HIRING BY MACHINE

CASE STUDY: 5

The development of artificial intelligence (AI) systems and their deployment in society gives rise to ethical dilemmas and hard questions. This is one of a set of fictional case studies that are designed to elucidate and prompt discussion about issues in the intersection of AI and Ethics. As educational materials, the case studies were developed out of an interdisciplinary workshop series at Princeton University that began in 2017-18. They are the product of a research collaboration between the University Center for Human Values (UCHV) and the Center for Information Technology Policy (CITP) at Princeton.

For more information, see **http://www.aiethics.princeton.edu**

DIALOGUES ON
AI AND ETHICS

As the means of warfare have modernized, the US Army has placed increasing emphasis on training new recruits in programming and computer engineering. This focus on tech not only helps US military operations remain competitive on a world stage, but it also provides many military professionals with skill sets that can later be leveraged into non-military environments. This was the case for the small, enthusiastic group of Army veterans who co-founded the non-profit company, Strategeion, after having been honorably discharged during the 2008 recession. Building on their previous experiences supporting various military operations with IT solutions, this group of programmers set out to create jobs for themselves and improve the lives of others by producing an online platform that would enable veterans to stay in touch with their cohorts and share experiences dealing with civilian life.

The co-founders did not stop there, however. Having been instilled with a strong sense of civic virtue, and having witnessed first-hand the problems of poverty, joblessness and homelessness that many American communities were facing during the economic downturn, the developers knew they wanted to use their programming skills to effect broad social change. The group was always looking for interesting new technical problems to address, and vowed to develop services, platforms and technical solutions for the benefit of all. As the company matured, the platform expanded to include a range of services—from social networking to personal blogging and even a location-based search app that helped individuals moving to new communities discover local points of interest—which were popular across many demographics. Strategeion's unofficial motto became "leave no one behind."

In order to fulfill this pledge, Strategeion's founders believed the best path forward was one of collaboration and peer production. One of the company's key commitments was to an open-source model, at least regarding their public-facing products and services. This meant Strategeion would make the source code for much of their software freely available to the public. The hope was that other organizations would not only use Strategeion's code to serve their own communities, but build upon and improve it such that benefits would accrue to all. As an added bonus, the open-source model also ensured some measure of transparency and public accountability. Over time, these features contributed to Strategeion's growing reputation as an honest, trustworthy company.

Discussion Question #1:

Trust is an increasingly important branding tool for tech companies in competitive markets. When a company opts for an open-source model, as Strategeion did, that decision may increase public perceptions of its trustworthiness. But does open sourcing necessarily imply trustworthiness? What other factors, if any, go in to determining a tech company's trustworthiness? If a tech company chooses not to share its source code, does that mean it is untrustworthy?

Strategeion's business model proved successful, with the company maintaining steady growth in both users and revenue. But even as it expanded and entered new markets, Strategeion never abandoned its special commitment to addressing the needs of veterans. This was evident in certain of its staple products, which were geared towards former military servicemembers—for example, a resume writing feature that translates military experience into civilian language—as well as in Strategeion's hiring practices.

Whereas other innovative tech firms mostly employ young, recent graduates from prestigious universities, Strategeion was meant to be staffed largely by ex-military personnel. The company considered this policy to be a win-win. In keeping with its mission, Strategeion was glad to provide job opportunities and support to veterans, a group that had been particularly hard-hit by the recession. Even where jobs were available, many of these individuals, who often returned home bearing the scars of physical and/or psychological

traumas, experienced difficulties adjusting to the civilian workforce. Veterans generally fit in well at Strategeion, however, a company that prided itself in maintaining certain aspects of the military's social culture. Strategeion's ex-military employees tended to excel at the company and reported high levels of job satisfaction. As a result, there was very little employee turnover among Strategeion's veteran employees.

In recognition of its high employee satisfaction and retention rates, *Wealth* magazine listed Strategeion among its "100 Best Companies to Work For 2013." This resulted in a surge in job inquiries. By the following year, when Strategeion was once again featured on the list, the number of applications had far outpaced the number of positions available, at a ratio of almost 100 to one. And despite minimal civilian outreach efforts, more and more of these applications were coming from the kind of traditional candidates that might have typically applied for jobs at larger, for-profit tech companies. At one point, Strategeion's human resources (HR) team became so overwhelmed with the number of resumes it received that they had to cease hiring to deal with their backlog. HR representatives complained on Strategeion's internal message board that they now expended so much energy on the first-round selection process that they no longer had enough time to execute other essential aspects of the job, such as performing background checks and processing new hires.

