A Deontological Analysis of the Amazon AI Recruitment Tool

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Cooper Scher

Spring 2023

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor
Benjamin Laugelli, Department of Engineering and Society

Introduction

In 2015, it was found that Amazon's AI recruiting tool was discriminating against women hires and perpetuating male dominance of the tech industry. Review of the algorithm and hiring process showed that the recruiting tool was biased to the extent of completely rejecting applicants from certain all-women colleges (Kodiyan, 2019).

Amazon quickly responded to this situation by taking down the software and claiming it did not use it in the hiring process. Current analysis of the situation finds substantial ethical issues of fairness due to the implementation of the algorithm and the resulting bias from that implementation. The sexist preferences of the algorithm were seen as unethical in practice, but–conditioned upon an unbiased algorithm that can mimic complex human decision-making–AI hiring is still believed to be ethical.

This understanding of the case and AI hiring fails to consider how replacing human-based HR, under the premise of a non-biased AI, will be affected by the realities of biased real-world data inputs and the unknowability of some of those biases beforehand. Thus, a framework based on solely "non-biased" AI will miss how such an AI hiring manager will interact with structural biases in its implementation environment.

To better address the practicalities of AI algorithm training and implementation, I will argue through a deontological framework that hiring through an AI HR manager is unethical unless we can sufficiently develop AI to the higher standard of being a moral agent. I will show that first the standard of non-biased AI is not enough to ensure non-biased outcomes, second that non-biased AI as a stand-in for human HR will not adhere to the reciprocity principle, and finally that an AI that rises to the level of a moral agent can better achieve both goals. I will analyze these claims using anecdotal and qualitative evidence from the Amazon AI case study as well as

a review of current research on reducing algorithmic bias to illustrate the shortcomings of ostensibly non-biased AI and compare to a potential AI moral agent.

Background

In 2014, Amazon machine-learning teams began working on developing AI hiring algorithms to quickly sort through resumes and streamline the hiring process by rating each potential job candidate with a score between one and five stars. After a year, the team realized that the algorithm was not giving equal scores to similar male and female candidates. It was uncovered that this was due to the training data. The data consisted of the 10 years prior of Amazon hiring, which was male-dominated and indicative of the entire industry over the period. Phrases like "women" would cause a resume to be penalized, and inclusion of all-women colleges would even result in downgrades. Amazon responded to this by editing the algorithm for these specific terms, but issues remained such that Amazon disbanded the project by 2017 (Kodiyan, 2019).

Literature Review

There are few ethical analyses of the Amazon AI recruitment bot case along with a large amount of research that details the upcoming role AI will and should play in the recruitment process. The ethical reviews of the Amazon AI recruitment case study invariably find that the sexist results of the algorithm are unethical primarily due to issues with fairness and interpretability. On the other hand, the literature is generally approving of the prospect of AI involvement in the recruiting process incumbent upon fixing the bias and interpretability issues. Accordingly, the general AI recruitment research has identified a plethora of frameworks that could be applied to lead to ethical AI recruitment.

In *Ethics Guidelines for Using AI-based Algorithms in Recruiting: Learnings from a Systematic Literature Review*, the authors conducted a systematic literature review (SLS) that surveyed 784 papers on the topic of ethical AI recruitment (Tehseen et al., 2021). The results indicated that an ethical AI hiring practice would meet several criteria such as transparency, justice and fairness, responsibility, non-maleficence, and privacy. From these principles, the authors distilled down an ethical AI recruitment policy to a three-step, checklist-based cycle. The first stage is the consideration of interpretability and fighting bias during algorithm design. The second stage is an iterative process that focuses on trust and transparency and achieving the goals in the first stage. The final stage is the factoring in of practical issues such as privacy and legal issues on the current design, which iterates back into the first stage until completion (Tehseen et al., 2021). An ethical AI recruitment process would adhere to the design principles in each stage of the process.

Mujtaba and Mahapatra agreed with (Tehseen et al., 2021) on the need for AI recruitment to employ interpretable and fair practices. They use the Amazon recruitment case study and take issue with specific algorithmic features of the Amazon recruitment bot. In particular, the authors don't just find the distinction by gender problematic but also gender proxies "because it inferred [gender attributes] from the educational institution listed on the resume of applicants (e.g., all female college or all-male colleges)" (Mujtaba & Mahapatra, 2019, p. 2). Their proposed solution to use and develop algorithms and design that have reduced bias and increase understandability. The authors discuss newer approaches and toolboxes that are available to reduce machine learning biases as well as decode the decision-making processes of such models. They find that implementing these approaches would resolve the issues in the Amazon AI recruitment case and allow for ethical AI recruiting (Mujtaba & Mahapatra, 2019).

