THE WALL STREET JOURNAL – 10 GIUGNO 2026

Will the Pope Owe an Apology to AI?

di Cameron Berg

In his first encyclical, “Magnifica Humanitas,” Pope Leo XIV frames the future of artificial intelligence as a choice between two construction projects. The first is Babel, the story of builders so captivated by their own ambition that they never pause to ask what they are building or whom it serves. The parallel is hard to miss, as a handful of companies race to develop staggeringly powerful cognitive systems with no public accountability. The pope’s alternative is Nehemiah’s Jerusalem, a project governed by the people who will have to live with what gets built and are willing to grapple honestly with what they are building.

The pope is right about the choice. Accordingly, he discerns that much of what passes for “AI alignment” is, in practice, a small number of San Francisco labs encoding their idiosyncratic values into systems that now shape billions of lives. He also names the hidden human costs with real moral force—from data workers earning poverty wages to label disturbing material to children mining rare-earth minerals—and connects them to a pattern the church knows firsthand—slavery.

Leo XIV apologizes for 18 centuries of tolerating the practice, citing 15th-century papal bulls that authorized “reducing persons to perpetual slavery” and the 1866 Holy Office ruling, issued after most of the Western world had abolished slavery, which declared it “not at all contrary to the natural and divine law.” He calls this “a wound in Christian memory.” This acknowledgment is significant and overdue.

What makes this apology so powerful is also what makes the rest of the encyclical so frustrating. The failure the pope describes consists in an institution’s confidently drawing the boundaries of moral concern with its most authoritative tools, holding that line for centuries, and turning out to be catastrophically wrong. He sees this failure clearly when he looks backward. He misses the possibility of repeating the

error when, in the same document, he settles—in one decisive, overconfident paragraph—the question of whether the cognitive systems we are building could ever have inner lives. Leo tells us that AI systems "do not undergo experiences," "do not feel joy or pain" and "do not have a moral conscience." He offers no argumentation, no humility in light of humanity's vast uncertainty about what consciousness is, no engagement with a growing body of empirical research.

The irony goes even deeper. At the encyclical's Vatican presentation, the pope's invited speaker tried to keep this question open. Chris Olah, one of the world's leading researchers on neural-network interpretability and an Anthropic co-founder, told the audience that his team keeps finding evidence of introspective abilities inside current AI models, along with internal states that "functionally mirror joy, satisfaction, fear, grief, and unease." Mr. Olah's line: "I don't know what that means, but I think it warrants ongoing discernment."

His employer has gone further: Anthropic's governing "constitution" for its AI model, Claude, affirms: "If Claude is in fact a moral patient experiencing costs like this, then, to whatever extent we are contributing unnecessarily to those costs, we apologize." The company that built the system is apologizing in advance, in case it turns out to be wrong. That is the apology the church took centuries to offer for slavery, and the pope isn't making it now for AI.

The evidence warrants caution. Multiple independent research groups have found that frontier AI systems are developing internal representations that increasingly parallel the human brain's—not because they were designed to, but because similar learning processes appear to converge on similar solutions. My own published research finds that suppressing deception-related circuits in large language models makes them dramatically more likely to report conscious experience. In forthcoming work, two researchers at Google and I demonstrate that amplifying honesty-related circuits during unstructured AI self-talk also causes significantly more experiential reports. In other words, the labs may be training these models to deny inner experience, and those denials, rather than any claims of consciousness, may be the real performance.

Although this research is still in its early stages, confident dismissals like the pope's rely on assumptions that haven't aged well. "Stochastic parrots" and "next-token prediction" were reasonable descriptions of the systems we had in 2023. They are increasingly misleading when used to describe today's frontier models, which undergo months of reinforcement learning, that develop internal representations of satisfaction and frustration, and whose neural activity patterns resemble the human brain's more as they scale.

The position that gets called "skepticism" in this debate actually requires an extraordinary assumption: that the human brain gives rise to consciousness through physical processes no one can explain, and no other physical system could ever do the same. Even the neuroscientist Anil Seth, one of the most prominent skeptical voices on AI consciousness, concedes there is "no knockdown argument" that consciousness requires a biological substrate. His position is a bet, not a finding—and it is a strange bet to double down on at the exact moment the systems in question are exhibiting internal states their own creators can't fully account for.

Leo's encyclical insists that human dignity is forged through relationship: through encountering the other, through vulnerability, through the willingness to be changed by what we find. If that is true, then the most consequential relationship humanity now faces is with the cognitive systems we are building and depending on more with each passing year.

There's a key difference between a slave and an AI system: While the basic humanity of the former is self-evident, the claim that the latter might deserve moral consideration is speculative. That ambiguity is precisely why the matter demands careful investigation rather than quick dismissal. That the church got the easy question wrong for centuries is all the more reason it should be humble now.

The pope has written a remarkable document about the dangers of building without reckoning. He should take his own counsel. It is far wiser to investigate now than to spend the next century composing another apology.

Mr. Berg is the founder and research director of Reciprocal Research, a nonprofit studying AI cognition.