diritto & religioni
Supplemento telematico – ISSN 1970-5301
https://www.rivistadirittocrcligioni.com
LUIGI PELLEGRINI EDITORE

Too Much Machine.
The Anthropic Case Between Faith, Algorithms, and Pluralism.

(3 July 2026)

In March 2026, the artificial intelligence company Anthropic brought together fifteen Christian leaders in San Francisco, announcing that it would subsequently invite representatives of other Abrahamic faiths, Hinduism, and various philosophical[1] traditions in order to develop a suitable method for providing ethical guidance to its AI model, Claude. This initiative forms part of a broader project already underway. From the outset, the company's founders[2] sought to orient the operation of their artificial intelligence according to the principles of AI Safety[3] by drafting an "ethical constitution[4]".

This raises a fundamental question: is it appropriate to provide an artificial neural network with ethical education?

To offer a preliminary answer, it must first be acknowledged that even when an AI system officially claims to operate without predetermined moral frameworks and presents itself as ethically neutral, some form of value orientation is almost inevitable.

The Anthropic case, however, presents distinctive features. The particular attention the American company has devoted to the Christian world—and especially to Catholicism—is clearly reflected in the document outlining Claude's principles. Alongside philosophers and AI experts, the drafting process involved the Irish priest Father Brendan McGuire[5] and Archbishop Paul Desmond Tighe, Secretary of the Section for Culture of the Dicastery for Culture and Education of the Roman Curia[6].

Through these choices, the corporation appears to have aligned its corporate ethical framework with theological and moral values characteristic of the Catholic tradition. Further confirmation can be found in its participation—alongside numerous other companies—in the 2020 "*Rome Call for AI Ethics*[7]" and, even more significantly, in the invitation extended exclusively to Anthropic to the Vatican for the presentation of Pope Leo XIV's encyclical *Magnifica Humanitas*.

On that occasion, Anthropic co-founder Chris Olah[8] addressed Vatican authorities, drawing attention to the difficulties AI developers face in interpreting neural network structures and highlighting the existence of significant and troubling forms of opacity[9].

For its part, the Holy See has long demonstrated a particular commitment to the ethical implications of intelligent technologies. Already in *Antiqua et Nova*, the document issued by the Dicastery for the Doctrine of the Faith in April 2025, reflections were offered on AI's inability to exercise genuine

---

[1] https://www.anthropic.com/news/widening-conversation-ai?ref=nextomoro.com
[2] https://ainews.it/chi-sono-dario-e-daniela-amodei-gli-italoamericani-delletica-tech/
[3] Modello da adottare ove si vogliano garantire sicurezza digitale e limiti dell'IA; https://www.ibm.com/it-it/think/topics/ai-safety
[4] https://www.anthropic.com/constitution
[5] https://www.omnesmag.com/it/focus/chi-e-brendan-mcguire-parroco-e-consulente-aziendale-in-ia/
[6] https://www.dce.va/it/authors/mons-tighe-paul-desmond.html
[7] https://www.vatican.va/roman_curia/pontifical_academies/acdlife/documents/rc_pont-acd_life_doc_20202228_rome-call-for-ai-ethics_en.pdf
[8] https://www.anthropic.com/news/chris-olah-pope-leo-encyclical
[9] https://www.youtube.com/watch?v=CQdzY_3apNQ

discernment or achieve a truly human understanding of reality[10]. Along similar lines, Pope Leo XIV argued in the encyclical *Magnifica Humanitas*: «We cannot simply call for the moralization of machines—the so-called alignment of AI with human values—without having the courage to establish a further condition: the possibility of openly debating the ethical framework to be adopted, subjecting it to shared criteria of social justice[11]».

Leaving aside the practical feasibility of any project aimed at ethically aligning artificial intelligence[12]—a goal that itself raises considerable technical uncertainties—it is worth considering whether entrusting the selection of ethical principles to representatives of theological thought, whose role would be to shape the outputs of a system intended for global use by religiously and ethically diverse users, might transform AI into a potential instrument of cultural hegemony. Would the creation of ethical codes based on particular conceptions of social justice ultimately result in an implicit restriction of freedom of expression?

Although Anthropic has pledged to include representatives of other religious and philosophical traditions, its evident proximity to the Catholic world and the promise of broader consultation only at a later stage may reasonably be interpreted as signs of a religiously unbalanced process of ethical alignment. Moreover, even if grounded in theological knowledge, would a religiously inspired ethical framework for AI be sufficient to guarantee genuine pluralism of values?

