



ORIGINAL

The risk of moral outsourcing: why artificial intelligence cannot and should not make our ethical decisions

El peligro de la externalización moral: por qué la inteligencia artificial no puede ni debe tomar nuestras decisiones éticas

Mohammed Zeinu Hassen¹  

¹Department of social sciences, Addis Ababa Science and Technology University. Addis Ababa, Ethiopia.

Cite as: Hassen MZ. The risk of moral outsourcing: why artificial intelligence cannot and should not make our ethical decisions. EthAlca. 2025; 4:428. <https://doi.org/10.56294/ai2025428>

Submitted: 08-03-2025

Revised: 07-06-2025

Accepted: 10-09-2025

Published: 11-09-2025

Editor: PhD. Rubén González Vallejo 

Corresponding author: Mohammed Zeinu Hassen 

ABSTRACT

Introduction: the delegation of ethical decision-making to artificial intelligence (AI), a practice termed ‘moral outsourcing,’ was examined.

Objective: this paper critically analyzes the philosophical and social implications of moral outsourcing to AI.

Method: a philosophical and theoretical analysis was conducted, synthesizing arguments from ontology, ethics, and social theory.

Results: the analysis revealed three core arguments against this practice. First, an ontological gap was identified; AI systems lacked the consciousness and subjective experience necessary for genuine moral agency. Second, the study found that the rich, contextual nature of human ethics could not be successfully reduced to formal logic without mechanizing historical biases and losing essential meaning. Third, it was argued that moral outsourcing would lead to an atrophy of human moral reasoning skills and an erosion of accountability.

Conclusions: it was concluded that AI should be developed as a tool to augment, not replace, human judgment, and that the final authority for ethical choice must remain a fundamentally human responsibility.

Keywords: AI Ethics; Moral Outsourcing; Computational Ethics; Decision-Making; Accountability.

RESUMEN

Introducción: se examinó la delegación de la toma de decisiones éticas a la inteligencia artificial (IA), denominada ‘externalización moral’.

Objetivo: Este artículo analiza críticamente las implicaciones filosóficas y sociales de la externalización moral a la IA.

Método: se realizó un análisis filosófico y teórico, sintetizando argumentos de la ontología, la ética y la teoría social.

Resultados: el análisis reveló tres argumentos centrales contra esta práctica. Primero, se identificó una brecha ontológica; los sistemas de IA carecían de la conciencia y la experiencia subjetiva necesarias para una genuina agencia moral. Segundo, el estudio encontró que la naturaleza contextual de la ética humana no podía ser reducida a lógica formal sin mecanizar sesgos históricos y perder su significado esencial. Tercero, se argumentó que la externalización moral conduciría a una atrofia de las habilidades de razonamiento moral humano y a una erosión de la responsabilidad.

Conclusiones: se concluyó que la IA debe ser desarrollada como una herramienta para aumentar, y no reemplazar, el juicio humano, y que la autoridad final para la elección ética debe seguir siendo una responsabilidad fundamentalmente humana.

Palabras clave: Ética de la IA; Externalización Moral; Ética Computacional; Toma de Decisiones; Responsabilidad.

INTRODUCTION

The modern world is captivated by the promise of artificial intelligence, with its growing presence seen in medicine, finance, and transportation. Its architects suggest it can solve our most difficult problems, proposing that AI can optimize systems and correct human errors. This line of thinking now extends into the domain of morality itself. Proponents of what is sometimes called computational ethics suggest that a sufficiently advanced algorithm, free from human emotion and inconsistency, could make fairer and more effective ethical choices than people can.⁽¹⁾ While this argument holds the appeal of achieving objective and consistent decision-making, it overlooks fundamental limitations.

The idea is seductive, offering a clean, technological solution to the messy, often painful work of human judgment. This delegation of moral judgment to a non-human, computational system is a form of moral outsourcing: the offloading of moral responsibilities onto machines. This is a dangerous and misguided path. Artificial intelligence cannot make our ethical decisions for us; more importantly, it should not. The attempt to do so represents a fundamental misunderstanding of both ethics and technology, with serious negative outcomes for human society.

