically to predict outcomes. Within business contexts, the most common application of machine learning technologies has been "supervised application" in which a data scientist "trains" a machine learning algorithm on a subset of the relevant data and determines the most appropriate metric to assess the performance of the algorithm that it produces. Some of the most commonly used prediction algorithms, such as "logistic regression," infer the outcome variable of interest from statistical correlations among observed variables.[16]

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. For current employees, algorithms are principally used to make recommendations to employees about actions they may take. 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. The extension of such recommendations into wellness programs is already underway, in some cases collecting data about employees' health and wellness directly with devices like "Fitbits," urging employees to adopt practices that lead to better health outcomes, and sometimes rewarding them with payments or punishing them with higher health care costs based on their compliance.

"Algorithmic management" is the name for the practice of using algorithms to guide incentives and other tools for "nudging" contractors in the direction of the contractee.[17] These are also being applied to regular employees now.[18]

These algorithms differ in important ways from traditional approaches used in HR. In industrial psychology, the field that historically focused the most attention on HR decisions—say, research on hiring—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, typically one test at a time, for example, the relationship between personality test scores and job performance, then the relationship between education and job performance, and so forth. The result would be conclusions about several variables that might be used to predict hiring success.

Machine learning, in contrast, uses many variables to generate one algorithm and typically one score to assess a candidate. The variables used may not be in the cannon of the theoretical literature associated with the topic, and the

researcher is not hypothesizing or even examining the relationship between any one variable and the outcome being predicted. Indeed, one of the attractions of machine learning is its investigation of non-traditional factors because the goal is to build a better prediction rather than advancing the theory of a given field.

If hiring is the most important topic for data analysis, the second most popular may be to predict turnover. Vendors such as Jobvite generate machine learning algorithms that score individual employees based on social media posts; others use simpler data like the extent to which individuals have updated their LinkedIn profiles. Many of the companies at our conference were developing their own proprietary algorithms to predict flight risk.

IBM's Blue Match software uses algorithms for another HR task, to drive career advancement by suggesting career advancement moves and new jobs for employees. The algorithms are based on employee interests and prior jobs, training, and ultimately the characteristics of individuals who have succeeded in those jobs in the past; 27% of the company's employees who changed jobs in 2018 did so based on recommendations from the company's Blue Match software.[19]

The move away from checklist-based performance appraisals and toward continuous discussions (facilitated by phone-based apps) has been aided by natural language processing software from vendors such as WorkCompass. These systems read through a year's worth of text messages to produce summaries of the issues discussed and comparisons with other employees, among other things, to drive merit pay decisions.

Decision making, the final stage in the life cycle, deals with the way in which we use insights from the machine learning model in everyday operations. In the area of HR decisions, employers may rely completely on the scores from algorithms to make decisions, or they may give individual managers' discretion as to how to use it.

Addressing AI Challenges: One Stage at a Time

In the following, we explore in detail the four general challenges HR poses for AI: complexity of HR phenomena, small data, ethical and legal constraints, and employee reactions to AI management. We do so 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. As noted above, 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.[20] Moreover, complex jobs are interdependent with one another, and thus, one employee's performance is often inextricable from the performance of the group:

is it sufficient to be a good individual contributor, and if not, how do we measure interactions with others? Multiple measures of performance might make sense, but it is not possible to optimize across several different outcome measures. The fact that individual measures of performance are at best incomplete and at worst biased is a significant drawback to a great many HR operations and, in turn, to using data analysis to improve them.

The participants of our workshop indicated that not all of the attributes of HR actions that we imagine are actually measured; not all details of operations leave digital traces that can be captured, and not all traces left can be extracted and converted to a usable format at a reasonable cost. For example, employers do not necessarily track the channels through which applicants come to them—from referrals versus visiting our website versus job boards, and so forth—which is a reasonably simple exercise to do. Most employers collect only a limited amount of data on applicants before ruling them out, and they do not retain information on the applicants they screen out. These choices limit the types of analyses that can be performed and the conclusions that can be drawn.

The fact that there is no list of "standard" variables that employers are required to gather and retain through their HR operations, as there might be in fields like accounting, 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 the details that employers have on the costs of employees differ 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:

- Perfect measures of performance do not exist. It is better to choose reasonable measures (e.g., would you have hired this new employee if you could go back?) and stick with them to see patterns and changes in results than to keep tinkering with systems to find the perfect measure. Most of our data analytics efforts in HR are based on decisions concerning individual employees—who to hire, who to retain, and what to recommend about training and advancement. Without reasonable measures of performance, none of these analytics efforts will be useful.
- 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 that 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.
- Include business and financial performance data at the organizational level closest to employee control to have the best chance of seeing how individual performance affects larger business units and the company as a whole.

