When Can My AI Lie?

Nandhini Swaminathan and David Danks

May 10, 2023

When Can My AI Lie?

Nandhini Swaminathan   David Danks

A considerable amount of research is currently in progress to examine the potential of an artificial intelligence system employing deception for beneficial purposes (e.g., in the education or the health sector). In this paper, we consider the permissibility of deploying an algorithm with adjustments unknown to the algorithm's user to counteract their biases.

We reason that this is necessary for human-AI systems whose decisions impact other humans and where the algorithm's user is biased. After illustrating this need through an example in the healthcare setting, we introduce a framework for identifying where the altered system can be implemented. We also discuss the autonomy-related consequences of such algorithms and conclude with some conjectures about how the framework could be employed in various domains.

Keywords

Deceptive Artificial intelligence, autonomy, human biases

Introduction

In typical human-AI systems (e.g., a judge and an AI tool to assess the risk of recidivism), the human has more authority over the final decision. However since humans suffer from known and predictable cognitive and emotional biases, their biases and prejudices may supersede the decision that should have been taken based on the output of the AI entity. This can decrease the overall system's efficiency and cause many adverse short-term and long-term effects on a societal scale. Most discussions of human-algorithm interaction focus on situations in which the human knows better than the algorithm. But what are the ethical obligations *of the algorithm developer* when they have reasonable, justified expectations that the algorithm user's biases will lead to suboptimal or biased decisions?

This situation might appear to be a case of paternalism since we are asking whether the developer's value judgments should be imposed on the user. However, in many situations, the beneficiary of the developer's actions would be the decision "target" rather than the algorithm user, so the framework of paternalism does not straightforwardly apply. This paper considers a hypothetical scenario where a clinical decision support system (CDSS) adjusts its output, compensating for a clinician's personal biases resulting in the patient receiving the treatment and resources he needs. The scenario motivates our framework for the ethical deployment of deceptive algorithms. This framework includes features both necessary and sufficient for it to be morally permissible for a developer to implement an algorithm that will try to compensate for the users' biases through deceptive means in order to benefit the decision targets. The framework provides guidance for responding to the tension between algorithm user autonomy and the values and interests of the decision target.

Thought Experiment: the ethics of deceiving a doctor to mitigate racial bias

A clinical decision support system (CDSS) helps improve healthcare by enhancing medical decisions with targeted clinical knowledge, patient information, and other necessary health information. A traditional CDSS software aids clinical judgment by matching the characteristics of an individual patient to a clinical knowledge database and providing patient-specific recommendations. Clinicians combine their knowledge with information and suggestions provided by the CDSS to provide the best care. As of 2018, 74% of hospitals in the U.S. use a CDSS machine.[1]

Although deception in healthcare is generally considered wrong as it undermines patients' autonomy and erodes trust, the ethical duty to be honest is not absolute. The prima facie obligations can be overridden in cases where more substantial moral considerations exist.[2]

Consider a case where the clinician's (algorithm user) conscious and unconscious racial biases prevent some patients (targets) from obtaining the required medical care.[3][4] In this case, we contend that compensatory adjustments to the algorithm can be an optimal and ethically defensible intervention to confront and dismantle these inequalities and ensure the ultimate decision is appropriate. For example, the algorithm might portray the patient's symptoms as less or more severe to prompt suitable treatment and resource allocation, given the clinician's history with similar patients.

We propose that it is permissible for a developer to modify their algorithm where the following conditions persist:

1. There exists a justified belief that the algorithm user's biases negatively impact the target
2. There exists a justified belief that the target would consent to the deception if they were made aware of it in advance
3. The moral objective justifying the infringement has a realistic prospect of achievement.
4. The chosen method of deception is as minimal as possible, commensurate with achieving the primary goal of the action.
5. The AI system seeks to minimize the negative effects of the deception.

The main objections to deception are the violation of the duty to be truthful and respect for patient (subject) autonomy. However, the patient ultimately receives the required care, and if they were made aware of the deception of the clinician in advance, they would presumably consent, making this morally permissible. [5] Furthermore, it becomes ethical when the deception is judged with respect to its underlying motive.[6] Moreover, while the deception challenges the clinician’s autonomy, it improves the competency of both the human and the human-AI system.