This article forwards a clear argument that the outsourcing of ethical judgment to AI is a categorical error built on a weak foundation that ignores the very source of moral reasoning. This position will be demonstrated by examining three core areas of failure. First, the article addresses the ontological gap between human agents and artificial systems; AI lacks the essential properties for moral agency, including consciousness and lived experience. Second, it investigates the fallacy of attempting to codify ethics. Human morality is fluid and contextual; it cannot be reduced to a set of programmable instructions without losing its meaning and mechanizing human failings. Third, the atrophy of human morality that results from this outsourcing is considered. Relying on machines for ethical choices weakens our own capacity for moral reasoning and creates a vacuum of accountability. AI can be a powerful instrument—it can provide data and model scenarios—but the final, difficult act of judgment must remain with humans.

DEVELOPMENT

The Ontological Gap

A machine has no stake in the world. It does not feel pain, experience joy, or understand love, loss, or loyalty. Its operations are a series of logical steps, executed with incredible speed and precision, yet they are entirely divorced from the subjective, conscious experience that gives rise to morality in the first place.

⁽²⁾ Ethics is not an abstract calculation but a deeply human activity, born from our existence as sentient beings who share a world and whose actions affect one another's well-being. An AI system can be programmed with the utilitarian principle to maximize good outcomes. It can process data about patient survival rates to decide who receives a scarce organ, perhaps arriving at a decision that appears logical. The machine, however, has no grasp of what a life is, or what its loss means to a family and a community. It is a system executing a function, not an agent making a genuine moral choice.

Proponents of AI ethics might argue that moral agency does not require consciousness, only functional equivalence—a position known as functionalism. However, this view is insufficient. This absence of sentience and consciousness creates an unbridgeable ontological gap between human beings and the machines they create. John Searle's famous Chinese Room argument provides a powerful illustration of this gap. Searle imagined a person who does not speak Chinese locked in a room, receiving Chinese characters and, following a rulebook, producing other Chinese characters as output. To an outside observer, the room appears to understand Chinese. Yet the person inside has no understanding whatsoever, merely manipulating syntactic symbols without any grasp of their semantic content.⁽³⁾ Modern AI systems function similarly. They are masters of syntax, identifying patterns and generating statistically probable responses that mimic human reasoning. But this mimicry is not understanding. An AI cannot be a moral agent for the same reason the person in the Chinese Room cannot be a Chinese speaker: it lacks the essential connection to meaning.

This issue extends to the core of what it means to have a subjective point of view. The philosopher Thomas Nagel famously asked, "What is it like to be a bat?"⁽⁴⁾ His point was that there is a subjective character of experience accessible only from the first-person point of view. This subjective quality, or qualia, is the raw material of our ethical world. An AI system has no such inner life. Delegating a moral choice to an AI is like asking a being that cannot see to make judgments about color. The necessary faculty is simply absent.

Human moral reasoning is not built from pure logic alone; it is formed through a lifetime of social interactions, cultural lessons, and personal experiences.⁽⁵⁾ An algorithm's "experience" is merely its training data—a static collection of past events. It cannot understand the subtle, unwritten social codes that govern our interactions.

The discussion of personhood is central here. Philosophers agree that personhood involves capacities for self-awareness, rational deliberation, and moral responsibility, all of which are fundamentally tied to conscious experience.⁽⁶⁾ Artificial systems do not possess these qualities. They are sophisticated tools, not persons. To treat them as moral peers is a category error. Granting them the authority to make ethical choices is to grant personhood to an object, elevating a tool to the status of a judge and devaluing the special standing of human beings as the sole authors of their moral world.

The Fallacy of Codification

The project of moral outsourcing rests on a second flawed assumption: that human ethics can be neatly translated into a formal system of rules and logic that a machine can execute. This is a fallacy of codification. The great ethical traditions are not instruction manuals but frameworks designed to guide the reasoning of conscious, social beings.⁽⁷⁾ They require interpretation, wisdom, and the ability to handle ambiguity.