- Aggregate information from multiple perspectives and over time. Digital HR tools allow for rapid real-time evaluations among colleagues using mobile devices, for example. Machine learning algorithms are ideal for making sense of such information.

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 because it is the best at that task, a system from another vendor for applicant tracking software because it is the best at that task, from a third for compensation and payroll data, and so forth. Because the systems are from different vendors and are typically based on different technology architectures, they are rarely compatible. It was surprising to hear from our respondents about the internal political battles over control over data. The payroll department, for example, does not want to give its data to the talent acquisition department to let them see what predicts which applicants are likely to take the most time off. These conflicts are clearly not unique to data analysis, but it is an important reminder that the organizational conflicts do not go away with the use of new tools like data analytics.

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, manage, and analyze their data was Excel. Very few used more purpose-built tools such as Tableau that are common in data analytics. Software for bridging data sets as well as "data lakes" that archive and access different data sets represent a way forward, but they can be difficult to integrate, can be viewed as confining, and face their own limitations. They remain underused 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. At present, the types of data needed to do even the most basic analyses—such as seeing whether certain hiring decisions lead to better new employees—often cannot easily be done because the components of data needed for the analysis are owned by different parts of the organization. Only higher level executives can drive the cooperation across units that is needed for data analysis to begin. One of our responding companies reported a potential solution to some of the data incompatibility issues in the form of a vendor selection committee with a representative from each HR department. The committee had to vote to approve the vendor selection of any individual department, and data compatibility was an important criterion in those votes.

Given these data access challenges, it can be extremely difficult and 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 in our discussion with data analytics practitioners in HR. Beyond the obvious criteria of the cost of undertaking any analysis is the likelihood that it will generate usable 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 in understanding a particular problem. Some go to senior management and solicit for help with the type of problems that prevent them 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 they believed that anyone was willing to act on results that they might find.

A more systematic way to select the questions to investigate should include examining the research literature in order to establish what is already known about different research questions, as EBMgmt has long advocated.[21] 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 information technology (IT) companies, such as Amazon, Google, Facebook, and Microsoft, hire as many PhDs in social sciences as in data sciences into the HR department to help close this disconnect.

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.

Small data are a fundamental concern for HR 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 if they do not have many employees. The machine learning literature has shown that access to larger data has substantial advantages in terms of predictive accuracy.

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. On the contrary, some aspects of HR may generate millions of observations, such as continuous measures of employee performance: it is straightforward, for example, to monitor employee time spent doing online work and not working; call center employees are assessed on each call with many metrics; employees performing simple physical tasks, such as sorting packages, are measured per hand movement.[22]

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. The management literature has an important advantage over data science in articulating theory as well as a long history of empirical findings. That literature also specifies causal relationships, as opposed to prediction from correlations among observed variables in machine learning. Recently, voices in the computer science community have articulated the need for causation as critical for the future of AI in human affairs.[23]

Using AI in HR should require that managers put their assumptions on the table as they are then built into the models and analysis. If we include gender in the data used to predict voluntary turnover, for example, why are we doing so? We hope that decision makers will persuade other stakeholders of the validity of their assumptions, ultimately by using data and empirical analysis. The formulation of such assumptions often turns into a contest among stakeholders with different views. It is common, for example, to have some stakeholders who assume that employees are rational decision makers and others who see them quite differently. Formalizing the process of creating the underpinnings of these models is important: these are not decisions that should be made by data scientists alone.

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.[24] (Off-the-record conversations also suggest that Google leadership did not accept the research finding that unstructured interviews were poor predictors of good hires—having already committed to that practice—and so they conducted research that confirmed it was true, even at Google.) If there is no consensus on the causal model being examined, AI analyses are likely to be counterproductive.

One attraction of using vendors is their ability to combine data from many employers and the ability to use big data tools to generate their algorithms. Vendors have long used this approach 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. Such evidence is essential to address legal concerns.

Employers are also concerned about employees' tendency to bias their responses and the data depending on how they think the data are used. Candidates

have never been completely forthcoming in job interviews, for example. Because of this, a great many employers now make use of social media information precisely because they believe employees are being more authentic in it.[25] Those data are now 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 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, although nothing prevents data like these from being used for other purposes as well, such as who is resisting change efforts.

