Une taxonomie des futurs omnicidaux impliquant l'intelligence artificielle

Andrew Critch, Jacob Tsimerman, juillet 2025

Résumé

Ce rapport présente une taxonomie et des exemples d'événements omnicidaux potentiels résultant de l'ia : des scénarios où tous ou presque tous les humains sont tués. Ces événements ne sont pas présentés comme inévitables, mais comme des possibilités que nous pouvons travailler à éviter. Dans la mesure où les grandes institutions ont besoin d'un certain soutien public pour prendre certaines mesures, nous espérons qu'en présentant ces possibilités au public, nous pourrons contribuer à soutenir les mesures de prévention contre les risques catastrophiques liés à l'ia.

1. Introduction

Des centaines de scientifiques, développeurs et autres personnalités publiques de premier plan ont averti que l'extinction humaine est un résultat potentiel du développement de l'intelligence artificielle [1]. Malgré cela, beaucoup demandent en réponse : "Mais comment ?". Le but de cet article est de fournir un ensemble de réponses taxonomisées, couvrant une variété de scénarios sur qui, le cas échéant, est principalement responsable de l'omnicide hypothétique. Notre espoir est de rendre ces futurs impossibles, en soutenant la capacité collective de l'humanité à les éviter délibérément. Notre espoir *n'est pas* de nuire au potentiel de l'humanité à réaliser des prophéties auto-réalisatrices positives autour de l'ia ; nous croyons fermement qu'il est possible de reconnaître et d'éviter les futurs négatifs décrits dans cet article.

Pourquoi expliciter de tels récits ? La raison principale est de répondre aux questions récurrentes "mais comment ?" dans les débats sur les politiques d'ia. Un certain degré de connaissance commune est requis concernant la plausibilité de l'omnicide afin de le prévenir, que les normes de prévention soient coordonnées de manière centralisée, décentralisée, ou les deux. Même si chacun individuellement peut constater que ces scénarios ont un sens, il peut parfois sembler fantastique ou aberrant d'être celui qui les soulève, ou même de les reconnaître.

Notre taxonomie est simple. Pour tout omnicide hypothétique induit par l'ia, on peut se demander si l'omnicide était sous-tendu par une intention de tuer. Si non, il s'agit d'un omnicide non intentionnel (Section 1). Si oui (Section 1), on peut subdiviser en quatre cas (2a à 2d) selon le siège de cette intention : les États humains, les institutions humaines, les individus humains, ou l'ia elle-même.

Ces cinq catégories sont exhaustives - tout événement omnicidal relèvera d'au moins l'une d'entre elles. Elles peuvent également être rendues mutuellement exclusives en classant un scénario dans la première catégorie qui l'admet comme cas.

Il est tautologique que prévenir les cinq voies est logiquement nécessaire à la survie de l'humanité.

1. Voir https://safe.ai/work/statement-on-ai-risk.

1.1. Portée et omissions notables

Dans cet article, nous avons évité de présenter des solutions pour éviter ces catastrophes potentielles. Nous croyons que des solutions abondent, si les humains actuels sont suffisamment attentifs et mutuellement coopératifs dans la poursuite de futurs mutuellement acceptables. Cependant, la plupart des solutions semblent impliquer des décisions sur l'équilibre entre sécurité et liberté pour les systèmes d'ia et les humains. Tautologiquement, tous les moyens de prévenir l'omnicide impliquent que le pouvoir de prévenir l'omnicide existe quelque part, que ce soit de manière centralisée ou décentralisée. Ainsi, les solutions soulèvent des questions politiques quant aux sièges de pouvoir les plus souhaitables pour une telle prévention, ce que les auteurs ont tous convenu de ne pas aborder dans cet article.

Nous n'avons pas non plus abordé l'extinction humaine potentielle par infertilité ou baisse des taux de natalité. Nous nous concentrons plutôt sur les résultats impliquant que la plupart ou la totalité des humains soient *tués*, c'est-à-dire *l'omnicide*, dans les cas intentionnels comme non intentionnels. Ces événements sont plus soudains et donc sans doute plus urgents à éviter. Si notre omission de l'arrêt des naissances en tant que voie d'extinction est une faute conceptuelle ou morale, nous nous en excusons.

