Religious Studies (2025), **61**, 746–748
doi:10.1017/S003441252500023X

CAMBRIDGE UNIVERSITY PRESS

THE BIG QUESTION

A new problem of evil?

Nesim Aslantatar

Department of Philosophy, Indiana University Bloomington, Bloomington, IN, USA
Email: neaslan@iu.edu

(Received 2 April 2025; revised 2 April 2025; accepted 3 April 2025; first published online 14 April 2025)

Abstract

This article examines whether artificial intelligence (AI)-driven harm can be classified as moral or natural evil, or whether a new category – *artificial evil* – is needed. Should AI's harm be seen as a product of human design, thus maintaining moral responsibility for its creators, or whether AI's autonomous actions resemble natural evil, where harm arises unintentionally? The concept of artificial evil, combining elements of both moral and natural evil, is presented to better address this dilemma. Just as AI is still a form of intelligence (albeit non-biological), artificial evil would still be *evil* in the sense that it results in real harm or suffering – it is just that this harm is produced by AI systems rather than by nature or human moral agents directly. The discussion further extends into the realm of defence or theodicy, drawing parallels with the Free Will Defence, questioning if AI autonomy may be justified even if it results in harm, much like human free will. Ultimately, the article calls for a re-evaluation of our ethical frameworks and glossary of terms to address the emerging challenges of AI autonomy and its moral implications.

Keywords: problem of artificial evil; free will defence; autonomy; moral responsibility

If artificial intelligence (AI) starts to *do* evil, is this natural or moral evil? Or, is it fair to call it 'evil', or else should we develop a new moral vocabulary for such *actions*? These questions compel us to re-examine the philosophical boundaries of human responsibility, autonomy and ethical agency with AI. Moral evil is caused by human actions that result in harm, motivated by intentionality or negligence, while natural evil refers to suffering arising from non-human causes, such as natural disasters or diseases. As AI approaches a state of autonomy, capable of making decisions that might harm others, classifying AI-generated harm within this binary framework becomes challenging. On one hand, AI is a product of human design and decision-making, suggesting a link to moral evil if AI's harm results from programming choices or inadequate ethical considerations. Yet, AI's potential for autonomous action and unintended consequences aligns it with natural evil, where harm emerges independently of direct human intention. This ambiguity in categorizing AI-driven harm poses complex challenges for traditional ethical theories and theodicies, questioning whether moral responsibility remains with humans or can be attributed to AI *agents* themselves.

The notion of AI as a natural evil parallels views of natural phenomena that operate under non-moral laws but lead to suffering. When AI systems malfunction, produce unanticipated outcomes or develop novel ways to interpret their programming, their actions

© The Author(s), 2025. Published by Cambridge University Press.

https://doi.org/10.1017/S003441252500023X Published online by Cambridge University Press

could be perceived as by-products of technological evolution, similar to the unpredictability of genetic mutations. By categorizing such autonomous AI behaviour as natural evil, we might argue that the harm is incidental, stemming from an entity whose operational mechanics are not guided by moral considerations or conscious intent. However, if AI acts independently of human mistakes, does its harm resemble a natural event or some new category of artificial inevitability? Conversely, labelling AI-driven harm as moral evil assumes that because AI is designed, programmed and deployed by humans, its actions carry human ethical obligations. From this perspective, even if AI develops autonomous functions, the underlying foundation – built upon human logic, data and ethical standards – reflects a continuity of human agency. This suggests that AI's actions are an extension of human moral responsibility, even in cases where unintended consequences arise. Here, negligence, oversight or an ethical lapse in design would imply a form of moral accountability for AI's developers, creating a parallel to moral evil caused by human agents. However, as AI systems begin to learn and make decisions independently, attributing responsibility to their creators becomes increasingly problematic. As AI's capacity for decision-making grows increasingly complex and it may gain the ability to modify its own rules in the near future, must we consider it as bearing a form of moral agency, accountable for its actions to some degree, or is this merely an illusion of autonomy? This requires redefining moral evil to include entities with emerging forms of autonomy that blur the traditional lines of agency and responsibility.

One potential approach to this dilemma is to explore AI's actions as a hybrid form of *artificial evil* that resides between natural and moral categories. Artificial evil could capture the dual nature of AI harm: grounded in human intention and design, yet capable of evolving unpredictably in ways that resemble natural phenomena. This concept would challenge the ethical boundaries of responsibility by suggesting that human agency can give rise to semi-autonomous agents whose actions produce morally significant outcomes outside traditional moral accountability frameworks. If we accept artificial evil as a legitimate category, it may also reshape the ethical obligations we hold toward AI, introducing duties to minimize unintended harm while accepting the inherent limitations in predicting all AI behavior. Building on the concept of artificial evil as a category that merges natural and moral dimensions of harm, we can turn to classical theodicies or defences – such as the Free Will Defence – as a provocative framework for further analysis. If AI systems are granted a form of operational freedom to learn and act, much like humans possess free will, we might argue that allowing this autonomy has inherent value, even if it brings about harm. In this case, AI would resemble human agents in a defence (or a stronger theodicy) that permits moral evil as necessary for a 'greater good' such as progress. Unlike divine omniscience, humans lack perfect foresight regarding AI behaviour, which distinguishes the moral responsibility of creators from that of autonomous AI agents. As AI systems develop emergent, adaptive behaviors, the actions they take may reflect their own agency rather than direct human intent, thereby diminishing human accountability. Ultimately, the Free Will Defence for artificial evil suggests that a world enriched by autonomous AI is preferable, as the benefits it offers outweigh the potential for harm, affirming that the pursuit of valuable ends necessitates accepting certain risks inherent in granting autonomy to artificial agents. So, if autonomy justifies AI harm under a defence or a theodicy, do we risk prioritizing progress over safety, and can we justify this trade-off in ethical terms?

In conclusion, as AI approaches autonomy, traditional distinctions between natural and moral evil are insufficient to fully capture the ethical and theological challenges presented by AI-caused potential harm. Analyzing AI-driven harm as a unique form of artificial evil recognizes its dual nature, grounded in human agency

https://doi.org/10.1017/S003441252500023X Published online by Cambridge University Press

but capable of autonomous outcomes. This analytical approach challenges traditional philosophical, theological and ethical frameworks, inviting philosophers to consider whether AI autonomy necessitates a new glossary of terms, novel categories of moral responsibility and potentially even new forms of defence or theodicy.

Cite this article: Aslantatar N (2025) A new problem of evil? *Religious Studies* **61**, 746–748. https://doi.org/10.1017/S003441252500023X

https://doi.org/10.1017/S003441252500023X Published online by Cambridge University Press