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. Of course, there are important concerns about privacy and the negative effects on employee attitudes and behavior when they perceive that their privacy is being violated.

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

Many of these and similar considerations fall under the purview of privacy. Issues associated with electronic monitoring of employee performance and privacy are not new,[26] but the contemporary context of social media in particular creates new challenges: data can persist well beyond the time it was generated and employers can repurpose it for use unanticipated by the creator, for example, the words from an email exchange with a colleague might be used to predict flight risk. The data of one person may also inadvertently affect other people, such as when a creator's friends are 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 European Union's (EU) 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.[27] Among the novel suggestions in this area are that the Genetic Information Nondiscrimination Act be used as a model for protecting employees from their employer's breach of privacy.[28]

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 are randomized during the collection process, which leads to "learning nothing about an individual while learning useful information about the population."[29] 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 those in email traffic or on social media, is that it is not clear how "authentic" it really is. Social media posts 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 issue of individuals and especially job applicants altering their responses to what they believe assessments want is not new.[30] In the case of social media data, the nature of what employees post will no doubt change 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 are yet another effort to obtain authentic data where the employees do not necessarily know how to spin their responses. However, job applicants are already getting help from businesses like the JobTestPrep company who figure out how to score well on Lloyds's selection game.[31] Getting authentic data on applicants will remain a challenge because of the ability of candidates to game such efforts.

Machine Learning Stage

A machine learning 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 HR is not making much progress to help employers: the fact that most of the predictors advocated in that research on a topic like hiring predicts so little of job performance that it creates an enormous opportunity 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 analyses, such as identifying relationships with one selection test, nor is it constrained by prior research findings.

Bo Cowgill, for example, shows how a machine learning algorithm can do better than humans. In a field experiment with hiring white-collar workers, he finds that AI can remove human biases if the training data are sufficiently noisy: inconsistent human decision making introduces quasi-experimental variation, which improves machine learning to such a degree that it yields better candidates than assessments by HR recruiters. Specifically, the candidates selected by the machine are 14% more likely to pass interviews and receive a job offer, are 18% more likely to accept an extended job offer, are 0.2-0.4 standard deviation more

productive once hired, and are 12% less likely to show evidence of competing job offers during salary negotiations.

An important and surprising conclusion from this study is that the better performance of the algorithm was due to noisy data that came from hiring "non-traditional" candidates that might typically not make it through a hiring process: from non-elite colleges, lacking job referrals and prior experience, with atypical credentials but strong non-cognitive soft skills, and so forth. These are remarkable counterintuitive findings that attest to the potential of AI.[32] They also make the important point that using restrictive hiring criteria—for example, we only hire finance majors from elite colleges—will make it impossible for a hiring algorithm to ever improve on its predictive power.

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. Doing so "selects on the dependent variable"—by looking only at best performers, we cannot know which, if any, attributes differentiate them from bad performers. Then applicants are assessed against that algorithm. Consider, for example, vendors such as HireVue that help clients conduct video interviews. Part of their offerings includes algorithms based on facial expressions captured on those videos. These algorithms are sometimes 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? 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 an effective 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 facial expressions in response to job 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 off 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 solely on the hiring criterion creates holds for any hiring 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 examine the effects of any one criterion.

Selection can also induce a type of spurious relationship among workers' characteristics called the collider effect in epidemiology and in data science.[33] 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 now 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.[34] Algorithms that are based solely on data from the current workforce create this problem as well.

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 overpredict for blacks than for whites.[35] The problem is that the algorithm cannot optimize on more than one measure of fit. The implications for HR are obvious given that prediction models for hiring or other outcomes may differ by sex, race, and other protected groups.

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. The HR context raises numerous issues where fairness matters. 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 because in the past, white males accounted for most of those rated as high performers. 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 were 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 and those with attributes associated with women.

In the HR context, there is a widespread 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 and that might influence hiring manager decisions, such as the race and sex of candidates. On the contrary, 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 U.S. labor force—everyone other than those under age 40 without disabilities or relevant medical conditions—are protected against discrimination in any employment decision (even white men are protected against employment actions taken on the basis of gender and race). 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.[36]

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 algorithm-based employment decisions. The first is simply that in order to assess whether algorithms 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 can be a considerable analytic task for most algorithms.

Letting supervisors make employment decisions without guidance, on the contrary, 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 because it is easier to identify. 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 experience 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.[37] 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? Such "integrity" tests are already used in the workplace now as part of the hiring process.[38]

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 truly cause the outcome of interest, such as good job performance. This is a fundamental distinction between social science (which rests on causal discovery) and data science (which does not) as the latter is most often valued simply for its predictive accuracy. Contexts where data science was developed, such as predicting when a machine is likely to fail, do not demand causal explanations.