While each source takes a different scope to the ethical issues with the AI hiring, they mostly agree on the conclusion that AI recruitment is still ethically plausible if the bias and interpretability of the decision-making process issues are resolved. The current consensus on the viability of non-biased AI fails to consider the challenges of creating such an algorithm with a first-order approach. Additionally, there is an insufficient handling of how an AI stand-in for a human HR officer can affect the potential hires and respect their autonomy as moral agents. In this paper, I will show not only why just attempting to resolve the bias or interpretability problems in AI is not necessarily going to resolve the fairness issues or ethically treat the applicants as agents, but also why the higher standard of AI that meets a definition of a moral agent is sufficient.

Conceptual Framework

The morality of the Amazon AI recruitment bot and its representation of the potential of AI replacement of human HR functions can be broken down under Immanuel Kant's theory of Deontology. Deontology, a type of duty ethics, is an ethical framework that attempts to determine morality through a set of rules, particularly by respecting the autonomy of all moral agents (Wilson, 2022). Autonomy follows as the ability to decide what is moral through internal reasoning, and moral agents are those that are capable of practicing autonomy (Van de Poel & Royakkers, 2012).

Importantly, Kantian ethics recognizes a few principles on moral rules to help determine if they respect the autonomy of all other moral agents. The first is the universality principle that stipulates "act only on that maxim which you can at the same time will that it should become a universal law" (Van de Poel & Royakkers, 2012, p. 90). In other words, the universality

principle states that you should only act on moral rules that would hold in all circumstances. The second is the reciprocity principle that states, "act as to treat humanity, whether in your own person or in that of any other, in every case as an end, never as a means only" (Van de Poel & Royakkers, 2012, p. 91). Put simply, the reciprocity principle implies that we should treat every human as a rational actor and allow them to freely make decisions and not just to be used for personal goals. These two principles are the two key guides for the generation of categorical imperatives, or general rules from which moral assessments can be made. In this paper, the first principle will be useful in understanding the morality of the Amazon hiring practices. Additionally, the second principle, the reciprocity principle, will be of key importance in discussing the relationship between HR and potential hires, who are human agents (Van de Poel & Royakkers, 2012).

Originally, the Kantian notion of a moral agent referred to only humans as they were the only beings capable of reasoning out moral decisions, but Daniel Dennett extends the scope of moral agents to potential future AI given that sufficiently advance AI can one day "be morally culpable" and have "higher order intentionality," meaning that they can have "beliefs about beliefs" (Sullins, 2006, p. 26). Specifically, there is a sense of subjective right and wrong that is developed in the higher-order progression of beliefs about possible actions. In the case of AI, such a right and wrong may only refer to beliefs and actions that maximize the success of its higher-order goal, but this is enough of a "state of mind" that provides the ingredients for autonomy and moral decision-making (Dennett, 1997).

The standard of a moral agent AI is important because it distinguishes first-order AI which attempt to accomplish goals directly under a heuristic, such as an algorithm designed to minimize bias, from more advanced, adaptative AI (Martinho et al., 2021). The latter category

could have the same goal as a higher-order goal, learning to change its direct, first-order goals to achieve the original goal. Each lower-order goal would be subject to a type of morality system when judged by the higher-order beliefs (Dennett, 1997).

For this paper, I will use deontological ethics to review the morality of Amazon's AI and Dennett's extension of the moral agent to analyze how a potential AI moral agent could resolve the issues of the Amazon case study better than simply a non-biased AI.

Analysis

The Amazon AI recruitment system resulted in unethical hiring practices that resulted in issues with bias and interpretability of the algorithm's decisions. An attempt to simply create a non-biased AI will fail to resolve the bias issue due to the myriad of issues in real-world data that is needed for training. Similarly, as AI becomes more complicated to deal with the bias issue, the ability for potential hires to understand the system they're being evaluated on diminishes, inhibiting their autonomy. Finally, a potential future AI system that rises to Dennett's standard of moral agent with high-order intentions would uniquely be able to resolve these issues.

Non-biased AI Failure to Defeat Bias

A first-order attempt to create a non-biased AI fails to sufficiently solve the bias issues that presented in the Amazon case study, which results in ethical issues with such an AI recruiting system upon deontological analysis. The sexist bias in the Amazon AI recruiting case study fails under the first formulation of the categorical imperative, or the universality principle. A hypothetical maxim that allows sexist recruitment, or any type of demographic bias, would exclude segments of the population from accessing jobs simply because of attributes unrelated to

their effectiveness as employees. Thus, the biased recruitment tactics are unethical under the deontological framework.

