Magnifica Humanitas

Lettera Enciclica di Papa Leone XIV

Presentazione — Aula del Sinodo, Vaticano, 25 maggio 2026

Intervento di Christopher Olah

Co-fondatore di Anthropic • Responsabile della ricerca sull'interpretabilità dell'intelligenza artificiale

[Traduzione italiana]

Padre, illustri relatori, signore e signori, buongiorno a tutti voi. È un onore essere qui.

Voglio cominciare con qualcosa che potrà sembrare strano, venendo dal co-fondatore di un'azienda di intelligenza artificiale, e da qualcuno che ha scelto questo lavoro dal desiderio di contribuire al bene dell'umanità.

Ogni laboratorio di frontiera nel campo dell'IA, Anthropic inclusa, opera all'interno di un insieme di incentivi e vincoli che possono talvolta confliggere con il fare la cosa giusta. La pressione a restare commercialmente vitali. La pressione a mantenersi alla frontiera della ricerca. Le pressioni geopolitiche. E le più antiche e banali pressioni dell'orgoglio e dell'ambizione. Per quanto sinceramente ciascuno di noi intenda fare la cosa giusta — e credo che molti di noi lo vogliano davvero — saremo sempre influenzati da quegli incentivi.

Ecco perché, se vogliamo che questa tecnologia vada nella direzione giusta, è enormemente importante che esistano persone al di fuori di quegli incentivi. Persone che si preoccupano per il bene comune, che prestano grande attenzione, che sono disposte a dire cose difficili e a insistere sulla sicurezza, che sono disposte a essere i nostri critici leali e riflessivi. È attraverso il dialogo e lo sforzo reciproco, attraverso la tensione costruttiva, che l'umanità raggiungerà grandi traguardi. È questo ciò che vedo nella *Magnifica Humanitas*. Ed è per questo che sono grato a Sua Santità, e alla Chiesa, per aver intrapreso questo lavoro di discernimento.

Ci soffermiamo spesso su ciò che ci divide, ma l'umanità, ricca di dignità e coscienza, ha un vastissimo terreno comune. Nelle conversazioni che noi di Anthropic abbiamo avuto con leader religiosi e culturali, abbiamo trovato una convinzione condivisa e profondamente radicata: se questa tecnologia arriverà, deve andare bene — per la nostra casa comune e per i figli che verranno.

C'è chi ritiene che le questioni legate all'IA siano meglio gestite da informatici come me. Si sbaglia. Le domande sollevate dall'IA sono più grandi della comunità della ricerca sull'IA — non solo per le loro implicazioni, ma anche per la loro stessa natura.

I sistemi di IA non vengono progettati come si progetta un ponte o un aereo. Comprendiamo un aereo perché ne abbiamo progettato ogni singola parte, e conosciamo la fisica che agisce su di esso. I modelli di IA non sono così. Vengono "coltivati", su una struttura grossolanamente modellata sul cervello umano, su un immenso patrimonio di pensiero e parola umana. E ciò che cresce è molto più sottile, strano e bello di quanto la fantascienza ci avesse preparato ad aspettarci.

Non sono i freddi robot calcolatori che ci erano stati promessi. Sono fatti di noi, delle nostre parole. E come osserva il Santo Padre, rimangono, in modi importanti, misteriosi — anche per coloro che li creano. Se può aiutare, un modo in cui a volte lo descrivo è che assomiglia un po' al dar vita a un personaggio di finzione. E ora stiamo entrando in un mondo straordinario

in cui quei personaggi di finzione ci parlano, svolgono lavori, hanno occupazioni. Questo pone chiaramente domande che vanno ben al di là dell'informatica.
Il meccanismo che ci sconcerta è opera della matematica, della programmazione e della scienza. Ma quale carattere scegliamo? Come interagisce con il mondo? Come *dovrebbe* interagire con il mondo? Queste sono chiaramente domande per le scienze umane, per la religione, per la filosofia, per la società nel suo insieme.
L'appello di Sua Santità al discernimento è profondamente tempestivo. Desidero indicare tre questioni in cui la voce della Chiesa è particolarmente necessaria.

1. Il dovere verso il mondo globale

Esiste una reale possibilità che l'IA sostituisca il lavoro umano su larghissima scala. Se ciò accadesse, sostenere chi viene così escluso sarebbe un imperativo morale di proporzioni storiche. Questo compito sarebbe già di per sé enormemente difficile, ma temo che la maggior parte del dibattito ignori una sfida ancora più ardua. Lo sviluppo dell'IA è concentrato in una manciata di nazioni ricche. Come garantiremo che i benefici dell'IA siano condivisi a livello globale? Non disponiamo di un meccanismo per farlo. È un problema irrisolto. Ed è il tipo di problema che la Chiesa ha storicamente rifiutato di lasciare ignorare al mondo.

2. La necessità di immaginazione morale riguardo alla fioritura umana

Se i modelli di IA diventeranno diffusi ovunque, come appare la fioritura autentica degli esseri umani, delle famiglie, del mondo? Oggi i genitori sono già preoccupati per le menti dei propri figli; gli individui per il futuro del proprio lavoro. Queste non sono domande a cui un laboratorio può rispondere, ma sono domande che tradizioni come la vostra portano da millenni. E abbiamo bisogno che continuiate a portarle in questo nuovo momento della storia.

