Practical and Ethical Perspectives on AI-Based Employee Performance Evaluation

Scott Pletcher

Purdue Polytechnic Institute

OLS 58100 — Human Capital Management

Dr. Laura Boehme

April 28, 2023

Contents

Practical and Ethical Perspectives on AI-Based Employee Performance Evaluation

For most, job performance evaluations are often just another expected part of the employee experience. While these evaluations take on different forms depending on the occupation, the usual objective is to align the employee's activities with the values and objectives of the greater organization. Of course, pursuing this objective involves a whole host of complex skills and abilities which sometimes pose challenges to leaders and organizations. Automation has long been a favored tool of businesses to help bring consistency, efficiency, and accuracy to various processes, including many human capital management processes. Recent improvements in artificial intelligence (AI) approaches have enabled new options for its use in the HCM space.

One such use case is assisting leaders in evaluating their employees' performance. While using technology to measure and evaluate worker production is not novel, the potential now exists through AI algorithms to delve beyond just piece-meal work and make inferences about an employee's economic impact, emotional state, aptitude for leadership and the likelihood of leaving. Many organizations are eager to use these tools, potentially saving time and money, and are keen on removing bias or inconsistency humans can introduce in the employee evaluation process. However, these AI models often consist of large, complex neural networks where transparency and explainability are not easily achieved. These black-box systems might do a reasonable job, but what are the implications of faceless algorithms making life-changing decisions for employees?

Literature Review

Employee Performance Evaluations Use and Efficacy

Employee performance appraisals are a Human Resources Management tool many organizations use in an attempt to align their employee activities with the overall organization's objectives and are often linked to a person's career progression at an employer (Rasch, 2004). They are implemented with varying degrees of formality and validity and can be a source of stress, conflict, and toil for employees and managers. Many variables exist in the employee evaluation process, and critics abound (Cappelli, Tavis, et al., 2016; Rasch, 2004; Roberts, 2003).

Cook (1995) argues that organizations can easily over-rely on performance appraisals as the only way to manage staff, despite the assertion that most performance appraisal processes are fraught with bias, politicking, and self-promotion. Moreover, Cook suggests that providing supportive, fair, and actionable feedback is a skill that is commonly lacking in leadership ranks. Unless this skill is specifically trained and developed, the staff who are effective at politicking and self-promotion will continue to be promoted into those management roles, creating a feedback loop that "simply perpetuate[s] an unsatisfactory status quo" (1995, p. 3).

Some organizations transitioned away from performance reviews entirely in the mid to late 2010s, citing administrative overhead and a transition toward employee development over employee admonishment as a motivational tool (Cappelli, Tavis, et al., 2016). Additionally, the move toward more team-oriented goal setting left the pure evaluation of the individual a tough proposition. This transition was not without its challenges, however. For example, some companies rely on the performance appraisal process as a documented standard process to record and manage poor performers out of the organization. Lacking such a process, organizations are potentially opening the door to improper termination allegations or inconsistent treatment of employees.

Automated Employee Evaluation Systems

The opportunity to offload an administratively intensive activity such as employee evaluations to automated systems has been enticing since the early days of workplace computing resources. Reavis (1973) hypothesized that the computer systems deployed at that time in the commercial banking industry could monitor loan officer activities and provide automated ratings of the loan officer's effectiveness, efficiency, and profitability. Computer-based evaluation systems are used regularly in education and certification scenarios where the participant is given a series of questions that they must then answer to prove their mastery of a topic (Malec, 2020). Historically, these computer-based tests were relatively simple, consisting of true/false or multiple-choice questions. In recent years, the methods by which the participant can respond have widened to include spoken words, diagrams, and other complex actions. These more complex feedback mechanisms have enabled computer-based evaluations to include foreign language mastery and artistic ability.

More recently, organizations are beginning to leverage AI for more overt employee performance evaluations, but these organizations are often not publicly forthcoming about their internal and external activities. Amazon, for example, has used AI models in many different and documented ways, including how it manages employees. In 2018, Amazon confirmed using a machine-learning model to evaluate and score applicant resumes since 2014 (Dastin, 2018).