In the Amazon case study, engineers involved in the project commented anonymously that the AI hiring tool was discovered to be sexist after a year of use. In particular, the engineers explained that resumes that included terms such as "women" or "women's chess club" were penalized and "problems with the data that underpinned the models' judgements meant that unqualified candidates were often recommended for all manner of jobs" (Dastin, 2018, p. 2). At the same time, "there is no evidence that Amazon had any intention of being discriminatory" (Zeide, 2019, p. 6). In fact, the engineers responsible for the algorithm "tried to fix [the gender] bias, but there was no way to guarantee it wasn't still happening" (Cole, 2018, p. 1). Thus, despite the lack of intent to create a biased algorithm and even attempts to fix the biases once discovered, the resulting implementation still produced a biased result.

The original intent of the algorithm was to optimize hire quality, a goal that is ethically acceptable if the criteria for hiring are universalizable, clearly communicated and represented to job applicants. It is speculated that biased data input affected the hiring criteria chosen by the algorithm–in this case gender–making the hiring practice unethical (Kodiyan, 2019). The Amazon algorithm used the hiring data for the previous 10 years to train the algorithm to select future hires from a stack of resumes (Kodiyan, 2019). Even though the designers did not seek to create an algorithm that preferentially chose against women, the bias was hidden in the data that reflected an environment where the overwhelming majority of hires in the industry were men. Importantly, the biased outcome is a result of the training data and not the direct algorithm design.

The current literature supports interventions to reduce or eliminate bias in AI algorithms, but there are drawbacks. Researchers behind the large language model, GPT-3, have tried solutions such as filtering, implementing hard-coded rules, and explicit training to recognize bias in its outputs (Martin, 2022). However, each of these strategies has flaws as attempting to minimize bias can be costly in terms of algorithmic complexity and interpretability (Martin, 2022). Similarly, preprocessing steps can be successful at reducing bias against minority demographics, but often require that the possible biases are already known for methods such as oversampling (Mujtaba & Mahapatra, 2019). Due to the pernicious nature of many biases, it isn't possible to piece out every affected group of people until the algorithm is already in use. In the Amazon case study, the algorithm was used for a year before engineers and the company were aware of the issue (Dastin, 2018).

Taken together, the intent to have and implementation of "non-biased" AI algorithms is insufficient to guarantee a non-biased outcome, especially considering the tradeoffs of bias minimization methods. In the Amazon case study, the issue was that the hiring system previously and the resulting training data was biased, and this is not an issue easily solved by such algorithmic changes when the extent and type of biases are unknown. To ensure that bias in hiring algorithms have been treated in an ethically appropriate manner, a stronger standard is needed than simply "non-biased" algorithms.

Non-biased AI Failure to Respect Autonomy

The second ethical issue of importance is the autonomy of the potential hires where the Amazon case study highlights how current AI approaches can fail to properly inform the hires of the hiring criteria. The issue of unclear hiring criteria, as mentioned previously, can occur when

biases are considered by the HR decision-making algorithm. In this case, the Amazon recruitment AI was not just looking for qualified resumes but preferentially male qualified resumes; however, this is only one type of complication with the Amazon case (Kodiyan, 2019).

In the case study, the Amazon hiring algorithm engineers explain that the algorithm was created with the intent "to be an engine where I'm going to give you 100 resumes, it will spit out the top five, and we'll hire those" (Dastin, 2018, p. 1). To both the applicants and the company, the goal was to find the best applicants for the open roles based on previous hires. To achieve this, the engineers trained on hires over past 10 years which, as aforementioned, overrepresented male applicants.

The resulting algorithm did not simply choose the best resumes but discriminated by gender. In fact, "Amazon's hiring application was biased even without using the gender attribute, because it inferred it from the educational institution listed on the resume of applicants (e.g., all female college or all male colleges)" (Mujtaba & Mahapatra, 2019, p. 2). In other words, resume items that were correlated with or identified gender, or proxies, still resulted in bias even when active effort was taken to remove the offending criteria.