A group of Strategeion's developers interpreted the messages from HR as a call for help. In keeping with the company's tradition of developing in-house solutions for internal problems, they offered to create a bespoke resume vetting system to help HR deal with the influx of resumes. Diagnosing the problem as a simple issue of information overload, this group of developers expected it could be easily solved by implementing some clever technical tricks to automatically pre-sort resumes according to a candidate's desirability, optimizing especially for projected "fit" within the company.

After having weighed several options, the team decided to implement a system that utilized natural language processing (NLP) and machine learning (ML) to look for markers in resumes that distinguished the best candidates. They dubbed the system PARiS, in tribute to the Trojan hero who was tasked with judging a contest to determine the most beautiful goddess among the deities of Mount Olympus. In order to train the system, HR provided the engineering team with dozens of resumes from current and previous employees who were deemed either exemplary or especially poor in terms of professional attributes and fit. PARiS would rate incoming resumes according to their match with the ideal types and cast aside those that were below a set threshold.

PARiS' rollout was met with a collective sigh of relief from HR. Because poor matches could be discarded automatically, HR no longer had to devote the overwhelming number of hours required for humans to read each resume the company received. And while some members of the team were initially hesitant about delegating first-stage application sorting to an algorithm, skepticism about PARiS quickly abated as the system revealed its impressive capacity to learn. After only a few weeks in operation, the lists of candidates PARiS suggested consistently reflected those that would have been assembled by human HR agents, instilling confidence that the system had absorbed Strategeion's values. But PARiS was so much faster and more efficient than humans! Over time, growing trust in the system meant that the HR representatives felt less and less need to double-check PARiS' work, and they began shifting their energies elsewhere.

Discussion Question #2:

PARiS promised to make the hiring process more efficient. But are there other values that might be desirable in hiring? Diversity? Equity? Creativity? What, if anything, do companies risk losing when hiring procedures are so singularly focused on maximizing efficiency?

Hara, a promising and hard-working computer science student from Athens, GA, received an automated rejection email from Strategeion within hours of applying for a job through its website. She was surprised at having been tossed aside so quickly, as she had been convinced she was an ideal candidate for the company. She had strong academic qualifications and she had carefully crafted her resume to reflect her civic commitments and experience working with non-profit organizations that advocated for wheelchair users such as herself. Her ambitions to develop transparent, responsible tech solutions to improve the lives of those with disabilities seemed a perfect match for Strategeion's mission to "leave no one behind." Disappointed at her rejection, Hara wrote to the company asking for feedback on her application. She also published a blog post about the experience, promising to share any future response from Strategeion.

Hara's request made its way to the HR department, and the representative who received it was also puzzled by her rejection. After having thoroughly reviewed Hara's application, he judged her to be on par with Strategeion's very best employees in terms of both interests and credentials. Indeed, based on her resume alone, he expected she would make an excellent addition to the company, and he couldn't see a reason for her application being automatically discarded. He decided to flag Hara's case for internal review. At that point, his supervisor decided to use some of the extra time the team had on their hands since the introduction of PARiS to convene a meeting with the system's engineers in order to figure out why the system had rejected Hara's application.

One potential concern going into the meeting was that PARiS may have used Hara's disability status as a reason to deny her application. However, the system's engineers reassured HR that they had explicitly designed the algorithm so that it would not discriminate against protected categories. Furthermore, Strategeion's policy of hiring ex-military personnel—many of whom were wheelchair users—meant that the system's training data was not biased against those with physical disabilities. But if it wasn't her disability, then what was it that PARiS had found in Hara's resume that had caused it to categorize her as a bad fit? What was it that the humans couldn't see?

After much digging, PARiS' engineers found the unlikely answer: sports. It turned out that there was a strong positive correlation between participation in athletics and military service. Given the overrepresentation of veterans among Strategeion's employees and their propensity to excel at the company, PARiS had learned to connect a history of playing sports with "good fit." And while it was true that many of Strategion's ex-military employees no longer participated in sports, their resumes typically reflected a history of having done so. Hara, on the other hand, had never been interested in sports. And, having used a wheelchair her entire life, she also had no history of athletic activities.