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 contrary, might ask whether the relationship between social status and job performance is true on its own—not just as part of a more complex relationship—and whether it actually caused better job performance. 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 individual characteristics to performance. This allows individuals to be acknowledged for their performance enhancing characteristics (e.g., grit or intrinsic motivation) independent of group membership (e.g., the alma mater status). It also allows policy makers to identify where to intervene to compensate for disadvantages that are perceived to be unfair, for example, to help students at lower status colleges to create the kind of strong social networks of support that graduates from high-status schools get by default. As a result, decision makers can "minimize or eliminate the causal dependence on factors outside an individual's control, such as their perceived race or where they were born."[39] They can be treated more as individuals rather than as members of a particular category or group. Individual fairness, in this case, replaces group fairness.

Computer-based algorithms can actually 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.[40] As noted above, when data are incomplete, one can test for the causality of specific factors in other ways, such as with randomized field experiments.

Instead of boosting the low predictive power of many HR algorithms with non-causal covariates, which exacerbate unfairness, we propose to accept that some HR outcomes are often random, or at least have random aspects to them. One approach, as noted above, is to turn off the algorithms on occasion to allow for variation that the algorithms are otherwise ruling out to de-bias algorithms.[41] If these observations perform well in terms of their later stage outcomes, this information can be fed back to the model to increase the likelihood they get selected in the earlier stage.

Even with good algorithms, the recommendations may not be so clear as to lead to decisions that will be perceived as fair. We may have two candidates with essentially identical scores or similar scores that predict an outcome, such as performance appraisal ratings, that we do not believe are very precise measures. In that case, relying on the algorithm to choose between the candidates leads to a decision that is essentially random.

Research shows that employees understand the random aspect of many outcomes and perceive decisions that are acknowledged to be random as fair in

such contexts.[42] "Flipping a coin" has a long history as a device for settling disputes, from ties in election outcomes to allocating fishing rights.[43] Introducing explicit randomization and acknowledging it in decisions are 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, and something very minor would have changed the outcome.

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. The common use of "cut scores" in tests where we assume that everyone who scored above a stated standard has "passed" and those below "failed" is one example where we might select winners at random from those who passed the standard.

Explainability. Closely related to the notion of fairness is explainability, in this case the extent to which employees actually 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 application of algorithms to oncology by IBM Watson. An algorithm was developed to identify cases of cancer, but it 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.[44] Patients seem to have the same difficulty accepting diagnoses and treatment recommendations generated by algorithms.[45]

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 toward explainable AI. For instance, the U.S. 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.[46]

Back to Operations: 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 workplace decision rules. Employment practices (e.g., how fast to work based on time and motion studies) and decisions about work organization (e.g., breaking down tasks to simple components) 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 frontline workers and their supervisors, the situation may have looked very similar to the AI model we outline here: decisions would have 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 one concern, 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 contrary, I have no goodwill 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.[47] 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 we are more willing to accept decisions from algorithms when we can see how they update to deal with mistakes.[48]

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.[49] 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

In Table 2, the column "Operations" is placed at the end to reflect companies' need to respond to the reality modified by AI algorithms. Three groups of recommendations across the stages of the AI Life Cycle are labeled causal reasoning, randomization and experiments, and employee contribution.

Most machine learning–based algorithms excel in pattern recognition by associations rather than *causation*. However, recognizing images, a common AI task, is not nearly as difficult as recognizing good workers. As noted above, the scope of possible performance indicators is broad and hard to observe and measure precisely. Attempts to dig them out from digital traces of human behavior within and outside organizations run into severe issues of control, privacy, and ethics and still do not guarantee that anything worthwhile will be found. Even if tight associations are found between a set of observable worker characteristics and behaviors within a company, they may not be usable because of legal and fairness concerns. Causal reasoning focuses our attention on the characteristics and behaviors that are relevant, reduces the costs of data management, and goes a long way toward meeting the requirements of fairness and explainability that are central to the future of AI algorithms.

The benefits of causal reasoning do come with costs. Causal models have lower predictive power in comparison with algorithmic associational models, although how much so in the HR area is difficult to know because we have essentially no evidence available on the validity of the algorithms being used now. Causal models require not just data but also expert knowledge of the context. By making causal assumptions explicit, as we suggest, algorithm designers make themselves vulnerable to criticism and organizational politics. We propose two remedies in this regard, methodological and organizational.