Amazon claimed that the machine learning model was not used solely to make hiring decisions but contributed to a candidate score which was only one aspect of the candidate evaluation. Later, in 2019, through court filings and Freedom of Information Act requests, reporters found that Amazon actively used another machine learning model to dynamically evaluate the productivity and efficiency of warehouse workers (Lecher, 2019). In cases where workers could not maintain efficiency standards, the machine learning model would generate automatic warnings and termination notices without management input. Amazon maintained, however, that the manager could override the process.

Bias in AI Evaluation Systems

Unfortunately, the revelation of Amazon's resume evaluation model became public as negative news in that Amazon had discovered this resume evaluation model was biased against women candidates. When Amazon was training its hiring model, engineers fed in the resumes of those hired in the past ten years, with the reasoning being that those were well-vetted by human hiring managers and served as a good model for future hires (Dastin, 2018). The data scientists creating the model failed to consider that, as with most technology companies of the early 2000s, most of their hires were male.

In the training process, the machine learning model shaped itself to tune in to the subtle terms included more frequently in male resumes, rating those words more preferred than other words that might occur in female resumes. This oversight is a type of sampling bias, known as coverage bias, where the training data sets do not fully and fairly represent the intended target population. Unfortunately, this scenario is not uncommon. Because many complex AI models lack transparency and explainability in making decisions, these biases can persist unknown until patterns are observed in their output.

AI-Based Employee Performance Evaluation

Patel et al. (2022) attempt to demonstrate a rather complex method for automated employee evaluation using an ensemble of machine learning algorithms against data points such as the employee's department, the distance they travel to work, an employee's compensation, age, and past performance ratings. While the researchers tout a high level of accuracy in their ensemble model, they intentionally excluded some seemingly relevant features from the initial training and testing dataset, such as the amount of training the employee received in the last year, employee job satisfaction, and employee job level. Further, the research focuses on calling out poor performers rather than identifying latent high performers. Birhane et al. (2022) brought attention to the tendency of research machine learning applications to evaluate themselves on statistical error rates versus the actual social, ethical, and human elements. The former research seems to fit in this bucket.

Other researchers have constructed more realistic models using simulated data, which claim to divine whole catalogs of employee predictive data, such as leadership potential, job suitability, and employee engagement (Ali et al., 2022; Mantello et al., 2023; Umadevi, 2021). Several commercially available employee evaluation platforms have been in place for some time, actively evaluating workers. Enaible is an AI startup that provides worker-tracking software and individual productivity coaching using monitoring agents loaded on the workers' computers, data analytics, and machine learning ("Enaible," 2020). MetLife Insurance trains and evaluates call center workers using realistic chatbot conversations driven by machine learning. The company also analyzes actual calls between call center workers and customers to extract the sentiment and efficiency of the interactions (Tong et al., 2021).

Benefits

One of the promises of automation is that it can offload tedious and time-consuming tasks, freeing humans for more value-add activities. Indeed, instances where AI models have been introduced have resulted in some reduction of administrative activities. However, numerous examples show how AI-based systems confidently provide wrong answers (Belanger, 2023; Ip, 2023). This error seems to be correlated with the complexity of the task and may be driven by overambitious applications of our current AI models and limitations.

A commonly cited benefit of AI-based performance evaluations is the belief that computers will be less prone to favoritism or other forms of bias, such as the recency effect or halo effect (Tong et al., 2021). However, this assumes that bias has not been introduced to the model through data selection in the training process or lack of data points to base the employee's performance score. Amazon's resume evaluation experience notwithstanding, Ali et al. (2022) demonstrated a natural-language model which could classify resumes with 98% accuracy compared to a human reviewer but do so in milliseconds at scale.

Machine learning evaluation models may also help identify desirable leadership traits in employees early in their careers. Umadevi (2021) found that machine learning models fed historical employee performance data could positively identify transformational leaders based on team leadership, problem-solving, and conflict management dimensions. Granted, an attentive manager could perform this same evaluation, but as the study's author points out, not all managers are attentive or open to the possibility that one of their direct reports might be ready for promotion.