Consider the three dominant traditions in Western ethics. Deontology, based on duties and universal rules like “do not lie,” appears simple, yet humans constantly apply judgment to navigate exceptions.⁽⁸⁾ An AI lacks this situational judgment. Utilitarianism demands acting to produce the greatest good for the greatest number.⁽⁹⁾ An AI might seem suited for this calculation, but it requires measuring subjective states like happiness and suffering, which a machine with no subjective experience cannot possibly quantify. Virtue ethics, which traces its roots to Aristotle, is perhaps the most difficult to automate. It is about developing a good moral character and asks what a virtuous person would do, an approach dependent on context, experience, and practical wisdom (*phronesis*).⁽¹⁰⁾ There is no algorithm for courage or compassion. A machine cannot be virtuous; it can only simulate the behavior of a virtuous person.

This problem is so deep that moral particularists argue there are no absolute, universal moral principles at all.⁽¹¹⁾ If this view is correct, the entire project of creating a general-purpose ethical AI is impossible. Furthermore, the data used to train AI systems is a mirror of our own flawed world, saturated with the biases and injustices of human society.^(12,13) An AI trained on biased data will learn to replicate that bias. When we ask an AI to make an ethical decision, we are often just asking it to mechanize our own prejudices. This is made worse by the opacity of many advanced systems. The “black box” nature of deep learning models means even their creators often cannot fully explain the reasoning behind a specific output.⁽¹⁴⁾ The immense technical challenge of the “value alignment” problem—encoding complex human values into machines—further underscores this philosophical impossibility. The system gives the cold, impartial authority of a machine, making the embedded biases harder to challenge. This is not a path to a fairer world, but one where our worst instincts are automated and entrenched.

The Atrophy of Human Morality: The Cost of Convenience

Ethical reasoning is like a muscle: it grows stronger with use and weakens with disuse. This process of struggle and reflection is how individuals and societies develop moral character.⁽¹⁵⁾ Moral outsourcing threatens to halt this development. By delegating our ethical decisions to machines, we are avoiding the very work that makes us better moral agents. Over time, our capacity for moral reasoning will atrophy. Nicholas Carr documented how automation can deskill experts, who come to rely on their instruments so heavily that their own intuitive abilities fade.⁽¹⁶⁾ A similar deskilling can happen in the moral sphere.

This dependence also creates a dangerous void of accountability. When an autonomous system makes a decision that leads to harm, who is responsible? This diffusion of responsibility creates what has been called an “accountability gap.”^(17,18) This is a morally corrosive state of affairs. This creates a fundamental crisis for legal frameworks built on the concept of *mens rea* (a guilty mind). How can we assign criminal or tort liability when the “mind” behind a harmful act is an algorithm with no intention, no consciousness, and no capacity for guilt?⁽¹⁹⁾ Moral outsourcing deliberately blurs this line of responsibility, allowing individuals and institutions to benefit from automated decisions while evading the moral weight of their consequences.

This erosion of responsibility has dire consequences for democratic society. Many of the most important questions a society faces are, at their heart, ethical questions of value that must be worked out through public deliberation and compromise.⁽²⁰⁾ Outsourcing these decisions to opaque AI systems transforms political questions into technical ones. This is an anti-democratic move. It proposes a form of technocracy—rule by technical experts—over democratic debate, disenfranchising citizens by taking away their power to shape the moral direction of their society. A healthy democracy requires an active, engaged citizenry, capable of reasoning together about the common good.^(21,22) Moral outsourcing encourages the opposite: a passive population that accepts the outputs of black-box systems without question. This is a recipe for a more authoritarian society.

Finally, the very struggle with ethical dilemmas is a source of human growth and meaning. A world where an AI provides the “optimal” ethical solution is a world with fewer opportunities for this kind of moral growth. It would be a morally shallower world. The goal of a good human life is not to eliminate difficulty, but to learn to face it with courage and wisdom. The struggle is not a problem to be solved by technology; it is an essential

part of the human condition.