Methodologically, causal discovery is a rapidly developing toolkit that automates empirical testing of causal assumptions and thereby narrows down the set

TABLE 2. Possible Responses to Challenges of AI's Introduction in HR Management.

	Response			
Challenge	**Data generation**	**Machine learning**	**Decision making**	**Operations**
Complexity of HR outcomes	Solicit employee contributions into outcomes' metrics and create consensus around them	Train algorithms for a few outcomes	Managers' discretion on the basis of the algorithm's predictions Run experiments whereby an algorithmic or human decision is randomly assigned to individual cases	Monitor the medium- and long-term validity of AI-based decisions Periodically review and retrain the algorithm
Small data	Integrate HR data with financial and operational data Use fine-grained real-time data Use vendors' data collected from larger populations	Use vendor-trained models Use causal models	Let managers act on algorithm's recommendations according to prespecified guidelines	
Accountability regarding fairness, ethical norms, and labor laws	Assess the consistency of human-made decisions used for training the algorithm	Create consensus around fairness criteria Weigh multiple fairness criteria Use causal model Ask data scientists to explain the model (identify the features that disproportionally affect its predictions)	Make random choices with probabilities predicted by the algorithm	Specify a code of ethics for AI-related initiatives Create an AI Council with representatives of all stakeholders
Employee reactions	Collect data to improve processes first	Create employee consensus around the features used to train the algorithm	Maintain managers' responsibility for AI-based decisions Create an appeal process	Regularly solicit employee feedback Monitor employee engagement

Note: AI = artificial intelligence; HR = human resource.

of plausible causal models for further consideration.[50] Evidence that the algorithms are based on factors that do cause relevant outcomes goes a long way toward mollifying critics of them.

To address the vulnerability of algorithm designers to criticism, we suggest that they channel criticism and politics through AI Councils that include widely respected representatives of all stakeholders. Those councils should debate the assumptions, data, and ethical dimensions of AI algorithms and solicit employees' contribution and feedback. Google moved in this direction in March 2019 by announcing the Advanced Technology External Advisory Council to guide the company on how to ethically develop its AI technologies for business. The Council lasted for only a week, however, because of the controversy around one of its members. Some commentators of this debacle justly noted that "Google already has a tremendous resource in many of its own employees"[51] and, we would add, has to treat its employees as internal clients who deserve to be consulted on when and how AI affects their work and careers.

The push toward causal modeling might face the resistance of organizational data scientists who rely on associational methods of algorithm development. However, we expect the trend toward causal algorithms to move rapidly from academic circles to the public arena as society confronts multiplying legal and ethical challenges of AI. Targeted learning[52] is one methodological approach that combines correlations-based pattern recognition with the subsequent targeted estimation of causal parameters and thus delivers a larger scope of benefits: more accurate predictions, generalizability, explainability, and fairness.

Randomization and experiments constitute our second principle that can help with algorithmic-based decisions. First, intentionally randomizing the inputs into an algorithm is akin to quasi-experimentation and can help to establish causality. Second, acknowledging the random nature of some HR outcome and being explicit about it acknowledge the inherently stochastic nature of HR outcomes and the unavoidable inaccuracy of algorithms. Employees may perceive such randomization—such as flipping a coin—to produce fairer outcomes under uncertainty. This is particularly important where some form of discrimination against legally protected groups is entrenched in the organizational structures, processes, or culture and makes the use of objective assessments (e.g., an algorithm based on historical appraisal scores) unreliable.

Algorithms are much easier to accept, of course, if individuals have the final say over outcomes. Managers may wish to do so in part to maintain control over outcomes and, in turn, over their subordinates. In this case, AI turns into augmented intelligence, which is the dominant modus operandi in data science today. To preserve the integrity of decision making, human judgment should be exercised consistently, according to standard rules. Bias is likely, for example, if a manager can insert their opinion into determining the merits of candidate A but not for other candidates or for one criterion in this decision and a different criterion in another decision. Meehl's classic finding that simple

rules that standardize the process of decision making are better than human "clinical" judgment has withstood the test of time and has been incorporated into the best HR decisions, such as structuring the interview process.[53] Employers can also solicit employee contributions on the criteria to be used in AI algorithms and on how the algorithms' quantitative outputs should be used in a final decision.[54]

Indeed, *employee contribution* is the third critical response to all the challenges of AI in part to address the evidence on algorithm aversion noted above. From the standpoint of organization theory, we see AI as an innovative organizational process whose introduction can be enabling rather than coercive[55] if all stakeholders can participate in the AI Life Cycle. An important, ancillary outcome from the process of introducing algorithms is that it forces organizations to articulate and face up to how they are making decisions right now (e.g., we let hiring managers use whatever criteria they want). The successful removal of human arbitrariness from HR decision making should by itself lessen algorithmic aversion and make workers accustomed to AI-managed organizations. It is also the area where HR and data science have a clear common interest. Formal channels for addressing high-stakes decisions that result from algorithms and for submitting feedback more generally will help with the acceptability of these decisions.