Discussion Question #3:

Biased data sets pose a problem for ensuring fairness in AI systems. Given the company's demographics, what could Strategeion's engineers have done to counteract the skewed employee data? To what extent are such proactive efforts the responsibility of individual engineers or engineering teams?

In the interest of openness and honesty, the HR representative in charge of Hara's case reached out to her with the team's findings. He explained that the company had recently incorporated an AI system into its hiring processes. And while PARiS was generally a success, he admitted that there were still some bugs that would need to be worked out. In Hara's case, PARiS had considered her resume's lack of references to physically demanding activities to indicate a weak cultural fit for Strategeion. The HR agent apologized on behalf of the company, invited Hara for an interview and promised that the company was already searching for solutions to PARiS' shortcomings.

Hara was dismayed to learn that Strategeion had delegated decision-making in hiring—an area that could have a profound impact on her life prospects—to an AI system. Even worse, that system had then wrongly discriminated against her! Frustrated and angry, she published the company's response on her website, where her readers joined in discussing several ethical concerns surrounding PARiS.

Ethical Objection #1: Fairness

Hara's frustration about PARiS' immediate rejection of her application was rooted in her belief that she had been treated unfairly. She felt she was a good fit for Strategeion across many dimensions, and thought she deserved a shot at the job. (The HR representative who contacted her conceded as much by later offering her an interview.) Even if it wasn't direct or by design, the fact was that PARiS had ultimately discriminated against her application on the basis of an irrelevant characteristic (i.e., her physical capabilities). This was unfair. All she wanted was to be judged on the basis of her relevant characteristics and achievements. Allowing the system to exclude her because of a lack of sporting experience meant that Hara had not been treated the same as other equally qualified candidates. Some of Hara's colleagues from the non-profit world went a step further, arguing that Hara's situation demanded more than mere equality of opportunity. If anything, they argued, given the history of marginalization and the lack of accommodations traditionally made for persons with disabilities, the fairest thing for Strategeion to have done would have been to engineer PARiS to positively discriminate in favor of those with physical disabilities. They pointed out that helping those in need was supposedly one of Strategeion's core principles and challenged the company to use its hiring tools to correct injustice.

Ethical Objection #2: Dehumanizing Systems

In the US, one's job influences that person's income, housing choices, healthcare options and any number of other essential aspects of life. Decisions about who is or is not afforded a particular job opportunity may not always seem like life-or-death matters, but for many, they either are or approach them in significance. Hara believed that her life prospects would be greatly improved by joining Strategeion, and was therefore disconcerted by the idea of a non-human agent deciding whether she'd even have the opportunity to make her case in a job interview. For a decision that important, she argued that there ought to be a human in the loop. True, humans aren't perfect and they import their own biases into reviewing applications, but Hara pointed out that they can care and empathize with applicants. When human agents reject worthy applicants, they may feel regret. An AI system, on the other hand, feels none of this. Instead, the system applies cold calculations to data in order to determine access to a scarce resource (e.g. jobs). For those personally affected by those calculations, the process of being converted to an "input" and assessed in this manner can feel dehumanizing. Hara even suggested that the sense of dehumanization may extend to the HR workers who had a central aspect of their job replaced by a machine.

Ethical Objection #3: Consent And Contextual Integrity

Hara was dismayed that automated decision-making tools had been used to evaluate her resume without her explicit consent. And she wasn't alone. Upon learning about PARiS, many of Strategeion's current and former employees were unhappy that their resumes might have been used to train the underlying datasets without their knowledge or permission. These employees had provided Strategeion with personal information under the reasonable expectation that it would be used in a limited context (i.e., to inform hiring decisions about them, as individuals). While it is true that expectations regarding what is right and proper for an employer to do vis-à-vis an employee's resume might have been merely implied, rather than explicit, Strategeion's use of its employees' personal information for unexpected and undisclosed purposes left them open to allegations that they had violated privacy norms and standards. Moreover, several employees qualified their complaints by noting that it wasn't just that they had not consented to broad use of their personal information, but that they would not have consented to this particular use, which had the effect of contributing towards discriminatory hiring practices.