Risks

Data analytics and machine learning models require large amounts of data for training and inference to produce accurate results effectively. With few exceptions, HCM processes are not usually associated with generating volumes of granular data (Tambe et al., 2019). This can create a condition data scientists call sparse dimensionality, where data is not fully present enough to use for prediction confidently. Furthermore, very few occupations can be boiled down to pure metrics which can be captured digitally and accurately. For example, it might be possible to track a warehouse worker's location, activity, and ultimate efficiency in purely objective and electronic ways accessible to AI models (Lecher, 2019). However, other factors would likely distinguish that employee as a 'valuable resource' above and beyond their pure motion, such as helping out another worker or having solid problem-solving skills. These ancillary attributes would only be visible to human evaluators.

Another challenge for AI-based employee evaluations is that AI models rely on associations and correlations rather than direct causal evidence (Tambe et al., 2019). The individual, the team, the management, the culture, the environment, and many other factors can influence the ability of an individual contributor to meet their performance goals. The entirety of the circumstances around an employee's performance would be difficult to quantify, if not impossible, to fully capture.

An AI-based employee evaluation model could provide a perfectly objective ranking of workers among their peers. However, it could fail to realize that one worker also has a medical condition that limits dexterity. Assigning a lower rating based on the impact of their medical condition is unfair to the worker and potentially exposes the company to legal issues (Egger, 2020). Improving such an AI model would typically involve training the model on data consistent with that employee's particular circumstances, but assembling a statistically significant population seems untenable. A human could fine-tune the model, but that would require the human to define just how efficient a disabled worker should be to avoid a low-performance rating. Should employers construct a sliding scale based on age, physical fitness or agility?

Perceptions of AI-Based Performance Evaluations

Employer Perceptions

Some researchers insist that employers have taken a decidedly "Theory X" approach to AI for evaluating their employees by creating an ever-present watchful eye to monitor activities (Mantello et al., 2023; Roberts, 2003). Another pessimistic view is that some employers doubt their management's ability to properly and fairly evaluate employee performance (Mantello et al., 2023). A turn to AI evaluations, despite the potential risks and pitfalls, could be seen as a way out for organizations that do not have well-established leadership and management development practices. While this may solve a short-term problem, the organization will likely face more problems down the road, just as any organization might by outsourcing some critical competitive practice–they can only hope to be mediocre at best.

Egger (2020) draws attention to the risk of automated AI evaluation models for those covered under the Americans with Disabilities Act. Consider the example of a company using an online aptitude test as part of an application process. For those who use assistive technologies, this online aptitude test may prove more challenging than for others who are not required to use such technology. As a result, while scoring cognitively the same as other candidates, the applicant

requires more time per question and thus cannot complete all the questions in the allotted time. The automated pre-screening algorithm records this score as-is and disqualifies the applicant. Additionally, even if there were an accommodation in this application process, the AI algorithms would likely not have been trained sufficiently on populations with disabilities or who are neurodivergent.

While automation may improve efficiency for some organizations, most complex AI models are not deterministic. They rarely return the same output every time, given the same input. For this reason, AI engineers and researchers often talk in terms of probabilities versus definitives. However, most end users of AI-based services expect answers and results, not vague probabilities. Without proper explainability behind the rating, this seems to be a recipe for conflict and mistrust among employees and management. Charas and Lupushor call on the Human Resources organization to be "the voice of the worker to ensure that there is no adverse impact, discrimination, bias, unethical use… of AI-enabled tools and solutions" (2022, p. 101). Further, the HR function has a responsibility to the workers of open and transparent communication about how, where and when AI tools are used and what that means for the employees.

U.S. Employer Perspectives

Research has indicated that employees harbor mixed feelings regarding AI for employee evaluations. Lee (2018) found that employees generally view algorithmic decisions with more distrust than a decision made by a human, primarily because they perceive algorithms to be unaware and unable to include context and other soft information in the decision process. Park et al. (2021) found via surveys that employees perceived the AI evaluation systems as generally objective and not influenced by workplace politics. However, the study also found that workers have concerns about AI evaluations which the authors call burdens, based on the framework Suh et al. (2016) used to signify the difficulties automated systems place on the human experiencing those systems. Tong et al. (2021) found that employees accepted AI as part of their evaluation process as long as it was openly disclosed and explained. In contrast, employees had strong negative sentiments when they found out after the fact that AI was used to judge their performance–especially when those evaluations were done in a 'black box' without explanation.