CONCLUSIONS

The push to create artificial intelligence systems capable of making ethical decisions is born from a noble desire to build a better, fairer world. It is, however, a profoundly mistaken project. It is not progress, but a form of surrender. Moral outsourcing cedes one of the most fundamental aspects of our humanity to unfeeling, unthinking machines. This is a path that must not be taken.

The argument presented here has shown this from three distinct angles. First, AI systems lack the ontological foundations of moral agency. Second, the attempt to codify the rich, contextual nature of human ethics into formal logic is a fallacy. Third, the long-term consequence of relying on AI for our choices will be the atrophy of our own moral capacities and a dangerous erosion of accountability.

This does not mean that AI has no role to play in our ethical lives. It can be an extraordinarily powerful tool for informing human decision-makers. AI can process massive amounts of data, model complex scenarios, and check for patterns of bias. In these roles, AI does not replace human judgment; it serves human judgment. The machine can provide the data, but the human must make the decision. The machine can show us the patterns, but the human must provide the interpretation. The machine can calculate the risks, but the human must bear the responsibility. The future of ethical AI is not a future where machines decide for us; it is a future where machines help us decide better for ourselves. The burden of choice is a heavy one. By synthesizing ontological, epistemological, and sociological critiques, this paper provides a comprehensive philosophical framework for rejecting the project of moral outsourcing. But it is our burden. We must not, and cannot, outsource it.

REFERENCES

1. Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, et al. The Moral Machine experiment. *Nature*. 2018;563(7729):59-64.
2. Santana-Soriano E. Ética y filosofía de la inteligencia artificial: debates actuales. *La Barca de Teseo*. 2023;1(1):47-64.
3. Searle JR. Minds, Brains, and Programs. *Behavioral and Brain Sciences*. 1980;3(3):417-57.
4. Nagel T. What Is It Like to Be a Bat? *The Philosophical Review*. 1974;83(4):435-50.
5. Siau K, Wang W. Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI. *Journal of Database Management*. 2020;31(2):74-87.
6. Hanna R, Kazim E. Philosophical foundations for digital ethics and AI Ethics: a dignitarian approach. *AI and Ethics*. 2021;1:405-23.
7. Prem E. Approaches to Ethical AI. In: Werthner H, et al., editors. *Introduction to Digital Humanism*. Springer; 2023.
8. Kant I. *Groundwork of the Metaphysics of Morals*. Gregor M, Timmermann J, translators. Cambridge University Press; 2012. (Original work published 1785).
9. Mill JS. *Utilitarianism*. Hackett Publishing Company; 2001. (Original work published 1863).
10. Aristotle. *The Nicomachean Ethics*. Thomson JAK, translator. Penguin Classics; 2004. (Original work c. 350 BCE).
11. Dancy J. *Ethics Without Principles*. Oxford University Press; 2004.
12. Crawford K. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press; 2021.
13. O'Neil C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown; 2016.
14. Burrell J. How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society*. 2016;3(1).

15. Kohlberg L. The Philosophy of Moral Development: Moral Stages and the Idea of Justice. Harper & Row; 1981.
16. Carr N. The Shallows: What the Internet Is Doing to Our Brains. W. W. Norton & Company; 2011.
17. Matthias A. The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology. 2004;6(3):175-83.
18. Siau K, Wang W. Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI. Journal of Database Management. 2020;31(2):74-87.
19. Sparrow R. Killer Robots. Journal of Applied Philosophy. 2007;24(1):62-77.
20. Habermas J. The Theory of Communicative Action, Volume 1: Reason and the Rationalization of Society. McCarthy T, translator. Beacon Press; 1984.
21. Sandel MJ. Justice: What's the Right Thing to Do? Farrar, Straus and Giroux; 2009.
22. Rueda J. ¿Automatizando la mejora moral humana? La inteligencia artificial para la ética. Daimon. Revista Internacional de Filosofía. 2023;89:199-209.

FINANCING

The authors did not receive financing for the development of this research.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Mohammed Zeinu Hassen.

Analysis: Mohammed Zeinu Hassen.

Methodology: Mohammed Zeinu Hassen.

Drafting: Mohammed Zeinu Hassen.

Writing and proof reading: Mohammed Zeinu Hassen.