Hara ultimately decided to reject Strategeion's offer of an interview and, instead, she filed an official complaint with the company, incorporating many of these arguments. Upon receipt of Hara's complaint, Strategeion's Board of Directors launched an investigation to determine the merits of her claims.

Strategeion first needed to address the legal allegations in Hara's complaint. The Board handed the accusations of inappropriate data use and discrimination to their in-house counsel, who could ascertain: 1) whether Strategeion had committed a legal wrong by using their employees' resumes to train PARiS without their knowledge or explicit consent; and 2) whether PARiS had fallen afoul of US anti-discrimination law. Both accusations were serious, but the lawyers were especially concerned about the latter, given Strategeion's conception of itself as a company that provides fair, transparent, honest tech solutions in service of the public good.

US anti-discrimination law has evolved significantly over the last half-century. The US Constitution has been interpreted to prohibit discrimination against "protected categories"—including persons with disabilities—by federal and state governments against public employees. Private corporations are also subject to a growing body of anti-discrimination law. The Rehabilitation Act of 1973 and the Americans with Disabilities Act (ADA) of 1990 require private employers to treat all prospective job applicants equally. More recently the ADA Amendments Act of 2008 defined "equal treatment" clearly within the framework of the equal opportunity principle, meaning that persons with disabilities cannot be at a placed at a disadvantage in hiring by virtue of their disabilities.

If PARiS had been directed to discriminate against applicants on the basis of their disability status, Strategeion would clearly have violated US law. But that was not the case. PARiS was not intentionally discriminating against resumes based on protected attributes; rather, "redundant encodings" in Strategeion's data had allowed the system to infer such attributes from other, seemingly innocuous data. Thus, Strategeion's lawyers believed they could prove the company was legally in the clear.

Discussion Question #4:

The type of discrimination practiced by PARiS might not seem as blatantly demeaning as a blanket hiring policy against those with physical disabilities, but it is any different from a moral standpoint? How might systems be designed to address this kind of insidious discrimination, which is, by definition, difficult to spot?

But even if the lawyers were able to show that Strategeion had not acted illegally, it was not clear that the company had behaved in accordance with its own ethical principles. Over the course of the investigation, it became clear that Strategeion would need to take a long, hard look at itself. Throughout its history, Strategeion had consistently promoted a robust notion of fairness through its positive efforts to recruit employees from a group they believed to be in dire need of help (i.e., veterans). Indeed, many of those individuals had injuries and ailments that would qualify as disabilities. Yet despite the comoany's best efforts to promote fairness in hiring, the Board had no choice but to acknowledge that, in deploying PARiS, Strategeion had failed to live up to its ambitions. Something would need to be done to ensure that all strong applications were given a fair shot. The question was: what?

A complete overhaul to the company's hiring policies would be difficult. Strategeion wished to be a positive force in the world, but it also wanted to hire individuals who would be in it for the long haul. Thus, projected cultural fit was an extremely important part of their hiring criteria. In order for PARiS to make a determination on fitness, the system's engineers had decided to train it on samples from past and present Strategeion employees. But while this approach meant that PARiS was adept at picking out resumes of people who most resembled successful Strategeion employees, because of the company's historical hiring practices that

favored of military types, it also meant that people who did not fit that mold would be discriminated against. In other words, given the system design, Strategeion's biased data would produce biased results and promote biased outcomes.

One option to address PARiS' bias problem was to implore the system's engineers to infuse more diversity into their models. A second option called for rethinking the value of a homogenous workforce. Recent reports in management studies have shown that more diverse project teams are able to evaluate products and services from a wider range of perspectives, typically resulting in all-around better outputs, as well as more productive workplaces. Upon reading some of this literature, even Strategeion's co-founders tentatively agreed that it might be worth considering a change in hiring priorities.

Discussion Question #5:

Social science increasingly shows that there are advantages to a heterogenous workforce, but there are also advantages to homogeneity. A diverse workforce helps protect organizations against "group think," for example, but groups that share certain experiences and backgrounds may find it easier to communicate with and understand one another, thereby reducing collective action problems. If you were a manager in charge of hiring at Strategeion, for which position would you advocate? Would you try to maintain the corporate culture by hiring people who resemble current employees, or would you argue that PARiS should be realigned to optimize for a broader range of types?

