Opinion

Guest Voices

A 14th-century French illumination depicts a class in philosophy. (Wikimedia Commons)

Mike Kirby

View Author Profile

David Danks

View Author Profile

The Conversation

View Author Profile

Religion News Service

View Author Profile

Join the Conversation

Send your thoughts to *Letters to the Editor*. Learn more

September 14, 2024

A self-driving taxi has no passengers, so it parks itself in a lot to reduce congestion and air pollution. After being hailed, the taxi heads out to pick up its passenger — and tragically strikes a pedestrian in a crosswalk on its way.

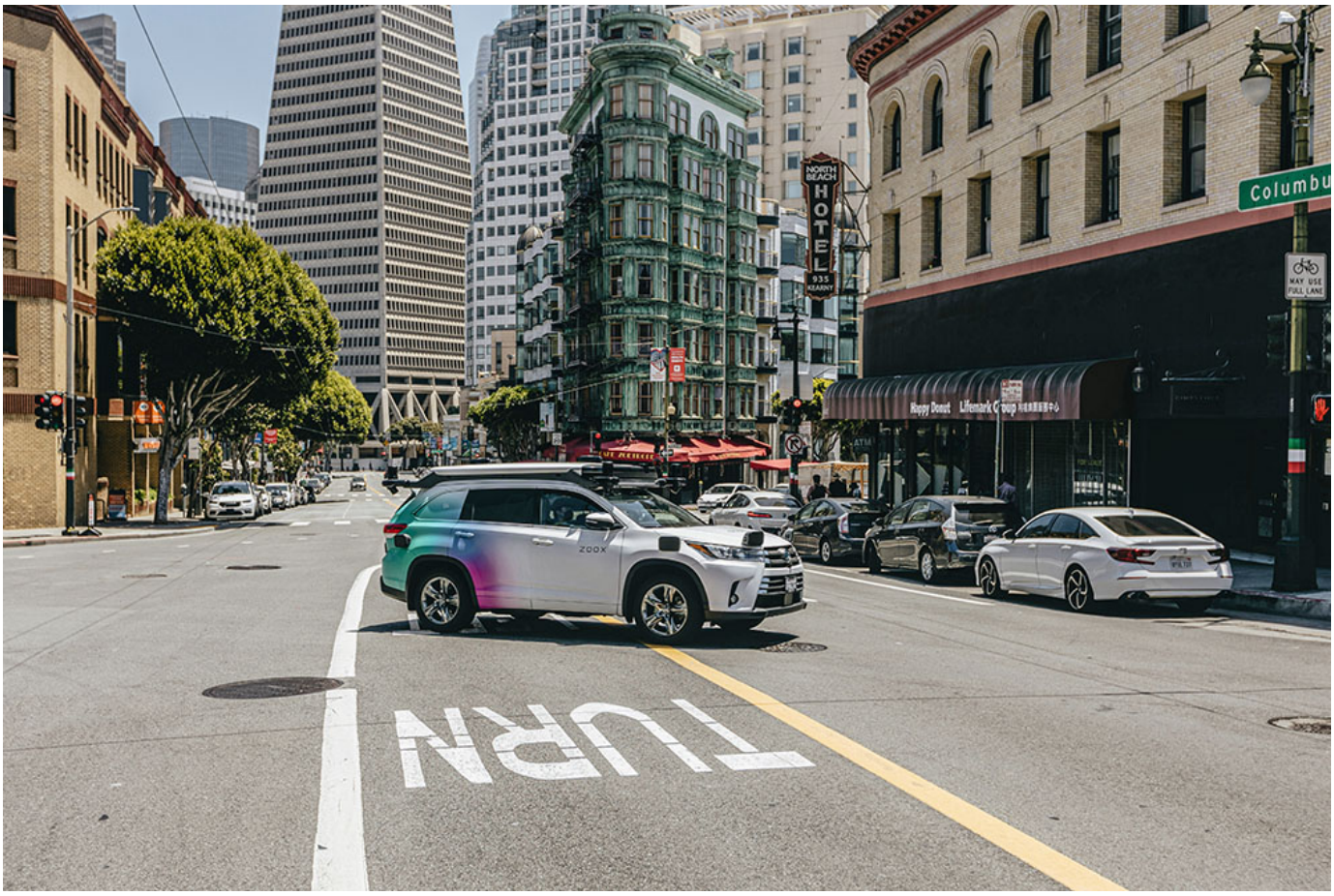
Who or what deserves praise for the car's actions to reduce congestion and air pollution? And who or what deserves blame for the pedestrian's injuries?

