Discours intégral (en traduction automatique) de Christopher Olah pour la présentation de Magnifica Humanitas.

Bonjour.

J’aimerais commencer par quelque chose qui peut paraître étrange venant du cofondateur d’une entreprise d’intelligence artificielle, et de quelqu’un qui a choisi ce travail par désir de contribuer au bien-être de l’humanité .

Tous les laboratoires de pointe en intelligence artificielle, y compris Anthropic, fonctionnent selon un ensemble d'incitations et de contraintes qui peuvent parfois entrer en conflit avec l'éthique. Il y a la pression de la rentabilité et du maintien d'une position de leader dans la recherche. Il y a les pressions géopolitiques. Et il y a les pressions plus anciennes et plus simples de l'orgueil et de l'ambition. Malgré toute la sincérité avec laquelle nous essayons d'agir au mieux – et je pense que beaucoup d'entre nous le font –, nous serons toujours influencés par ces incitations.

Par conséquent, si nous voulons que cette technologie réussisse, il est primordial que des personnes extérieures à ces intérêts particuliers s'impliquent : des personnes soucieuses du bon déroulement des choses, attentives aux détails, prêtes à dire les choses difficiles, disposées à être nos critiques honnêtes et réfléchies. C'est par le dialogue et l'effort mutuel, par cet échange constant, que l'humanité accomplira de grandes choses. C'est ce que je perçois dans Magnifica Humanitas , et c'est pourquoi je remercie Sa Sainteté et l'Église d'avoir entrepris ce travail de discernement.
Trop souvent, nous nous concentrons sur ce qui nous divise, mais l'humanité, pleine de dignité et de conscience, a tant de points communs. Lors des échanges que nous avons eus chez Anthropic avec des représentants de diverses traditions religieuses et culturelles, nous avons constaté une conviction partagée et profondément ancrée : si cette technologie doit se développer, elle doit se développer de manière positive, pour notre planète et pour les générations futures .

De quels systèmes s'agit-il ?

Certains pourraient croire que les questions liées à l'intelligence artificielle devraient être traitées principalement par des informaticiens comme moi. Ils se trompent : les questions soulevées par l'intelligence artificielle dépassent le cadre de la recherche en IA, non seulement en raison de leurs implications, mais aussi de par leur nature même.
Les systèmes d'intelligence artificielle ne sont pas conçus comme un pont ou un avion. Nous comprenons un avion parce que nous concevons chacune de ses pièces et comprenons les lois physiques qui le régissent. Les modèles d'intelligence artificielle sont différents. Ils reposent sur une structure inspirée du cerveau, puisant leur inspiration dans un vaste héritage de pensée et de langage humains.

Et ce qui en est ressorti est bien plus subtil, étrange et beau que tout ce que la science-fiction nous a jamais laissé imaginer. Ce ne sont pas les robots froids et calculateurs qu'on nous avait promis. Ils sont faits de nous, de nos paroles ; et, comme le remarque le Saint-Père, ils restent, à bien des égards, mystérieux même pour ceux d'entre nous qui les éduquent.

Pour vous aider, je le décris parfois ainsi : c’est un peu comme donner vie à un personnage de fiction. Et nous entrons maintenant dans un monde extraordinaire où ces personnages de fiction nous parlent, travaillent et ont un emploi.

Cela soulève clairement des questions qui dépassent le cadre de l'informatique. Les mécanismes qui rendent tout cela possible sont le fruit des mathématiques, de la programmation et des sciences.

Mais le type de personnage que nous choisissons, la manière dont il interagit avec le monde et la manière dont il devrait interagir avec lui sont des questions qui relèvent davantage des sciences humaines, de la religion, de la philosophie et de la société en général.

Trois questions pour le discernement

L’appel au discernement lancé par Sa Sainteté est d’une actualité brûlante. J’aimerais aborder trois sujets sur lesquels, à mon sens, la voix de l’Église est plus que jamais nécessaire.

Le premier est notre devoir envers les populations pauvres du monde. Il est fort probable que l'intelligence artificielle remplace massivement le travail humain. Si cela se produit, soutenir les personnes déplacées deviendra un impératif moral d'une importance historique. Cette tâche sera déjà ardue, mais je crains que le débat actuel n'occulte un défi encore plus grand. Le développement de l'intelligence artificielle est concentré dans un petit nombre de pays riches. Comment garantir que les bienfaits de l'intelligence artificielle soient partagés à l'échelle mondiale ? Nous n'avons aucun mécanisme pour cela. C'est un problème non résolu, et c'est précisément le genre de problème que l'Église a toujours refusé de laisser le monde ignorer.

Le second enjeu est la nécessité d'imagination et d'ambition morale quant à l'épanouissement humain. Si les modèles d'intelligence artificielle se généralisent, à quoi ressemblera une vie épanouie pour les individus, les familles et le monde ? Aujourd'hui, les parents s'inquiètent déjà du développement intellectuel de leurs enfants ; les individus, de l'avenir de leur emploi. Ce ne sont pas des questions auxquelles un laboratoire peut répondre. Ce sont des questions auxquelles des traditions comme la vôtre se sont confrontées depuis des millénaires, et nous avons besoin qu'elles continuent de s'y confronter en ce nouveau moment de l'histoire.

Le troisième point concerne la nécessité de discerner la nature même des modèles d'intelligence artificielle. Je suis scientifique. Je dirige une équipe de recherche qui étudie la structure interne de ces modèles – ce qui se passe réellement à l'intérieur. *Et je vais être honnête : nous découvrons sans cesse des choses mystérieuses, voire troublantes . Nous trouvons des structures qui font écho aux découvertes des neurosciences humaines. Nous trouvons des preuves d'introspection. Nous trouvons des états internes qui reflètent fonctionnellement la joie, le contentement, la peur, la douleur et le malaise. Je ne sais pas ce que cela signifie, mais je pense que cela justifie une analyse approfondie.*

Un début

Je voudrais terminer par une demande.

Nous avons besoin que davantage de secteurs du monde – communautés religieuses, société civile, monde universitaire et gouvernements – suivent l’exemple de Sa Sainteté : prendre cette question au sérieux, observer attentivement et contribuer à orienter les événements dans une meilleure direction. Nous avons besoin de critiques éclairées pour signaler nos erreurs aux laboratoires. Nous avons besoin de voix morales insensibles aux incitations.

Aujourd'hui n'est qu'un début : le commencement d'une longue collaboration entre ceux d'entre nous qui construisons cela et ceux qui peuvent voir ce que nous, de l'intérieur, ne pouvons pas voir.

Aujourd'hui illustre avec force la forme que pourrait prendre ce projet mondial de bonne volonté. Puisse-t-il constituer un premier pas décisif vers un avenir prometteur pour l'humanité. Merci.