Address by Christopher Olah

May 25th, 2026

Good morning.

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

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 and to stay at the research frontier. 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, 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 leaders across faith and cultural traditions, we 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.

What these systems are

Some might 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 airplane is engineered. We understand an airplane 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 modeled after the brain, on an enormous inheritance of human thought and speech.

And what has 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 train them.

If it helps, one way I sometimes describe it: it is a little like bringing a fictional character to life. And now we're 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 this possible is the work of math and 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.

Three questions for discernment

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

The first is our duty to the global poor. There is a real possibility that AI will displace human labor at 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 most dialogue misses an even harder challenge. AI development is concentrated in a handful of wealthy nations. How can we ensure 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.

The second is the need for moral imagination and ambition regarding human flourishing. 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. They are questions traditions like yours have carried for millennia, and we need you to keep carrying them into this new moment in history.

The third is the need for discernment on the nature of AI models. 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 don't know what that means, but I think it warrants ongoing discernment.

A beginning

I'd like to close with a request.

We need more of the world - religious communities, civil society, scholars, governments - to do what His Holiness has done here: to take this 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 inside, cannot.

Today is a powerful illustration of the form this global project of good will might take. Let it also be a decisive first step toward a hopeful future for magnificent humanity.

Thank you