As Roberts (2003) writes, employee participation in the performance assessment process is a vital part of the overall efficacy of the process. Open and honest two-way communication contributes significantly to the process. It is unclear how AI-enabled evaluation systems could carry on such dialog, much less a manager who is provided an evaluation with little explanation. Another critical aspect of sound performance appraisal systems is how employees perceive their rater (i.e., manager) on their ability to provide a proper, accurate evaluation (Roberts, 2003). In traditional appraisal systems, a secondary review of the manager's evaluation or specific training on best practices in employee evaluations could provide this piece of mind for employees. That same option does not exist with an AI-based evaluation model.

Global Perspectives

Globally, the perception of AI differs across cultures which can impact how employees accept AI in the workplace. For example, in both Western and Japanese pop culture, AI is often depicted in human form as a robot, anthropomorphized with human traits. However, while Western media most frequently portrays this AI robot as the aggressive Terminator from the

movie franchise, Japanese culture and media frequently portray the same AI robot as a friendly, helpful sidekick (Cave et al., 2018).

Employees' expectations of privacy also vary by culture and can impact how a worker perceives the use of AI to evaluate them in the workplace. For example, a Chinese construction company implemented AI-enabled close-circuit camera systems on construction sites to monitor for dangerous conditions or lapses in safety protocols. The cameras also alerted supervisors when workers were loitering for too long instead of working (Chen, 2020).

One cross-cultural study of workers' perceptions of AI found "that being managed by AI is the greatest AI risk perceived by the future international job-seeker" (Mantello et al., 2023, p. 110). While all populations voiced some concern, specific cultures varied in their intensity. Those from Africa and Central Asia were much less concerned from a criticality standpoint. One theory is the influence of the predominant local religion on one's perception of self. Confucianism, for example, places less emphasis on the individual and more on upholding the collective, with deference to a perceived more knowledgeable entity–in this case, an AI model (Mantello et al., 2023). Those from Japan indicated the least concern, which correlates with the Japanese work culture of loyalty, and complete deference to the manager's authority. Those from Indian, Bangladeshi, and Indonesian cultures reported the most anxiety around AI used for evaluations.

In the EU, regulations around AI in the workplace are starting to form. Recently, the EU has drafted some text that suggests using AI for worker monitoring and performance management is a "high-risk" activity, implicitly cautioning organizations that this may soon be an area of more regulation (Mantello et al., 2023, p. 115). Existing EU worker regulations under EU Directive 2002/14/EC require organizations to consult with worker unions and representatives when any new AI tool has the potential to significantly impact the organization and workers (De Stefano & Wouters, 2022). Additionally, organizations are also required to perform risk analysis for any new measures implemented in the organization, and under some interpretations, AI evaluation systems could pose psychological risks.

However, applying some EU laws is not straightforward in this scenario. AI systems can produce materially feasible decisions based on a robust and accurate data set because an algorithmic strategy will find correlations that amount to an intended output. Additionally, designers of AI models–at least the ethical ones–would not directly and intentionally use gender or race as an influential factor. This creates a challenge in applying current EU anti-discrimination laws because there was no explicit choice to use a protected class attribute to make decisions (Ntoutsi et al., 2020). The burden is on the aggrieved to prove discrimination happened, but this would not be feasible as many commercial AI-based evaluation models are regarded as trade secrets and thus private.

Implications and Conclusion

While AI is undoubtedly changing the workplace, some use cases present technical and ethical challenges. Employee performance appraisals have long been the source of consternation and complexity. Conducting them properly and effectively requires skill, experience, objectivity, subjectivity, and considerable emotional intelligence. AI-based models are becoming more advanced each month, and it may one day be possible to simulate these human skills in an intelligent system.

However, the limiting factor in this development, especially for high complexity, high

emotional quotient activities, seems to be the availability of consumable data that ultimately represents a worker's holistic performance. There is room for AI in the evaluation process, such as using specially trained chatbots to help train new managers on having critical conversations with employees. Currently, wholesale outsourcing of employee performance evaluation seems to hold more risk than benefit. Organizations would be advised not to commit fully to AI employee performance evaluation yet. Instead, they can leverage AI-based skill development and coaching products to help their human managers improve in this complex and sensitive task.

References

Ali, I., Mughal, N., Khand, Z. H., Ahmed, J., & Mujtaba, G. (2022). Resume classification system using natural language processing and machine learning techniques. *Mehran University Research Journal of Engineering and Technology*, *41*, 65+.

