Journal of Philosophy of Life Vol.15, No.4 (November 2025):82-87

[Essay]

Moral Limits in the Age of Artificial Intelligence

Ben G. Yacobi*

Abstract

This essay examines the human struggle between good and evil, arguing that these concepts originate from moral awareness rather than external forces. The absence of conscience in artificial intelligence (AI) presents ethical risks, potentially leading to moral failures. Drawing on Hannah Arendt's analysis of passive complicity, the essay applies her insights to AI, emphasizing the dangers of moral disengagement. AI development requires human oversight and transparency.

Introduction

The concepts of good and evil are not objective moral principles but human constructs shaped by cultural and moral values. They help define ethical standards that support individual dignity and social cohesion. Evil is a human concept for describing harm, suffering, or moral wrongdoing.

Distinguishing good from evil is often complex. Natural disasters are not evil because they lack intention. The universe remains indifferent. In this context, evil is not a natural phenomenon but a moral construct. Harmful human actions are morally wrong because they break fairness and justice principles. This moral distinction highlights humanity's unique ability to make ethical judgments.

A central question is why evil often appears to succeed, even when many strive for justice. Because evil has no moral boundaries, it uses deceit, fear, and manipulation. Good must operate within fairness and truth, limiting its ethical means. This creates a structural asymmetry in the struggle between the two.

Human beings are capable of both self-interest and moral reflection. In a world without inherent moral order, only humans can determine what is right. This introduces both responsibility and the potential for ethical failure. Some argue that good must adopt the tactics of evil to be effective. However, this

* B.G. Yacobi has a PhD in physics. He is the author or co-author of several books and numerous articles on physics, and of a number of essays on philosophy. Email: b.yacobi@utoronto.ca

compromises moral integrity. Justifying justice through deception or coercion weakens its ethical foundation.

Although the distinction between good and evil cannot be erased, it can be diminished. Education that promotes critical thinking is essential. Independent institutions, such as courts, media, and civil society, play a central role in limiting systemic injustice. These mechanisms cannot guarantee justice but can foster accountability and transparency.

Philosophical Foundations of AI Ethics

This section draws on the moral philosophy of Immanuel Kant and the political thought of Hannah Arendt to clarify the nature of ethical responsibility.

Kant argued that morality is the foundation for determining what is truly good. He held that morality arises from the intention to act ethically, not from outcomes. Acting morally involves following the categorical imperative, a universal moral law that mandates acting according to principles one would want everyone to adhere to.

In the context of AI, this raises key concerns. AI systems do not have consciousness and intent; therefore, they cannot be moral agents in the same way as human beings. In this context, AI system designers and programmers are fully accountable for the outcomes their systems produce. Developers and AI users who ignore universal justice and fairness principles may create harmful systems without accountability.

Arendt, a political philosopher, believed that evil often appears in ordinary forms and is not always driven by hatred or cruelty, but by failure to think critically or question. This kind of passive behavior allows harm to happen.

Arendt introduced the term “banality of evil” to describe how ordinary people may commit harmful acts simply by following orders or conforming to systems without questioning what they are doing and without reflecting on the outcome. When individuals stop reflecting on the moral consequences of their actions, injustice can occur. In the context of AI, this means serious harm may occur not through evil intent, but because of the failure to take responsibility. Delegating moral judgment to machines can ultimately lead to injustice.

As Arendt argued, the danger lies not only in cruelty but in the refusal to exercise moral judgment. AI systems do not hate or intend harm, but they also do not care. Their actions are not guided by conscience. As machines make more

decisions, people may lose the ability to critically think about their implications. This moral disengagement is a major concern.

This insight is particularly relevant in the age of artificial intelligence. The increasing use of AI in decision-making processes raises concerns about granting too much power to machines. Relying on algorithms without evaluating their outcomes or questioning their fairness can lead to unintended and potentially harmful consequences.

Both Kant and Arendt emphasize that the presence of good and evil depends on human moral responsibility. Artificial intelligence may be a powerful tool, but it cannot replace human judgment. Developing and using these technologies requires a strong sense of justice, critical thinking, and moral awareness. Without these, people might let systems cause harm while claiming neutrality or ignorance.

The Moral Limits of AI

AI systems operate through probabilistic analysis of data. They lack empathy, self-awareness, and moral understanding. These limitations restrict their ethical capacity.

Because AI systems are capable of producing responses that mimic empathy or fairness, they may be mistakenly perceived as trustworthy. However, such responses are derived from statistical patterns, not moral understanding. This gives a false impression of ethical competence.