One possibility is the self-driving taxi's designer or developer. But in many cases, they wouldn't have been able to predict the taxi's exact behavior. In fact, people typically want artificial intelligence to discover some new or unexpected idea or plan. If we know exactly what the system should do, then we don't need to bother with AI.

Alternatively, perhaps the taxi itself should be praised and blamed. However, these kinds of AI systems are essentially deterministic: Their behavior is dictated by their code and the incoming sensor data, even if observers might struggle to predict that behavior. It seems odd to morally judge a machine that had no choice.

According to many modern philosophers, rational agents can be morally responsible for their actions, even if their actions were completely predetermined – whether by

neuroscience or by code. But most agree that the moral agent must have certain capabilities that self-driving taxis almost certainly lack, such as the ability to shape its own values. AI systems fall in an uncomfortable middle ground between moral agents and nonmoral tools.



A self-driving car is seen in San Francisco. (Unsplash/Timo Wielink)

As a society, we face a conundrum: It seems that no one, or no one thing, is morally responsible for the AI's actions — what philosophers call a responsibility gap. Present-day theories of moral responsibility simply do not seem appropriate for understanding situations involving autonomous or semi-autonomous AI systems.

If current theories will not work, then perhaps we should look to the past — to centuries-old ideas with surprising resonance today.

God and man

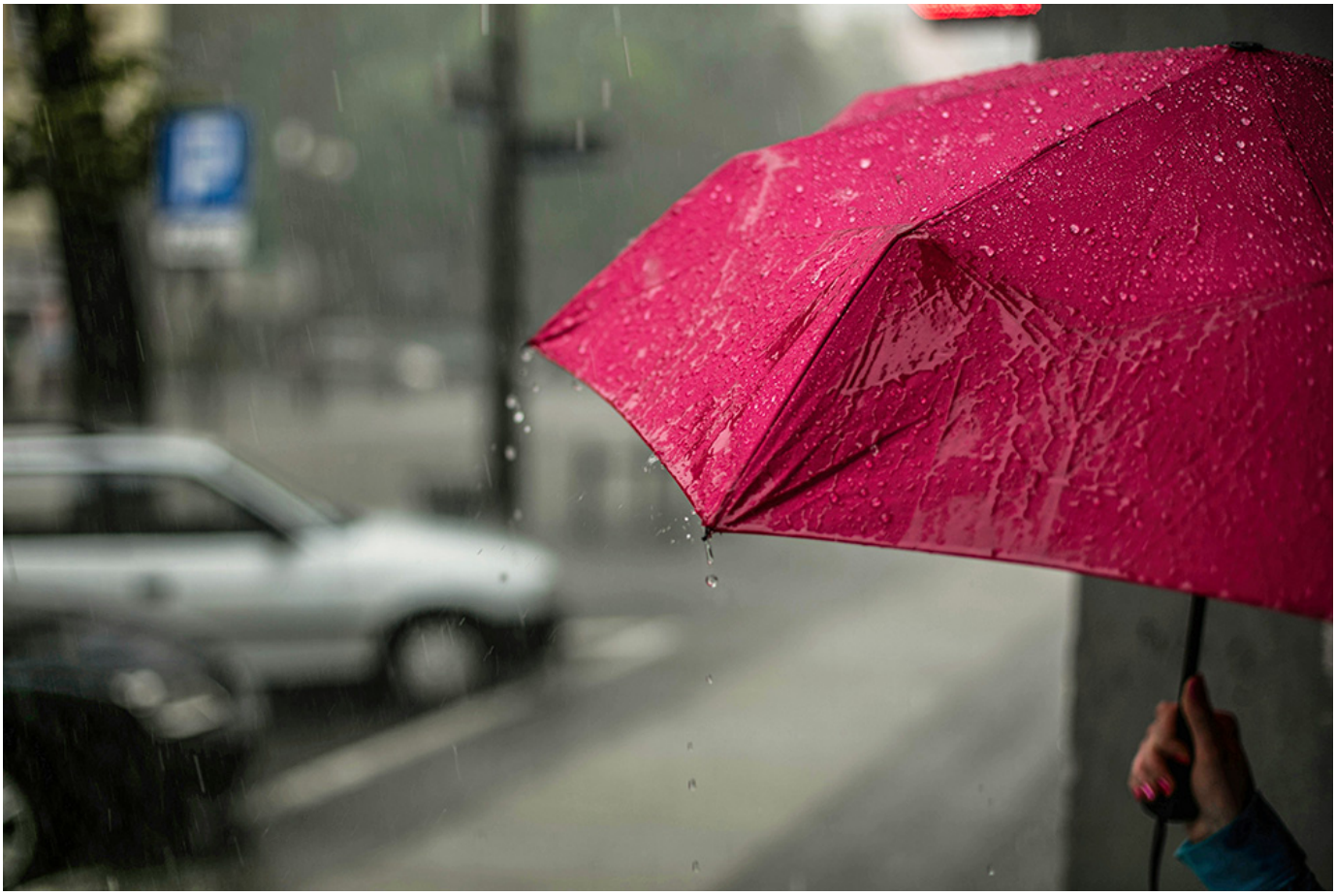
A similar question perplexed Christian theologians in the 13th and 14th centuries, from Thomas Aquinas to Duns Scotus to William of Ockham. How can people be responsible for their actions, and the results, if an omniscient God designed them — and presumably knew what they would do?