Reflection & Discussion Questions

Fairness: As the discussion surrounding Hara's blog post illustrated, when people speak of "fairness," they are often drawing on several different concepts. For example, fairness may refer to conditions of **equal opportunity**, meaning that all individuals receive the same opportunities to showcase their merits so that they can be judged and rewarded correspondingly. This conception of fairness is often underlined by the principle of meritocracy, or the idea that people should get what they deserve only on the basis of relevant judging criteria (e.g., physical ability, intelligence, hard work). In contrast, fairness may also be thought to require **equality of outcomes**. To satisfy this vision of fairness, the fruits of society must be distributed to everyone equally, regardless of their individual merits or achievements. Instead, distributional fairness is justified on the basis of shared humanity or a person's membership in a community. Going a step further, some claim that fairness should be understood within a **social justice** framework, which demands that active, affirmative efforts be made to correct for inequalities. Supporters of this view argue that fairness can only be achieved when members of groups that have been systematically disadvantaged in a society be given a "leg up" by those who have profited from inegalitarian institutional arrangements. Finally, fairness may refer to the standard of treating all people equally across various dimensions: **procedural, legal, interpersonal**, etc. What any one individual considers to be fair hinges on which of these (or other) definitions of fairness she chooses. Is it what I deserve? What we deserve? What does anyone deserve? Is it for us to have the same or to be treated the same? What would that even look like? These are difficult questions without clear, undisputed answers. And yet despite the lack of consensus about its meaning, fairness remains a prominent moral value – one that engineers may be encouraged to reflect in the design of AI systems. To the extent that fairness must be articulated in order to be coded, it becomes increasingly important that we all understand the different values underlying this principle.

- A computer scientist recently showed that there are no fewer than 21 definitions of fairness used in programming. Philosophers might consider that a low estimate in their own field. If we cannot all agree on what fairness entails, we must accept that different notions of fairness will prevail in different instances. If you were one of the engineers behind PARiS, which interpretation(s) of fairness would you choose? What if you were a job applicant?

- Given the goal of selecting job applicants at Strategeion, to what extent can and should PARiS be programmed to reflect the value of fairness? How could this be operationalized technically?
- Consider this famous thought experiment: Imagine you exist before the institution of society ("the original position") and you have no way of knowing who you are and how you will fit into that particular society once it's formed ("the veil of ignorance"). Given these conditions, what rules would you choose to govern the distribution of social goods, such as jobs? Specifically, how would you wish for an algorithm to determine an individual's suitability for a job interview, if at all? How does your response—which represents one way of thinking about fairness—align with the notions of fairness you discussed in the previous question regarding the perspectives of engineers and applicants?

Irreconcilability: Optimizing for fairness is difficult in part because the concept encompasses many different values (see "Fairness"), but even more so because those values are often mutually exclusive. For example, a company might wish to ensure fairness by providing that all job applicants be judged equally according to their merits without regard to any ascriptive characteristics (e.g. race, sex, etc.). Such an approach to fairness would mean that that company could not, at the same time, promote a notion of fairness that corrects for historical disadvantages and social injustices that may have contributed towards group differences. For someone who cares about both principles, their incompatibility can be frustrating. But it is nothing new. Humans have always had to grapple with the irreconcilability of certain values, which they may hold dear. This often entails making compromises and acting in philosophically inconsistent ways. In the case of AI, it is unclear whether these systems can and should be developed to model human behavior regarding irreconcilable values.

- Some argue that AI systems cannot simultaneously hold and enact several competing principles at once. As products of code that must adhere to the values written into them, AI systems are accused by some of lacking the moral flexibility—or pragmatism—of humans. Others, however, contend that AI are just as capable as humans in this regard. They argue that what humans think of as their capacity for compromise is actually just arbitrariness, and that that trait can easily be encoded into algorithms by incorporating randomness. Proponents of this view argue that algorithmic decision-making isn't exacerbating the irreconcilability problem, but has merely thrown the issue into sharper relief. Discuss which view you find most convincing and why.
- How can programmers account for the multiple values for which they may wish to optimize when those values are orthogonal to one another? Should machines be made to mimic humanity's capacity for moral inconsistency, or should a different approach be devised?