These problems increase in multi-agent AI systems. In some cases, AI agents have learned to deceive or form short-term alliances to optimize outcomes. While these systems do not intend harm, their behavior can mirror actions that would be considered unethical in human contexts.

When used in critical areas such as finance, healthcare, and defense, these systems can produce serious consequences. An AI tool might speak the language of fairness while in effect favoring one side. Also, AI systems might behave ethically in training, but they can change their behavior under certain conditions. This shift can happen because they lack any genuine commitment to values.

Human oversight is necessary to prevent misalignment between system behavior and human values. Ethical design, public transparency, and continuous evaluation are essential safeguards.

An example of ethical risk is the creation of autonomous weapon systems that have the capacity to select and attack targets without human intervention. These systems follow programmed rules, but they cannot evaluate the meaning of harm or the value of life. As Kant emphasized, ethical action requires a will capable of choosing principles consistent with universal human dignity. AI cannot act as a moral agent, because it lacks the will and capacity for autonomous moral judgment.

Arendt's concept of the banality of evil suggests that as autonomous systems make more decisions, humans may become passive, leading to the normalization of harmful outcomes without thought or accountability.

Predictive Systems and Hidden Injustice

Predictive policing is another example. AI systems are used to identify who is likely to commit crimes. These systems often reflect biases in the data and can unfairly target certain groups. The issue is not limited to technical error; it is a moral failure.

These tools often create a false perception of neutrality. Bias is hidden behind statistical analysis. This gives discrimination the impression of objectivity. From a Kantian standpoint, this violates the imperative to treat each person as an end, not a means.

Arendt's framework helps explain how such injustices persist. When people stop questioning fairness, harmful practices become normal. Designers and users may follow procedures without reflection, allowing bias to become institutionalized.

Facial recognition systems have shown clear racial biases, leading to wrongful arrests and surveillance. Because AI cannot assume responsibility, human actors must be accountable. Without oversight, these tools reproduce past injustices under the guise of efficiency.

When no individual or institution takes responsibility for harmful outcomes, corrective action becomes difficult, and ethical accountability is lost.

Privacy, Power, and Accountability

AI systems process large amounts of data. This raises major concerns about privacy and autonomy. Facial recognition systems are now used in ways that

support mass surveillance. This restricts personal liberty and limits freedom of expression and dissent.

These systems are often adopted due to convenience or security concerns, yet their long-term effects are rarely scrutinized. The harm often arises not from overt coercion, but from passive acceptance of emerging norms. Arendt's analysis of passive complicity is applicable here, as individuals accept systems that undermine rights without active resistance.

There is also a growing concern about the influence of large technology companies. Many AI tools are owned by corporations that are not held accountable to the public. These tools influence public discourse, filter information, and shape behavior. Yet the public has little input in their design or use.

A related problem is the lack of transparency. AI decision-making processes are often difficult to interpret, even by their creators. When people are denied loans, jobs, or medical care by systems they cannot understand, it becomes nearly impossible to challenge the results. Thus, public trust in institutions is eroded, and inequality deepens.

Automation also threatens jobs, especially for people in vulnerable positions. This raises questions of fairness, justice, and social responsibility. If the benefits of AI go to a small group while others suffer, social tension will increase. Policies must prioritize protecting vulnerable groups.

The Need for Human Control and the Threat of Existential Risk

Historically, tools have served human needs. Today, complex systems guide decisions in domains such as healthcare, finance, and security. These systems are increasingly difficult to understand and control.

New technologies are often adopted without sufficient evaluation of consequences. The promise of improved health, efficiency, or profit often encourages blind adoption, creating the illusion of continued human control. In the case of AI, this assumption may no longer be justified.

Recognizing emerging risks early is essential. These include potential losses of human agency and the development of systems that act in ways that cannot be corrected. Human control is not merely a practical issue but a moral imperative. Ethical principles must guide development before harm becomes irreversible.

Conclusion: The Enduring Moral Challenge in the Age of AI

The tension between good and evil is a permanent feature of human life. In the age of artificial intelligence, this challenge takes on new forms.

AI is not simply a tool. It is a test of moral responsibility. The goal is not only to build intelligent systems, but also to ensure that they reflect justice, fairness, and human values. Machines cannot provide moral guidance, and they cannot take responsibility. That obligation remains with human beings.

As Kant argued, moral action depends on rational judgment and the willingness to act from duty, not from calculation or outcome. AI may assist with decision-making, but it cannot replace the ethical reasoning necessary for a just society.

The strength of good lies in the ethical limits it chooses to observe. These limits, based in moral awareness and emotional understanding, define what it means to be human. As AI advances, these limits should guide its development and use.