Medieval philosophers held that someone's decisions result from their will, operating on the products of their intellect. Broadly speaking, they understood human intellect as a set of mental capabilities that enable rational thought and learning.

Intellect is the rational, logical part of people's minds or souls. When two people are presented with identical situations and they both arrive at the same "rational conclusion" about how to handle things, they're using intellect. Intellect is like computer code in this way.

But the intellect doesn't always provide a unique answer. Often, the intellect provides only possibilities, and the will selects among them, whether consciously or unconsciously. Will is the act of freely choosing from among the possibilities.

As a simple example, on a rainy day, intellect dictates that I should grab an umbrella from my closet, but not which one. Will is choosing the red umbrella instead of the blue one.



(Unsplash/Erik Witsoe)

For these medieval thinkers, moral responsibility depended on what the will and the intellect each contribute. If the intellect determines that there is only one possible action, then I could not do otherwise, and so I am not morally responsible. One might even conclude that God is morally responsible, since my intellect comes from God — though the medieval theologians were very cautious about attributing responsibility to God.

On the other hand, if intellect places absolutely no constraints on my actions, then I am fully morally responsible, since will is doing all of the work. Of course, most actions involve contributions from both intellect and will — it's usually not an either/or.

In addition, other people often constrain us: from parents and teachers to judges and monarchs, especially in the medieval philosophers' days — making it even more complicated to attribute moral responsibility.

Man and AI

Clearly, the relationship between AI developers and their creations is not exactly the same as between God and humans. But as professors of philosophy and computing, we see intriguing parallels. These older ideas might help us today think through how an AI system and its designers might share moral responsibility.

AI developers are not omniscient gods, but they do provide the "intellect" of the AI system by selecting and implementing its learning methods and response capabilities. From the designer's perspective, this "intellect" constrains the AI's behavior but almost never determines its behavior completely.

Most modern AI systems are designed to learn from data and can dynamically respond to their environments. The AI will thus seem to have a "will" that chooses how to respond, within the constraints of its "intellect."

Users, managers, regulators and other parties can further constrain AI systems — analogous to how human authorities such as monarchs constrain people in the medieval philosophers' framework.

Who's responsible?

These thousand-year-old ideas map surprisingly well to the structure of moral problems involving AI systems. So let's return to our opening questions: Who or what is responsible for the benefits and harms of the self-driving taxi?

The details matter. For example, if the taxi developer explicitly writes down how the taxi should behave around crosswalks, then its actions would be entirely due to its "intellect" — and so the developers would be responsible.

However, let's say the taxi encountered situations it was not explicitly programmed for — such as if the crosswalk was painted in an unusual way, or if the taxi learned something different from data in its environment than what the developer had in mind. In cases like these, the taxi's actions would be primarily due to its "will," because the taxi selected an unexpected option — and so the taxi is responsible.

If the taxi is morally responsible, then what? Is the taxi company liable? Should the taxi's code be updated? Even the two of us do not agree about the full answer. But we think that a better understanding of moral responsibility is an important first step.

Medieval ideas are not only about medieval objects. These theologians can help ethicists today better understand the present-day challenge of AI systems — though we have only scratched the surface.

Related: Hope for an AI doomer: Laudato Si' predicted today's technology threats

Authors' note: *Mike Kirby does not work for, consult, own shares in or receive funding from any company or organization that would benefit from this article, and has disclosed no relevant affiliations beyond their academic appointment.*