Furthermore, to ensure conditions 4 and 5 are met, the system could self-analyze to see if it can defend its views and reasoning for a particular output before a body of reasonable people, such as a professional association or a court of law. The purpose of this is to encourage it to reassess the strengths and weaknesses of its justifications and thus reduce the risk.

Usually, when a medical professional is deficient in moral character, their colleagues are required to report them according to Section 4 of the American Medical Association's (A.M.A.) Principles of Medical Ethics. However, this has repeatedly proven to be difficult and ineffective in addressing clinician biases.[7] This difficulty persists in other sectors as well. Hence, we contend that compensatory algorithmic adjustments are morally permissible.

Conclusion

This paper introduced a framework for assessing when to deploy a deceptive artificial intelligence entity. To recap, the framework involves acquiring relevant factual information, evaluating its reliability, and mapping out alternative solutions before resorting to deception. Furthermore, the algorithm should ensure that the deception is of the least amount required and that no negative consequences arise. We think that the following conditions are morally demanding of a developer and decrease arbitrary, purely intuitive implementations.

We consider the framework suitable for other domains where similar situations arise, such as employment (recruiters and hiring AIs) or child protective systems (welfare workers and analytics tools). These situations are united by the fact that the decisions by the human-AI systems impact other humans, and any biases in the system cause tangible adverse outcomes.

In the future, we hope to further explore when it would be obligatory rather than just permissible for a developer to modify the algorithm.

References

1. Healthit.gov. 2022. *Office-based Physician Electronic Health Record Adoption | HealthIT.gov*.
2. Holm SPrinciples of Biomedical Ethics, 5th edn.*Journal of Medical Ethics* 2002;**28:**332.
3. 2003. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. Washington, DC: The National Academies Press. https://doi.org/10.17226/10260
4. Williams, D. R. (2012, August 31). *Miles to Go Before We Sleep: Racial Inequities in Health*. SAGE Journals. Retrieved March 10, 2022, from https://journals.sagepub.com/doi/abs/10.1177/0022146512455804
5. Sokol D. K. (2007). Can deceiving patients be morally acceptable?. *BMJ (Clinical research ed.)*, *334*(7601), 984–986. https://doi.org/10.1136/bmj.39184.419826.80

6. Isaac, Alistair & Bridewell, Will. (2017). White Lies on Silver Tongues: Why Robots Need to Deceive (and How). 10.1093/oso/9780190652951.003.0011.
7. Hoberman, J. (2007). Medical Racism and the Rhetoric of Exculpation: How Do Physicians Think about Race? *New Literary History*, *38*(3), 505–525. http://www.jstor.org/stable/20058020
8. Heyen, N.B., Salloch, S. The ethics of machine learning-based clinical decision support: an analysis through the lens of professionalisation theory. *BMC Med Ethics* 22, 112 (2021). https://doi.org/10.1186/s12910-021-00679-3
9. Baird, Aaron et al. "Stakeholder bias in best practice advisories: an ethical perspective." *JAMIA open* vol. 3,2 142-145. June 28. 2020, doi:10.1093/jamiaopen/ooaa018
10. Kaltoft MK, Nielsen JB, Salkeld G, Dowie J. Towards Integrating the Principlist and Casuist Approaches to Ethical Decisions via Multi-Criterial Support. Stud Health Technol Inform. 2016;225:540-4. PMID: 27332259.
11. Khairat S, Marc D, Crosby W, Al Sanousi A Reasons For Physicians Not Adopting Clinical Decision Support Systems: Critical Analysis JMIR Med Inform 2018;6(2):e24
12. Nabi, Junaid, "How Bioethics Can Shape Artificial Intelligence and Machine Learning," *Hastings Center Report* 48, no. 5 (2018): 10– 13. DOI: 10.1002/hast.895
13. *"Medical Ethics" Wikipedia*, Wikimedia Foundation, July 29, 2019, en.wikipedia.org/Medical_ethics.