Belanger, A. (2023). OpenAI threatened with landmark defamation lawsuit over ChatGPT false claims [newspaper]. *Ars technica*. https://arstechnica.com/tech-policy/2023/04/openai-may-be-sued-after-chatgpt-falsely-says-aussie-mayor-is-an-ex-con/

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., & Bao, M. (2022). The Values Encoded in Machine Learning Research. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 173–184. https://doi.org/10.1145/3531146.3533083

Cappelli, P., Tavis, A., et al. (2016). The performance management revolution. *Harvard Business Review*, *94*(10), 58–67.

Cave, S., Craig, C., Dihal, K., Dillon, S., Montgomery, J., Singler, B., & Taylor, L. (2018, December). *Portrayals and perceptions of AI and why they matter*. Apollo - University of Cambridge Repository.

Charas, S., & Lupushor, S. (2022, September 13). *Humanizing Human Capital: Invest in Your People for Optimal Business Returns*. Matt Holt.

Chen, S. (2020). Chinese construction firms using AI to monitor workers' safety … but also to spot 'loiterers' [newspaper]. *South china morning post*. https://www.scmp.com/news/china/science/article/3091738/chinese-construction-firms-using-ai-monitor-workers-safety-also

Cook, M. (1995). Performance appraisal and true performance. *Journal of Managerial Psychology; Bradford*, *10*(7), 3–8.

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

De Stefano, V., & Wouters, M. (2022, March). *AI and digital tools in workplace management and evaluation: An assessment of the EU's legal framework*.

Egger, K. E. (2020). Artificial intelligence in the workplace: Exploring liability under the americans with disabilities act and regulatory solutions student scholarship. *Washburn L.J.*, *60*(3), 527–560.

Enaible. (2020, August). enaible. https://enaible.io/

Ip, G. (2023). The Robots Have Finally Come for My Job [newspaper]. *Wall Street Journal: Economy*. Retrieved April 5, 2023, from https://www.wsj.com/articles/the-robots-have-finally-come-for-my-job-34a69146

Lecher, C. (2019, April 25). *How Amazon automatically tracks and fires warehouse workers for 'productivity'*. The Verge. Retrieved April 4, 2023, from https://www.theverge.com/2019/4/25/18516004/amazon-warehouse-fulfillment-centers-productivity-firing-terminations

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

Malec, W. (2020). Computer-Based Testing: A Necessary Evil or a Sensible Choice? *The Modern Higher Education Review*, (5), 100–113.

Mantello, P., Ho, M.-T., Nguyen, M.-H., & Vuong, Q.-H. (2023). Bosses without a heart: Socio-demographic and cross-cultural determinants of attitude toward Emotional AI in the workplace. *AI Soc.*, *38*(1), 97–119.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., … Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, *10*(3).

Park, H., Ahn, D., Hosanagar, K., & Lee, J. (2021). *Human-AI interaction in human resource management: Understanding why employees resist algorithmic evaluation at workplaces and how to mitigate burdens*. Association for Computing Machinery.

Patel, K., Sheth, K., Mehta, D., Tanwar, S., Florea, B. C., Taralunga, D. D., Altameem, A., Altameem, T., & Sharma, R. (2022). RanKer: An AI-Based Employee-Performance classification scheme to rank and identify low performers. *Sci. China Ser. A Math.*, *10*(19), 3714.

Rasch, L. (2004). Employee performance appraisal and the 95/5 rule. *Community College Journal of Research and Practice*, *28*(5), 407–414.

Reavis, D. R. (1973, May). *Computer-based management information systems: An application to commercial banking* (Doctoral dissertation). The University of Montana.

Roberts, G. E. (2003). Employee performance appraisal system participation: A technique that works. *Public Personnel Management*, *32*(1), 89–98.

Suh, H., Shahriaree, N., Hekler, E. B., & Kientz, J. A. (2016). Developing and validating the user burden scale: A tool for assessing user burden in computing systems. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 3988–3999.

Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*, *61*(4), 15–42.

Tong, S., Jia, N., Luo, X., & Fang, Z. (2021). The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance. *Strategic Management Journal*, *42*, 1600–1631.

Umadevi, K. (2021). Identifying transformational leaders using machine learning. *Zeichen Journal*, *7*(2), 126–141.