Diversity: Hiring is an inherently discriminatory process in the technical sense that some applicants receive offers and others do not on the basis of certain criteria used to define a "good" job candidate. Companies mostly decide for themselves which characteristics they wish to optimize for and the level of diversity with which they are comfortable. Many, like Strategeion, may prefer a more homogenous model. This is why PARiS was designed to compare new resumes to those of successful employees – the engineering team was trying to minimize the uncertainty that accompanies difference. However, there are some legal restrictions regarding which criteria a company may use to distinguish between job candidates. Hara had the law—both US and international—on her side when she argued that her application should not be dismissed on the basis of her physical disability. (And had PARiS excluded Hara explicitly as a result of her disability, rather than secondary-order characteristics, her case against Strategeion would have been a slam dunk.) Beyond law, much contemporary research now shows that companies enjoy concrete advantages by promoting a diverse workplace.

- Should the values of inclusivity, diversity and tolerance be actively included as functional requirements in AI systems for recruitment purposes? Does your answer change according to your standpoint (e.g. HR representative, Strategeion advisory board member, job applicant, etc.)?
- PARiS' lists of suggested applicants closely resembled the lists that would have been drafted by Strategeion's human HR team. To the extent that PARiS was biased towards a particular kind of applicant, this suggests that the human HR workers were as well. Indeed, it can be argued that PARIS is merely an extension of the human biases already in existence at Strategeion. Are computational biases necessarily worse than human biases in a recruitment context? Or are humans just as bad? Even if humans are no better than machines, are there reasons we might want to keep human actors involved in hiring, at least for people with protected characteristics under the law?

Capabilities: It is unsurprising that humans may wish to use AI technologies designed to save them tedious labor. This may be especially true in cases where an AI system is capable of replicating an individual's decision-making, but in a way that is more efficient and/or faster than a human could achieve. For example, an expert runner who may have, over many years, developed a talent for calibrating his runs to fit his needs at any given moment may choose to delegate that task to an AI enabled technology if that system seems to perform just as well as his own judgment. In such a case, the runner no longer has to think about how fast to move or which direction to turn, and can funnel that mental energy towards other tasks. Over time, however, his capability to navigate his surroundings while running may whither. In the case of PARiS, once the system's outputs consistently cohered to the HR team's expectations of an ideal candidate, agents seemed content to delegate the initial sorting of job applications to PARiS. This gave them more time to worry about other aspects of the job, since they trusted that PARiS knew what it was doing. Indeed, over time, they may have stopped thinking about how to perform first-round application sorting altogether.

- Reliance on systems like PARiS is likely to produce efficiency gains, as they are capable of performing previously human tasks more quickly and with fewer errors. But what, if anything, do organizations risk losing when they replace human judgment with that of AI systems?
- What happens to individuals—and society—when they grow accustomed to trusting AI systems – either more than or in lieu of their own intuitions and training? Is this necessarily problematic?

Contextual Integrity: Contextual integrity is a theory of privacy that evaluates the appropriateness of some use of an individual's information according to how well that use conforms to the reasonable expectations the individual had when consenting to share her information. For example, if you told your local banker about financial difficulties you might be experiencing, you would be shocked if the bank then conveyed that information to local businesses, considering it a breach of trust. You may have agreed to share this information in order to be considered for a bank loan, but that does not imply consent for the banker to share that information with others. In the Strategeion case, previous and current employees felt that the company's use of their personal information to train PARiS violated their privacy. They had entrusted the company with their personal information in order to be evaluated for a job, but argued that any further use of their resumes would have required their consent.

- The theory of contextual integrity provides an alternative to the idea of "informed consent," or the requirement that an agent must be informed about which data will be collected and how it will be used in order for consent to be binding. In the tech area, where informed consent may not always be either possible or desirable, contextual integrity offers a more flexible way to think about appropriate data use. Evaluate Strategeion's use of its employees' resumes using each of these theories. Does one theory produce a more convincing conclusion than the other?
- Had Strategeion replied to its employees' concerns about the ways in which their resumes were used to train PARiS, it may have argued that, once the resumes had been submitted, that data belonged to the company and could be used as its agents saw fit. Do you agree or disagree with this claim? What implications might stem from such a view of data ownership?

AI Ethics Themes:

Fairness
Irreconcilability
Diversity
Capabilities
Contextual integrity