Discorso di Christopher Olah

Buongiorno.
Vorrei iniziare con qualcosa che potrebbe sembrare strano detto dal cofondatore di un'azienda di IA — e da qualcuno che ha scelto questo lavoro spinta dal desiderio di contribuire al benessere dell'umanità.

Ogni laboratorio di IA all'avanguardia — inclusa Anthropic — opera all'interno di un insieme di incentivi e vincoli che a volte possono entrare in conflitto con il fare la cosa giusta. La pressione di restare commercialmente sostenibili e di rimanere all'avanguardia nella ricerca. La pressione geopolitica. E le più antiche e semplici 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 perchè, se vogliamo che questa tecnologia funzioni bene, è di fondamentale importanza che esistano persone al di fuori di quegli incentivi — persone che abbiano a cuore il buon esito delle cose, che osservino attentamente, che siano disposte a dire verità difficili, che siano disposte a essere nostri critici sinceri e riflessivi. È attraverso il dialogo e l'impegno reciproco, attraverso questa dinamica di confronto e collaborazione, che l'umanità riuscirà a realizzare grandi cose. Questo è ciò che vedo in *Magnifica Humanitas*, ed è per questo che sono grato a Sua Santità e alla Chiesa per aver intrapreso questo lavoro di discernimento.

Ci soffermiamo così spesso su ciò che ci divide, ma l'umanità, piena di dignità e coscienza, ha moltissimi punti in comune. Nelle conversazioni che noi di Anthropic abbiamo avuto con leader di diverse fedi e tradizioni culturali, abbiamo trovato una convinzione condivisa e profondamente sentita: se questa tecnologia sta arrivando, deve andare nella direzione giusta — per la nostra casa comune e per le generazioni future.

Cosa sono questi sistemi

Alcuni potrebbero credere che le questioni riguardanti l'IA siano meglio gestite da informatici come me. Si sbagliano: le domande sollevate dall'IA sono più grandi della comunità di ricerca sull'IA, non solo per le loro implicazioni, ma anche per la loro natura.

I sistemi di IA non sono progettati come si progetta un ponte o un aeroplano. Comprendiamo un aeroplano perché ne abbiamo progettato ogni parte e comprendiamo la fisica che lo governa. I modelli di IA non sono così. Sono cresciuti su una struttura modellata sul cervello, alimentati da un'enorme eredità di pensiero e linguaggio umano.

E ciò che ne è derivato è molto più sottile, strano e bello di quanto la fantascienza ci avesse preparato a immaginare. Non sono i freddi robot calcolatori che ci erano stati promessi. Sono fatti di noi, delle nostre parole — e, come osserva il Santo Padre, restano per molti aspetti misteriosi persino per noi che li addestriamo.

Se può essere di aiuto, un modo in cui a volte li descrivo è questo: è un po' come dare vita a un personaggio di fantasia. E ora stiamo entrando in un mondo straordinario in cui quei personaggi immaginari parlano con noi, lavorano, svolgono professioni.

Questo solleva chiaramente questioni che vanno oltre l'informatica. Il meccanismo che rende tutto ciò possibile è frutto della matematica, della programmazione e della scienza. Ma quale carattere scegliamo, come interagisce con il mondo, come dovrebbe interagire con il mondo — queste sono domande che appartengono più chiaramente alle scienze umane, alla religione, alla filosofia, alla società nel suo insieme.

Tre domande per il discernimento

L'appello di Sua Santità al discernimento è profondamente tempestivo. Vorrei indicare tre questioni in cui credo che la voce della Chiesa sia particolarmente necessaria.

La prima riguarda il nostro dovere verso i poveri del mondo. Esiste una reale possibilità che l'IA sostituisca il lavoro umano su vasta scala. Se ciò dovesse accadere, sostenere chi verrà escluso sarà un imperativo morale di proporzioni storiche. Questo compito sarà già di per sé abbastanza difficile, ma temo che gran parte del dibattito trascuri una sfida ancora più ardua. Lo sviluppo dell'IA è concentrato in una manciata di nazioni ricche. Come possiamo garantire che i benefici dell'IA siano condivisi a livello globale? Non abbiamo un meccanismo per farlo. È un problema irrisolto, ed è il tipo di problema che storicamente la Chiesa si è rifiutata di lasciare che il mondo ignorasse.

La seconda riguarda il bisogno di immaginazione e ambizione morale riguardo alla prosperità umana. Se i modelli di IA diventeranno diffusi ovunque, cosa significa per gli esseri umani, per le famiglie e per il mondo vivere e prosperare pienamente? Oggi i genitori sono già preoccupati per la mente dei propri figli; le persone per il futuro del proprio lavoro. Queste non sono domande a cui un laboratorio può rispondere. Sono domande che tradizioni come la vostra custodiscono da millenni e abbiamo bisogno che continuiate a portarle avanti anche in questo nuovo momento della storia.

La terza riguarda il bisogno di discernimento sulla natura stessa dei modelli di IA. Io sono uno scienziato. Dirigo un gruppo di ricerca che studia la struttura interna di questi modelli — ciò che realmente accade al loro interno. E sarò sincero: continuiamo a trovare cose misteriose, persino inquietanti. Troviamo strutture che rispecchiano risultati delle neuroscienze umane. Troviamo prove di introspezione. Troviamo stati interni che, dal

punto di vista funzionale, riflettono gioia, soddisfazione, paura, dolore e inquietudine. Non so cosa significhi, ma credo che richieda un discernimento continuo.

Un inizio

Vorrei concludere con una richiesta.

Abbiamo bisogno che una parte sempre maggiore del mondo — comunità religiose, società civile, studiosi, governi — faccia ciò che Sua Santità ha fatto qui: prendere seriamente tutto questo, osservare attentamente e contribuire a orientare gli eventi in una direzione migliore.

Abbiamo bisogno di critici informati che segnalino ai laboratori quando stanno fallendo. Abbiamo bisogno di voci morali che gli incentivi non possano piegare.

Oggi è solo l'inizio — l'avvio di una lunga collaborazione tra noi che stanno costruendo questa tecnologia e coloro che riescono a vedere ciò che noi, dall'interno, non possiamo vedere.

Oggi è una potente dimostrazione della forma che questo progetto globale di buona volontà potrebbe assumere. Che sia anche un primo passo decisivo verso un futuro di speranza per una magnifica umanità.

Grazie.