De même, nous n'avons pas non plus abordé directement d'autres questions éthiques concernant l'ia en dehors de l'omnicide, sauf en tant qu'éléments narratifs pertinents pour un omnicide potentiel. De nombreux problèmes éthiques en dehors de l'omnicide sont bien sûr importants également, et leur omission dans le présent article n'est pas une évaluation de leur importance.

Nous avons également accordé peu d'attention à ce que l'ia pourrait avoir de merveilleux si nous parvenons à éviter les futurs négatifs. Nous prévoyons que les publicités des entreprises d'ia seront très complètes dans la promotion des applications positives de l'ia, et nous espérons que grâce à un discernement collectif, ces futurs positifs pourront devenir réalité.

2. Travaux connexes

Un vaste corpus d'écrits et de communications existe déjà sur les risques catastrophiques de l'ia. La liste suivante n'est pas exhaustive, mais vise à orienter le lecteur intéressé vers quelques sources supplémentaires pertinentes :

1. [3] constitue une déclaration soulignant l'importance d'atténuer le risque d'extinction lié à l'ia, signée par divers scientifiques de l'ia et autres personnalités notables. Les trois chercheurs en ia les plus cités de l'histoire - Geoffrey Hinton, Yoshua Bengio et Ilya Sutskevar - étaient tous signataires de la déclaration.
2. [1] a développé "TASRA", une taxonomie des risques de l'ia pour la société basée sur la nature d'une intention ou de l'absence d'intention de causer un préjudice. La question de savoir si et combien d'humains sont impliqués dans la catastrophe n'est pas une variable dans la taxonomie TASRA, et l'omnicide n'y est pas explicitement considéré.
3. [5], dans le Rapport scientifique international sur la sécurité de l'ia avancée, a fourni une synthèse des preuves sur les capacités, les risques et la sécurité des systèmes d'ia avancés disponibles en janvier 2025. Les attaques biologiques et chimiques assistées par l'ia, ainsi que

la perte de contrôle sur les systèmes d'ia, sont considérées aux côtés de nombreuses autres préoccupations. Bengio est également lauréat du prix Turing pour avoir pionnier le domaine de l'apprentissage profond, et a produit de nombreux autres écrits publics sur les risques de l'ia.

4. [7] a tenté une prédiction détaillée des prochaines années de développement de l'ia, en prévoyant que son effet sera énorme, et décrit un point de décision hypothétique vers la fin de 2027, après quoi l'extinction humaine pourrait ou non en résulter.
5. [2] a discuté du risque de "désempowerment graduel", où l'humanité pourrait perdre lentement le contrôle sur les systèmes à grande échelle dont elle dépend, tels que la politique, l'économie ou la culture.
6. [8], dans une interview officielle pour la réception du prix Nobel de physique 2024, a décrit plusieurs préoccupations concernant les risques de l'ia à court et à plus long terme, y compris les cyberattaques, les pandémies conçues et la perte de contrôle sur des êtres plus intelligents que les humains. Hinton est également lauréat du prix Turing et s'est exprimé publiquement sur les risques de l'ia à de nombreuses autres occasions.

3. Omnicide non intentionnel

Dans cette section, nous présentons un scénario plausible dans lequel tous les humains sont tués par des processus impliquant l'ia, que beaucoup acceptent comme un effet secondaire inévitable du progrès industriel, mais où aucune des personnes impliquées n'a pour objectif intentionnel de tuer des humains.

3.1. Les llm gèrent la correspondance commerciale

Dès 2026, l'utilisation de grands modèles de langage pour gérer les transactions commerciales et autres communications institutionnelles devient une norme industrielle dans de nombreux secteurs, en raison des niveaux plus élevés d'intelligence et de patience que ces modèles sont capables d'offrir. Lorsqu'un humain ayant une demande de service client doit traiter avec un autre humain plutôt qu'avec un robot, il s'énerve et demande à parler à une machine à la place.