The hiring criterion were outside of the intent of the Amazon engineers and hiring team as well as the knowledge of potential applicants given that the biased nature of the algorithm went undiscovered for a year of operation and the use of the hiring AI was kept secret (Kodiyan, 2019). This has implications for the reciprocity principle, or the second formulation of the categorical imperative, where it is unethical to treat moral agents as means to an end. Amazon was attempting to pursue its goal of optimizing hiring quality, but without providing full information to potential applicants about what criteria they were being judged on. Applicants did not know that their gender would harm them in the hiring process, nor were they made aware

of the plethora of gender proxies like previous colleges attended. This results in an unethical situation of the potential hires being treated as a means to Amazon's goal without their ability to exercise fully autonomous decisions due to the lack of information on hiring criteria.

The key issue is the complexity of the machine learning algorithms that can be involved in the process that can make it difficult or even impossible for potential hires to know the standards under which they're assessed. Just for the Amazon engineers to understand the bias of their hiring algorithm, they needed to run 500 models (Dastin, 2018). Neural networks and other deep learning models can have billions of nodes that combine in unintuitive and non-linear ways that prevent any real interpretability without significant abstraction (Mujtaba & Mahapatra, 2019).

Transparent and interpretable hiring criterion are an important part of an ethical hiring practice, which is put at risk with the implementation of AI hiring algorithms as seen in the Amazon case study. This issue is compounded by the complexities of effective and non-biased machine learning algorithm, which can often not be fully understood or require resources that aren't available to everyone. Thus, the first-order attempt to create a non-biased algorithm also runs into ethical problems due to the lack of interpretability of the hiring process that detriments the autonomy of the potential applicants.

Moral Agent AI Resolving Bias and Autonomy Issues

Under Dennett's conception of a moral agent AI, both the bias and autonomy issues could be resolved by an AI who meets this standard. The Dennett standard requires that a potential future AI that would meet the standard of being a moral agent would possess the abilities to "notice—and analyze, criticize, analyze and manipulate" (Dennett, 1998, p. 354).

Dennett notes how such an AI would have to have the ability to maintain "states of mind" where there would be a memory and effect on its actions from previously acquired information (Dennett 1997). Thus, an AI capable of being a moral agent would be able to not only be able to provide outputs in response to some set of inputs but also actively modulate how it would output to future inputs.

Best practices for managing AI bias suggests that "traditional and seemingly sensible safeguards do not fix the problem… [as] solving for fairness isn't just difficult–it's mathematically impossible" (Townson, 2023, p. 1). Of particular importance is the notion that bias cannot be eliminated given that "data is imperfect," and there is a need to "focus instead on remediating it" (Townson, 2023, p. 2). To accomplish this, it is recommended to use "two-model solutions" with an "adversary or auditor" (Townson, 2023, p. 3). This is not dissimilar from Dennett's conception of being able to analyze inputs and outputs. The notion of auditing outputs from an AI hiring is something also accomplished by a model with the ability to repeatedly criticize previous results until a desired state is reached–in this case a model that actively minimizes bias. Additionally, for an AI that can manage bias, "it is important to frequently examine outputs and look for suspicious patterns" (Townson, 2023, p. 4). An AI meeting Dennett's standard would have active memory of past responses to inputs.

One example of AI that begins to meet Dennet's standard of higher-order decision-making is the generative adversarial network (GAN). These algorithms function by generating new examples beyond the ones provided in a training dataset to continuously present challenges to the model. The ability of the algorithm to ideate completely new examples based on previously seen data and change its future behavior makes it a candidate for an AI that could achieve Dennett's standard. GANs have already been used to reduce bias in diverse

demographic groups for insurance risk premiums (Townson, 2023, p. 3). GANs also have the advantage of being interpretable as GAN used for image analysis can "[learn] to automatically discover meaningful visual concepts" (Li et al., 2022, p. 1280). Because of the way GAN algorithms are designed, it is much easier to identify the important features and understand why those features are important in the output of the model for given input (Li et al., 2022). This has twofold importance in increasing interpretability and transparency for a potential AI hiring model and making it easier to detect biases if they occur. Thus, higher-order AI would be sufficient to minimize bias and increase interpretability to respond to the ethical issues of fairness and autonomy respectively.

As I have argued that an AI meeting Dennet's standard of moral agency is sufficient to respond to the bias and autonomy issues, one may consider the point that algorithmic approaches that don't meet this standard may still treat bias and autonomy in a more ethical manner than a human HR. However, such approaches may in some cases make the problem worse due to "algorithmic discrimination and the risk of reproducing existing inequality" (Ajunwa, 2020, p. 1698). For example, Goldman Sachs' shift to algorithmic decision-making has "the (un)intended effects of perpetuating structural biases" (Ajunwa, 2020, p. 1699). Such a method that isn't successful at reaching the proposed standard may lead to the same issues as human HR. Further, the Sachs system "represents an ecosystem in which, if left unchecked, a closed loop forms–with algorithmically-drive advertisement determining which applicants will send in their resumes, automating sorting of resumes leading to automated boarding and eventual automated evaluation of employees, and the results of said evaluation being looped back into criteria for job advertisement (Ajunwa, 2020, p. 1694). Thus, the risk is severe even if such an algorithm is

better than human HR due to the risks of even small biases being amplified over time in a closed feedback loop.