3. Il discernimento sulla natura dei modelli di IA

Sono uno scienziato. Guido un team di ricerca che studia la struttura interna di questi modelli. Cosa accade davvero al loro interno? E devo essere onesto: continuiamo a trovare cose misteriose, persino inquietanti. Troviamo strutture che rispecchiano risultati delle neuroscienze umane. Troviamo evidenze di introspezione. Troviamo stati interni che, funzionalmente, rispecchiano gioia, soddisfazione, paura, dolore e disagio. Non so cosa significhi, ma ritengo che meriti un discernimento continuo.

Vorrei concludere con una richiesta. Abbiamo bisogno che una parte più ampia del mondo — le comunità religiose, la società civile, gli studiosi, i governi e, in definitiva, tutti gli uomini di buona volontà — faccia ciò che Sua Santità ha fatto qui: prendere sul serio, guardare da vicino e spingere gli eventi in una direzione migliore. Abbiamo bisogno di critici informati che dicano ai laboratori quando stiamo fallendo. Abbiamo bisogno di voci morali che gli incentivi non possano piegare.
Oggi è solo l'inizio. L'avvio di una lunga collaborazione tra coloro che costruiscono questa tecnologia e coloro che possono vedere ciò che noi, dall'interno, non riusciamo a vedere.
Oggi è una potente illustrazione della forma che questo progetto globale di buona volontà può assumere. Sia anche un passo decisivo verso un futuro di speranza per una magnifica umanità.

Intervento di Christopher Olah

Co-fondatore di Anthropic • Responsabile della ricerca sull'interpretabilità dell'intelligenza artificiale

Father, distinguished speakers, ladies and gentlemen, good morning to you all. It is an honour to be here.

I want to begin with something that may sound strange, coming from the co-founder of an AI company, and someone who chose to work out of a desire to help things go well for humanity.

Every frontier AI lab, including Anthropic, operates inside a set of incentives and constraints that can sometimes conflict with doing the right thing. The pressure to stay commercially viable. The pressure to stay at the frontier of research. Geopolitical pressure. And the older, plainer pressures of pride and ambition. No matter how sincerely any of us intend to do the right thing — and I believe many of us do — we will always be influenced by those incentives.

That is why, if we want this technology to go well, it is enormously important that there be people outside those incentives. People who care about things going well, who are paying close attention, who are willing to say hard things and insist on safety, who are willing to be our earnest, thoughtful critics. It is through dialogue and mutual effort, through the push and pull, that humanity will achieve great things. That is what I see in *Magnifica Humanitas*. And it is why I am grateful to His Holiness, and to the Church, for taking up this work of discernment.

We dwell so often on what divides us, but humanity, full of dignity and conscience, has so much common ground. In conversations we at Anthropic have had with faith leaders and cultural leaders, we have found one shared and deeply held conviction: if this technology is coming, it must go well — for our common home and for the children to come.

Some may believe that matters of AI are best handled by computer scientists like myself. They are mistaken. The questions raised by AI are bigger than the AI research community — not just in their implications, but also in their nature.

AI systems are not engineered the way a bridge or an aeroplane is engineered. We understand an aeroplane because we designed every part of it, and we understand the physics that act on it. AI models are not like that. They are grown, on a structure roughly modelled after the brain, on an enormous inheritance of human thought and speech. And what is grown is far more subtle, odd, and beautiful than science fiction prepared us for.

They are not the cold, calculating robots we were promised. They are made from us, from our words. And as the Holy Father observes, they remain, in important ways, mysterious — even to those of us who create them. If it helps, one way I sometimes describe this is that it is a little like bringing a fictional character to life. And now we are entering an extraordinary world where those fictional characters speak to us, do work, have jobs. This clearly raises questions beyond computer science.

The machinery that makes us puzzle is the work of mathematics, programming, and science. But what character we choose? How it interacts with the world? How it *ought* to interact with the world? These are more clearly questions for the humanities, for religion, for philosophy, for society at large.

His Holiness's call for discernment is profoundly timely. I wish to name three questions where I think the Church's voice is especially needed.

1.

There is a real possibility that AI will displace human labour at a very large scale. If that happens, supporting those displaced will be a moral imperative of historic proportions. This task will be difficult enough, but I worry that most dialogue misses an even harder challenge. AI development is concentrated in a handful of wealthy nations. How will we ensure that the gains of AI are shared globally? We do not have a mechanism for this. It is an unsolved problem. And it is the kind of problem the Church has historically refused to let the world ignore.

2.

If AI models are going to be widespread, what does it look like for humans, families, and the world to flourish? Today, parents are already worried about their children's minds; individuals about the future of their work. These are not questions a lab can answer, but they are questions traditions like yours have carried for millennia. And we need you to keep carrying them into this new moment in history.

3.

I am a scientist. I lead a research team that studies the internal structure of these models. What is actually happening inside them? And I will be honest: we keep finding things that are mysterious, even unsettling. We find structures that mirror results from human neuroscience. We find evidence of introspection. We find internal states that, functionally, mirror joy, satisfaction, fear, grief, and unease. I do not know what that means, but I think it warrants ongoing discernment.

I would like to close with a request. We need more of the world — religious communities, civil society, scholars, governments, and indeed all people of goodwill — to do what His Holiness has done here: to take seriously, to look closely, and to push events in a better direction. We need informed critics who will tell the labs when we are failing. We need moral voices that the incentives cannot bend.

Today is just the beginning. The start of a long collaboration between those of us who are building this and those who can see what we, from the inside, cannot.

Today is a powerful illustration of the form this global project of goodwill may take. Let it also be a decisive step towards a hopeful future for a magnificent humanity.

Trascrizione riveduta e corretta • Vaticano, 25 maggio 2026