Un Chemin Étroit

Comment assurer notre avenir ?

Andrea Miotti, Tolga Bilge, Dave Kasten, James Newport
Traduit en français par Flavien Chervet avec l'aide de GPT-4o,
et assistance éditoriale par Adam Shimi

Octobre 2024

Merci à Anthony Aguirre, Connor Leahy, Max Tegmark, Eva Behrens, Leticia Garcia Martinez, Gabriel Alfour, Adam Shimi, Pedro Serodio, et aux nombreux autres contributeurs pour leurs apports et pour les discussions vitales que nous avons eues. Merci à tous ceux qui ont proposé des retours sur nos nombreux brouillons.
Merci à Eleanor Gunapala pour les graphiques et la publication.

Nous contacter : contact@narrowpath.co
www.narrowpath.co

Résumé Exécutif

Une vérité simple s'impose à nous : l'extinction de l'humanité est possible. L'histoire récente nous a également révélé une autre vérité : nous sommes capables de créer une intelligence artificielle (IA) qui peut rivaliser avec l'humanité.

Bien que la majorité des développements en IA soient bénéfiques, la superintelligence artificielle menace la survie de l'humanité. À l'heure actuelle, nous ne disposons d'aucune méthode pour contrôler une entité dotée d'une intelligence supérieure à la nôtre. Encore pire, nous ne pouvons même pas prédire le niveau d'intelligence des IA avancées avant leur développement, et nos moyens pour évaluer leur compétence une fois développées sont totalement insuffisants.

Voilà qui pose le contexte de ce moment critique où nous nous trouvons. Partout dans le monde, des entreprises investissent pour créer une superintelligence artificielle – qu'elles estiment capable de surpasser les capacités collectives de tous les humains. Elles affirment publiquement qu'il ne s'agit pas de savoir "si" une telle superintelligence existera, mais "quand".

Comme nous l'avons pointé, notre civilisation ne sait pas comment contrôler une IA qui nous surpasse immensément en puissance. Si ces efforts pour engendrer une superintelligence réussissent, c'est notre extinction qui est en jeu. Mais l'humanité peut choisir un autre futur : il existe un étroit chemin à travers ce terrain miné.

Un nouvel avenir ambitieux se dessine au-delà de ce chemin étroit. Un avenir porté par le progrès humain et technologique. Un avenir où l'humanité réalise les rêves et aspirations de nos ancêtres d'éradiquer les maladies et la pauvreté extrême, d'obtenir une énergie pratiquement illimitée, de vivre plus longtemps et en meilleure santé, et d'explorer le cosmos. Cet avenir exige que nous gardions le contrôle de nos créations, y compris les IAs.

Actuellement, nous avançons sans aucun garde-fou vers la création d'une IA menaçant l'extinction de l'humanité. Ce document propose notre plan pour sortir de cette voie dangereuse et emprunter un autre chemin. **Pour ce faire, nous avons développé des propositions d'actions pour les décideurs politiques, réparties en trois phases :**

Phase 0 : Sûreté - D'abord, les nations devraient immédiatement créer de nouvelles institutions, législations et stratégies politiques afin de prévenir le développement d'une superintelligence artificielle que nous ne pourons pas contrôler. Si ces mesures sont correctement exécutées, **elles devraient empêcher quiconque de développer une superintelligence artificielle pendant les 20 prochaines années.**

Phase 1 : Stabilité - Ensuite, les nations devraient créer des institutions internationales à même de garantir la pérennité des mesures de contrôle du développement de l'IA ne s'effondrent pas en raison de rivalités géopolitiques ou de développements anarchiques par des acteurs étatiques et non étatiques. Si ces

mesures sont correctement mises en œuvre, elles devraient assurer la stabilité et conduire à un système international de supervision de l'IA qui perdure dans le temps.

Phase 2 : Épanouissement - Enfin, une fois contenu le développement de superintelligences scélérates et établi un système international stable, l'humanité pourra se concentrer sur les fondations scientifiques d'une IA transformative sous contrôle humain. C'est-à-dire entre autres construire une science et une métrologie robustes de l'intelligence, une ingénierie de l'IA sécurisée par conception, et le reste des bases pour la création d'IAs à même de transformer la société tout en restant sous contrôle humain.