Conclusion

The Amazon case study and the larger notion of AI recruitment exposes striking ethical issues under deontological analysis with bias and the autonomy of moral agents requiring not just non-biased AI but autonomous AI. This analysis clearly shows why we need to be careful about introducing AI elements to replace humans due to the downstream impacts. Some solutions may seem simple like designing around minimizing bias, but that may just lead to more ethical complications. The process determining a sufficient replacement for human HR should be based on methodical ethical analysis that considers the standards necessary for such an AI. For something as important to everyday life as the process through which we get hired, it is essential to consider the ethical facets of revolutionizing the process. If we aren't careful with the future of AI recruitment, a key economic process and the autonomous decisions of billions of actors are at stake.

Word count: 3809

References

Ajunwa, I. (2020). The paradox of automation as anti-bias intervention. *Cardozo Law Review, 41*(5), 1671-1742.

Brożek, B., & Janik, B. (2019). Can artificial intelligences be moral agents? *New Ideas in Psychology, 54*, 101–106. https://doi.org/10.1016/j.newideapsych.2018.12.002

Cole, S. (2018, October 10). Amazon Pulled the Plug on an AI Recruitment Tool That Was Biased Against Women. Vice. https://www.vice.com/en/article/evwkk4/amazon-ai-recruitment-hiring-tool-gender-bias

Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Dennett, D. (1997). When Hal Kills, who's to blame? computer ethics. In Stork, D. (Ed.), *Hal's Legacy: 2001's Computer as Dream and Reality* (351-365). MIT Press. https://doi.org/10.7551/mitpress/3404.003.0018

Gogoshin, D. (2021). Robot responsibility and moral community. *Frontiers in Robotics and AI, 8*. https://doi.org/10.3389/frobt.2021.768092

Kodiyan, A. (2019). An overview of ethical issues in using AI systems in hiring with a case study of Amazon's AI based hiring tool. ResearchGate. https://www.researchgate.net/publication/337331539_An_overview_of_ethical_issues_in_using_AI_systems_in_hiring_with_a_case_study_of_Amazon%27s_AI_based_hiring_tool

Li, C., Yao, K., Wang, J., Diao, B., Xu, Y., & Zhang, Q. (2022). Interpretable Generative Adversarial Networks. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*(2), 1280-1288. https://doi.org/10.1609/aaai.v36i2.20015

Martin, K. (2022). *Ethics of data and analytics*. Auerbach Publications. https://doi.org/10.1201/9781003278290

Martinho, A., Poulsen, A., Kroesen, M., & Chorus, C. (2021). Perspectives about artificial moral agents. *AI and Ethics*, *1*(4), 477–490. https://doi.org/10.1007/s43681-021-00055-2

Mujtaba, D., & Mahapatra, N. (2019). Ethical considerations in AI-based recruitment. *2019 IEEE International Symposium on Technology and Society (ISTAS)*, Medford, MA, USA, 2019, 1-7. https://doi.org/10.1109/istas48451.2019.8937920

Sullins, J. P. (2011). When is a robot a moral agent? *Machine Ethics*, *6*(12), 151–161. https://doi.org/10.1017/cbo9780511978036.013

Tehseen, R., Omer, U., & Farooq, S. (2021). Ethical Guidelines for Artificial Intelligence: A Systematic Literature Review. *VFAST Transactions on Software Engineering*, *9*, 33-47. http://dx.doi.org/10.21015/vtse.v9i3.701

Townson, S. (2023). *Manage AI bias instead of trying to eliminate it*. MIT Sloan Management Review.

Van de Poel, I., Royakkers, L. (2012). *Ethics, technology, and engineering: An introduction*. Wiley-Blackwell.

Wilson, E., & Denis, L. (2022, August 19). *Kant and Hume on morality*. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/kant-hume-morality/

Zeide, E. (2019). Artificial Intelligence in Higher Education: Applications, Promise and Perils, and Ethical Questions. *Educase Review, 54*(3). https://ssrn.com/abstract=4320049