Beyond these questions, closer attention to the actual functioning of artificial intelligence may challenge the assumption that ethically aligning AI is inherently desirable. Particularly relevant in this respect is the argument advanced by scholars who propose describing AI instead as "Artificial Communication[13]". This perspective critically emphasizes the risks of excessively anthropomorphizing artificial intelligence.

AI, according to this view, does not think as human beings do; rather, it processes and computes information through algorithms[14]. The mechanisms of "machine learning[15]" and "deep learning[16]" are now well known. Their functioning depends upon databases, which in turn rely on the information and data generated by digital users. Chatbots—described as genuine «competent communicative partners[17]»—produce outputs according to the digital ecosystem from which they retrieve, process, and integrate information. Algorithms are capable of carrying out these operations at a speed far beyond that of the human brain[18].

---

[10]https://www.vatican.va/roman_curia/congregations/cfaith/documents/rc_ddf_doc_20250128_antiqua-et-nova_it.html cfr. 32; in particolare ove si specifica che «tramite l'IA e i mezzi tecnologici non si possono risolvere tutti i problemi dell'uomo, altrimenti rischieremmo di trovarci dinanzi ad una nuova idolatria e a considerarle come un nuovo dio».

[11] https://www.vatican.va/content/leo-xiv/it/encyclicals/documents/20260515-magnifica-humanitas.html cfr.107

[12] Vedi: https://www.avvenire.it/rubriche/artifici/le-intelligenze-artificiali-sono-figlie-di-dio-la-silicon-valley-cerca-le-chiese_107131

[13] ELENA ESPOSITO, *Comunicazione artificiale. Come gli algoritmi producono intelligenza sociale*, Bocconi University Press, Milano, 2022, p. 97

[14] https://www.newmediaeuropeanpress.eu/2025/11/27/linvenzione-dellalgebra-lalgoritmo-di-al-khwarizmi/

[15] https://www.ibm.com/it-it/think/topics/machine-learning

[16] https://www.ibm.com/it-it/think/topics/deep-learning

[17]ELENA ESPOSITO, *Comunicazione artificiale. Come gli algoritmi producono intelligenza sociale*, Bocconi University Press, Milano, 2022, p. 97

[18] E ne è un esempio la "mossa 37" durante la partita di "*Go*" del 2016 tra l'algoritmo di "*AlphaGo*" e il campione mondiale Lee Sedol; https://www.repubblica.it/tecnologia/2026/03/26/news/google_deepmind_ia_alphago_lee_sedol_alphafold-425247945/

In other words, it is human interaction itself that contributes to the continuous modification of the algorithm. Furthermore, the fact that users receive responses contingent upon the questions they ask[19]—and the manner in which they formulate them—also demonstrates that no algorithm simply invents its outputs from nothing.

Ultimately, the possibility that the future development of AI may come to rely upon uniform ethical models producing predetermined answers should not be underestimated.

Certainly, the risks associated with the improper, violent, or discriminatory use of artificial intelligence cannot be ignored. Yet increasing prohibitions through stricter regulation or through forms of algorithmic ethical alignment based on standardized religious frameworks may ultimately prove ineffective—or even counterproductive. By contrast, strengthening fact-checking mechanisms and implementing algorithmic models specifically designed to counter users' passivity[20] could help reduce social and political polarization[21] without undermining the diversity of information or the exchange of competing perspectives[22].

MARCO SPINA

[19] ELENA ESPOSITO, *Comunicazione artificiale. Come gli algoritmi producono intelligenza sociale*, Bocconi University Press, Milano, 2022, p.29

[20] ELENA ESPOSITO**,** *Comunicazione artificiale. Come gli algoritmi producono intelligenza sociale*, Bocconi University Press, Milano, 2022, p. 97

[21] Si veda per esempio Flipfeed, https://www.media.mit.edu/projects/flipfeed/overview

[22] In alcuni si introducono nei news feed contenuti personalizzati secondo prospettive politiche che l'algoritmo identifica come contrarie all'ideologia dell'utente; ELENA ESPOSITO, *Comunicazione artificiale. Come gli algoritmi producono intelligenza sociale*, Bocconi University Press, Milano, 2022, p. 97