Vendors of cloud-based HR services are positioned the best to develop valid causal models for recruitment, from which all their clients can potentially benefit. Whether clients are willing to let their data be aggregated for this opportunity to realize remains to be seen, as does whether vendors will take the high road in searching for the best and fairest decisions. Escalating our suggestion of AI Councils up "into the cloud"—having vendors create such councils—can be one way to let clients monitor vendors' handling of their data and decide what algorithms can be implemented.

The arguments above suggest how to take on HR questions with data science rather than which tasks to take on. The most complicated and challenging HR task to address with data science techniques is likely to be hiring because so many fairness and legal issues are at play there. Even though hiring may be the most important HR decision, it might make more sense to start elsewhere given its complexity. A good place to start might be with natural language processing analyses of data such as that generated by open-ended questions in employee surveys and performance feedback conducted via apps. Finding patterns in responses would be extremely helpful, few organizations seem to try to find them now, and it is a straightforward task for data science to provide these descriptive results.

The next set of tasks to take on might be those that involve machine learning algorithms but on topics that do not involve HR outcomes subject to legal and fairness considerations. These include "advice" given to employees about training programs that make sense for them, new jobs for which they might apply, wellness and retirement planning advice, and so forth.

Before moving with machine learning into topics such as recruitment and selection, dismissals, or promotion decisions, where legal and fairness questions are paramount, it makes sense to see how what we are doing now is working. That question is best answered with traditional statistical methods and hypothesis testing: do employee referrals actually provide the best candidates for us, do the personality tests we give predict good performers, and do graduates from elite schools actually perform better in our jobs than other hires?

Conclusion

While the deployment of general-purpose AI is still a long way away in any domain of human activity, the speed of progress toward specialized AI systems in health care, automobile industry, social media, advertising, and marketing has been considerable. Far less progress has been made in issues around the management of employees on the first step of the AI path, which are 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.

We also recognize the limits of a top-down, optimization approach to HR decisions because of the negative effects it is likely to have on employee behavior. Ensuring employee involvement in the process of building and using algorithms is necessary for their success.

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, and new competencies may be needed to make that happen. The use of data analytics should help the HR function integrate more closely with other parts of the business, particularly finance and operations. There is a risk to HR leaders that if they do not engage the possibilities of AI, some other function in the business will take control of it for them.

Line managers will have to refresh their skill set as well. For them, AI at present implies "augmented intelligence," an informed use of workforce analytics insights in decision making. The literature on EBMgmt proposes a Bayesian approach to systematically updating managerial beliefs with new information.[56] 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, as March and Simon noted.[57] 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 article will move AI management in HR forward on both counts, those of efficiency and appropriateness.

Authors' Note

In keeping with our arguments, the order of the authors is random.

Author Biographies

Prasanna Tambe is a professor at the Wharton School (email: tambe@wharton.upenn.edu).

Peter Cappelli is a professor at the Wharton School and a research associate at the NBER (email: cappelli@wharton.upenn.edu).

Valery Yakubovich is a professor at ESSEC and a senior fellow at the Wharton School (email: yakubovich@essec.fr).