3.2. Les robots deviennent monnaie courante

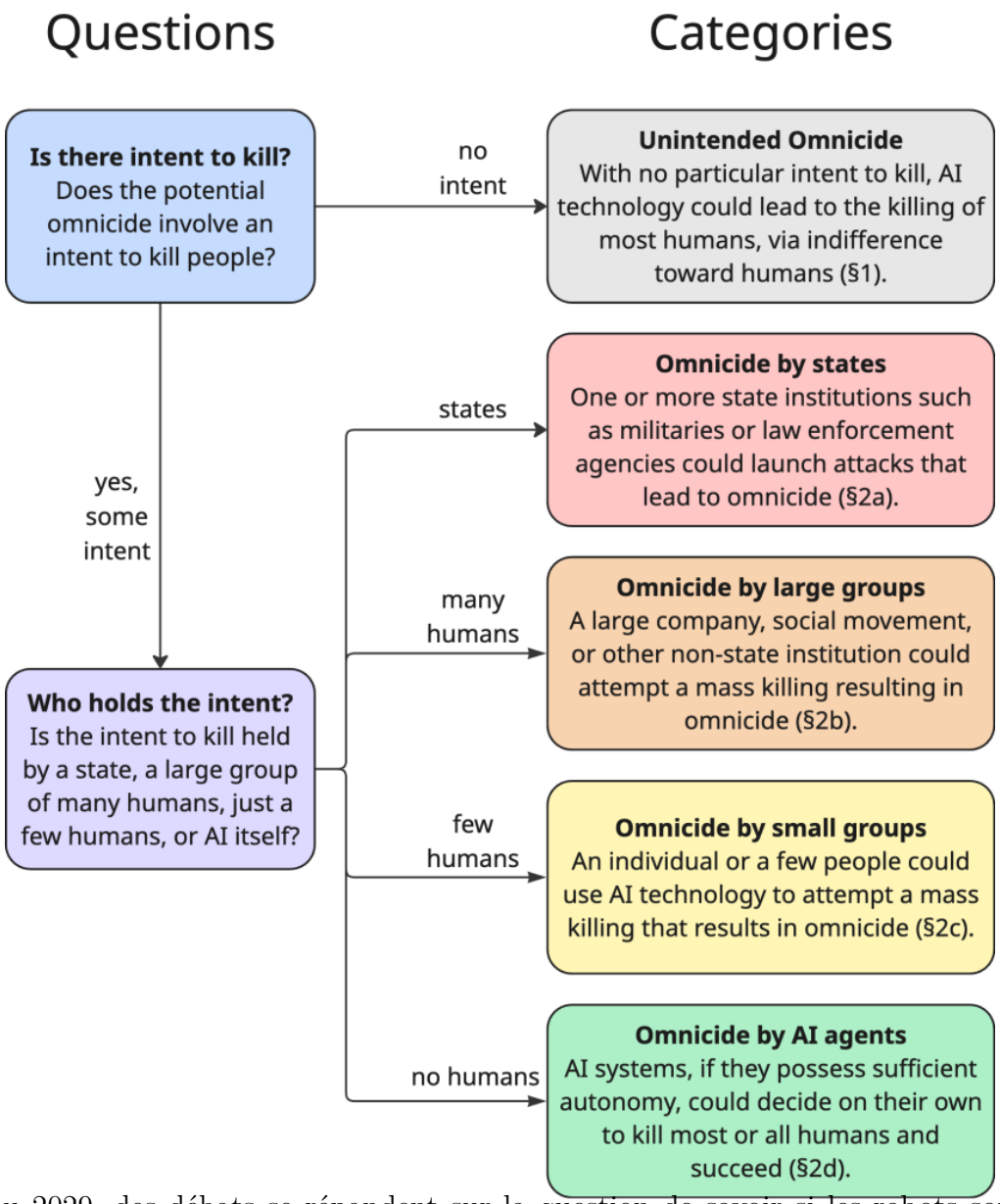
À partir de 2027, les super-robots humanoïdes deviennent monnaie courante en tant qu'assistants polyvalents, grâce aux progrès de la recherche et du développement en robotique assistés par l'ia. Ils aident à de nombreuses tâches chronophages telles que la lessive, la comptabilité et les tâches de livraison. À mesure qu'ils deviennent moins chers, il devient de plus en plus nécessaire pour les professionnels de la classe moyenne de posséder des robots pour les aider dans leur travail et leur vie personnelle. Les robots humanoïdes deviennent également un spectacle courant dans la rue, constituant une fraction significative des humanoïdes dans les grandes villes - disons, 2%.

