AI must not encroach on human dignity

BY MICHAEL R. STRAIN

WASHINGTON, DC: Magnifica Humanitas, Pope Leo XIV's encyclical on artificial intelligence (AI), has accelerated the debate on the fundamental nature of the technology. Could AI have a conscience? How does it think? What does its rapid advance portend?

In the encyclical's most provocative section, Pope Leo argues that AI systems "do not undergo experiences", "do not feel joy or pain" and do not "have a moral conscience", since they do not judge good and evil, grasp the ultimate meaning of situations, or bear responsibility for consequences."

He continues: "They may imitate language, behaviour and analytical skills, or even simulate empathy and understanding, but they do not understand what they produce."

Whether you agree with the Pope's interpretation depends, in part, on your answers to foundational questions about what it means to be human and the nature of consciousness.

Those formed by the Biblical tradition will argue that being human means having been created in God's image and likeness. Catholics—including me and, of course, the Pope—believe that God is inherently relational, one essence in three persons. To glimpse part of the truth of this great mystery, Saint Augustine, one of Christianity's foremost theologians, proposed an analogy in the 5th century: the Father as the great and eternal mind; the Son as the eternally begotten, perfect self-knowledge of the Father; and the Spirit as perfect self-love.

Humans, created in the divine image, share with God the ability to form an understanding of ourselves through the act of self-knowledge. And, like God, we can love ourselves. In this way, our inner life—our consciousness—is relational, like God's.

AI does not come anywhere close to meeting these criteria. Will it ever? I am doubtful.

Many disagree. At the presentation of the encyclical in Vatican City, Anthropic co-founder Chris Olah claimed that his research team, which studies the internal structure of these models, has found "evidence of introspection" and "internal states that functionally mirror joy, satisfaction, fear, grief and unease." While Olah admitted that he does not know what that means, he posits that "it warrants ongoing discernment."

Well, I would like to see some evidence. The burden of proof that AI can engage in introspection—can form a self-image, as humans can—and experience emotions is on the believer, not the sceptic.

What about cognition? The Pope argues that AI tools "merely imitate certain functions of human intelligence," and that they are "entirely tied to data processing." Many technologists disagree. But Pope Leo is right to stress a distinction between AI's data processing and human cognition. Generative AI tools excel at pattern recognition. The statistical models that power them use an inductive approach, relying on huge data sets and massive computing power to imbue AI systems with tacit knowledge.

This differs from human learning in important ways. We do not train our minds on enormous quantities of data with the goal of using these inputs to predict outputs. Instead, we theorise and hypothesise based on a small number of examples, often from our own experience. We are tribal, learning from our families and communities, often adopting the conclusions of those around us.

In Magnifica Humanitas, the Pope seeks to shift attention away from the marvels and terrors of AI and toward the magnificence of humanity. With all eyes now on the technology, this message is needed and welcome. AI tools are impressive. But they are pedestrian compared to the grandeur of a human being.

Michael R Strain, Senior Fellow; Paul F Oreffice Chair in Political Economy; Director of Economic Policy Studies, Project