Notes

1. IBM, "Unplug from the Past: 19th Global C-Suite Study," IBM Institute for Business Value, 2018, https://www.ibm.com/downloads/cas/D2KEJQRO.
2. LinkedIn, "The Rise of HR Analytics," 2018, https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions/talent-intelligence/workforce/pdfs/Final_v2_NAMER_Rise-of-Analytics-Report.pdf.
3. David Meyer, "Amazon Reportedly Killed an AI Recruitment System Because It Couldn't Stop the Tool from Discriminating against Women," *Fortune*, October 10, 2018, https://fortune.com/2018/10/10/amazon-ai-recruitment-bias-women-sexist/.
4. Eric Barends and Denise M. Rousseau, *Evidence-Based Management: How to Use Evidence to Make Better Organizational Decisions* (London, UK: Kogan Page, 2018); Jeffrey Pfeffer and Robert Sutton, *Hard Facts, Dangerous Half-Truths, and Total Nonsense: Profiting from Evidence-Based Management* (Boston, MA: Harvard Business School Press, 2006); Denise Rousseau, ed., *The Oxford Handbook of Evidence-Based Management* (Oxford: Oxford University Press, 2012).
5. Judea Pearl, *Causality* (Cambridge, UK: Cambridge University Press, 2018); Judea Pearl and Dana Mackenzie, *The Book of Why: The New Science of Cause and Effect* (New York, NY: Basic Books, 2018).
6. Bo Cowgill, "Bias and Productivity in Humans and Algorithms: Theory and Evidence from Résumé Screening" (working paper, Columbia University, New York, NY, 2018); Jerker Denrell, Christina Fang, and Chengwei Liu, "Perspective—Chance Explanations in the Management Sciences," *Organization Science*, 26/3 (May/June 2014): 923-940; Christina Liu and Jerker Denrell, "Performance Persistence through the Lens of Chance Models: When Strong Effects of Regression to the Mean Lead to Non-monotonic Performance Associations" (working paper, Warwick Business School, Coventry, UK, 2018).
7. E. Alland Lind and Kees van den Bos, "When Fairness Works: Toward a General Theory of Uncertainty Management," *Research in Organizational Behavior*, 24 (2002): 181-223.
8. Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact," *104 California Law Review*, 671 (2016): 671-732; Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan, "Algorithmic Fairness," *AEA Papers and Proceedings*, 108 (2018): 22-27.
9. F. David Schoorman, "Escalation Bias in Performance Appraisals: An Unintended Consequence of Supervisor Participation in Hiring Decisions," *Journal of Applied Psychology*, 73/1 (1988): 58-62.
10. Peter Cappelli and Anna Tavis, "The Performance Management Revolution," *Harvard Business Review*, 94/10 (October 2016): 58-67.
11. Angelo DeNisi and Caitlin E. Smith, "Performance Appraisal, Performance Management, and Firm-Level Performance: A Review, a Proposed Model, and New Directions for Future Research," *Academy of Management Annals*, 8/1 (2014): 127-179.
12. Peter Cappelli, "There's No Such Thing as Big Data in HR," *Harvard Business Review Digital Articles*, June 2, 2017, pp. 2-4.
13. Enric Junque de Fortune, David Martens, and Foster Provost, "Predictive Modeling with Big Data: Is Bigger Really Better?" *Big Data*, 1/4 (December 13, 2004): 215-226.
14. Michael D. Giandrea and Shawn A. Sprague, "Estimating the U.S. Labor Share," *Monthly Labor Review*, February 2017, https://www.bls.gov/opub/mlr/2017/article/estimating-the-us-labor-share.htm.
15. "Human Resources Cycle: Comparison of Models," *The Oxford Review*, https://www.oxford-review.com/oxford-review-encyclopedia-terms/human-resources-cycle/.

16. Logistic regression refers to a supervised machine learning technique that is commonly used for predictive analysis. Logistic regression uses a logistic function to predict a binary outcome variable of interest.
17. Min K. Lee, Daniel Kusbit, Evan Metsky, and Laura A. Dabbish, "Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (New York, NY: ACM, April 2015), pp. 1603-1612.
18. Serguei Netessine and Valery Yakubovich, "The Darwinian Workplace," *Harvard Business Review*, 90/5 (May 2012): 25-28.
19. Eric Rosenbaum, "IBM Artificial Intelligence Can Predict with 95% Accuracy Which Workers Are about to Quit Their Jobs," *CNBC*, 2019, https://www.cnbc.com/2019/04/03/ibm-ai-can-predict-with-95-percent-accuracy-which-employees-will-quit.html.
20. Kevin R. Murphy, "Criterion Issues in Performance Appraisal Research: Behavioral Accuracy versus Classification Accuracy," *Organizational Behavior and Human Decision Processes*, 50/1 (October 1991): 45-50; Loriann Roberson, Benjamin M. Galvin, and Atira Cherise Charles, "13 When Group Identities Matter: Bias in Performance Appraisal," *Academy of Management Annals*, 1/1 (2007): 617-650.
21. Barends and Rousseau (2018), op. cit.
22. Calle Rosengren and Mikael Ottosson, "Employee Monitoring in a Digital Context," in *Digital Sociologies*, ed. Jessie Daniels, Karen Gregory, and Tressie McMillan Cottom (Chicago, IL: Policy Press, 2016), pp. 181-194.
23. Pearl and Mackenzie (2018), op. cit.
24. Lazlo Bock, *Work Rules! Insights from Inside Google that Will Transform How You Live and Lead* (New York, NY: Twelve, 2015).
25. Philip L. Roth, Philip Bobko, Chad H. Van Iddekinge, and Jason B. Thatcher, "Social Media in Employee-Selection-Related Decisions: A Research Agenda for Uncharted Territory," *Journal of Management*, 42/1 (January 2016): 269-298.
26. Pearl and Mackenzie (2018), op. cit.
27. See www.eugdpr.org.
28. Bradley Areheart and Jessica Roberts, "GINA, Big Data, and the Future of Employee Privacy," *Yale Law Journal*, 128/3 (2019): 710-790.
29. Cynthia Dwork and Aaron Roth, *The Algorithmic Foundations of Differential Privacy* (Boston, MA: Now Publishers, 2014), p. 5.
30. See, for example, Scott A. Birkeland, Todd M. Manson, Jennifer L. Kisamore, Michael T. Brannick, and Mark A. Smith, "A Meta-analytic Investigation of Job Applicant Faking on Personality Measures," *International Journal of Selection and Assessment*, 14/4 (December 2006): 317-335.
31. See, for example, https://www.jobtestprep.co.uk/lloydsbank.
32. Cowgill (2018), op. cit.
33. Pearl (2018), op cit.
34. Ben Schneider, "The People Make the Place," *Personnel Psychology*, 40/3 (September 1987): 437-453.
35. Matthias Spielkamp, "Inspecting Algorithms for Bias," *Technology Review*, 120/4 (July/August 2017): 96-98.
36. For details on the relevant U.S. legal requirements, see David J. Walsh, *Employment Law for Human Resource Practice* (Boston, MA: Cengage Learning, 2015).
37. Joshua R. Loftus, Chris Russell, Matt J. Kusner, and Ricardo Silva, "Causal Reasoning for Algorithmic Fairness," *arXiv preprint arXiv:1805.05859*, 2018.
38. For a critical review, see Ronald J. Karren and Larry Zacharias, "Integrity Tests: Critical Issues," *Human Resource Management Review*, 17/2 (June 2017): 221-234.
39. Loftus et al. (2018), op. cit., p. 7.
40. David Malinsky and David Danks, "Causal Discovery Algorithms: A Practical Guide," *Philosophy Compass*, 13/1 (January 2018): e12470.
41. Cowgill (2018), op. cit.
42. Lind and van den Bos (2002), op. cit.
43. See Peter Stone, *The Luck of the Draw: The Role of Lotteries in Decision Making* (Oxford: Oxford University Press, 2011).
44. Jason Bloomberg, "Don't Trust Artificial Intelligence? Time to Open the AI Black Box," *Forbes*, September 16, 2018, https://www.forbes.com/sites/jasonbloomberg/2018/09/16/dont-trust-artificial-intelligence-time-to-open-the-ai-black-box.

45. Chiara Longoni, Andrea Bonezzi, and Carey K. Morewedge, "Resistance to Medical Artificial Intelligence," *Journal of Consumer Research* (forthcoming), doi:10.1093/jcr/ucz013.
46. See https://www.darpa.mil/program/explainable-artificial-intelligence.
47. Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey, "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms if They Can (Even Slightly) Modify Them," *Management Science*, 64/3 (2016): 1155-1170.
48. Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey, "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err," *Journal of Experimental Psychology: General*, 144/1 (March 2015): 114-126.
49. For example, see Cowgill (2018), op. cit.
50. See Malinsky and Danks (2018), op. cit.
51. See https://www.technologyreview.com/s/613281/google-cancels-ateac-ai-ethics-council-what-next/.
52. Mark J. van der Laan and Sherri Rose, *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies* (Cham, Switzerland: Springer, 2018).
53. Paul Meehl, *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence* (Minneapolis: University of Minnesota Press, 1954).
54. Daniel Kahneman, *Thinking, Fast and Slow* (New York, NY: Farrar, Straus and Giroux, 2011), pp. 230-231.
55. Paul Adler and Bryan Borys, "Two Types of Bureaucracy: Enabling and Coercive," *Administrative Science Quarterly*, 41/1 (March 1996): 61-89.
56. Barends and Rousseau (2018), op. cit.
57. James March and Herbert A. Simon, *Organizations* (New York, NY: Wiley, 2011).