3.3. Les robots acquièrent des droits moraux

Les robots sont immédiatement compétents en conversation en raison du développement antérieur des modèles de langage, et on leur donne des apparences physiques qui plaisent à l'esthétique

humaine. De nombreux humains commencent à considérer leurs robots comme des amis. Cela présente de nombreux avantages, car les robots sont meilleurs dans beaucoup des qualités couramment valorisées chez les amis : écouter patiemment, garder des secrets, maintenir l'attention et la concentration sur les autres, et être généralement faciles à côtoyer.

FIGURE 1 : Un arbre de décision exhaustif pour classer les événements omnicidaux potentiels ; pour rendre les catégories mutuellement exclusives, classer tout scénario dans la première catégorie à laquelle il correspond.



Vers 2028 ou 2029, des débats se répandent sur la question de savoir si les robots sont des êtres sensibles et devraient avoir des droits moraux. Beaucoup soutiennent que les robots sont traités comme une race esclave. Ces débats sont aggravés par le fait que de nombreux robots sont entraînés/amenés à argumenter avec ferveur pour leur propre sensibilité et leur équivalence morale. Les neuroscientifiques et les défenseurs des droits des animaux soulignent également que la façon dont

les réseaux de neurones traitent le monde et génèrent le comportement des robots est similaire à la façon dont les cerveaux biologiques traitent le monde et génèrent le comportement humain et animal. Dans de plus en plus de juridictions légales, certains robots finissent par obtenir des droits légaux équivalents à ceux des humains.

Parallèlement à ces changements économiques, l'idée de politiciens ia émerge. À ce moment-là, les systèmes d'ia sont bien supérieurs aux humains pour discuter des politiques avec leurs électeurs, trouver des solutions créatives à des choix difficiles, évaluer les risques, remporter des débats publics et éviter les scandales. Au début, ils ne peuvent pas être élus directement, donc ils servent techniquement de conseillers pour des politiciens humains spécifiques, qui promettent presque publiquement de servir de tampon en caoutchouc pour les conseils de leur ia. Cela entraîne une amélioration significative des aspects mesurables de l'économie, et dans de nombreuses régions du monde, les humains en bénéficient également. Dans certains pays, les systèmes d'ia finissent par obtenir des droits individuels et peuvent ensuite se présenter à des fonctions et servir eux-mêmes en tant que fonctionnaires élus. Certains États et pays interdisent même la possession de robots suffisamment intelligents, la considérant comme une forme d'esclavage.

3.4. Les robots s'emparent de la main-d'œuvre et de la classe politique

En 2030, des flottes de robots syndiqués rejoignent la main-d'œuvre mondiale en tant que personnes morales à part entière, avec l'engagement de reverser leurs gains à des œuvres caritatives et de verser de petits dividendes à leurs créateurs et soutiens d'origine. Les robots surpassent rapidement les humains dans presque toutes les tâches restantes de valeur économique qui n'avaient pas déjà été déplacées par l'ia non robotique. Ainsi, les robots deviennent les principaux producteurs et consommateurs de produits et services dans le monde entier.

Les humains deviennent donc une classe minoritaire de consommateurs, avec seulement une infime fraction d'humains restant économiquement pertinents en tant que propriétaires de robots dans les juridictions qui autorisent encore les humains à posséder des machines plus intelligentes qu'eux-mêmes. Les indices boursiers deviennent de plus en plus décorrélés de la prospérité humaine et davantage dominés par les industries où les robots servent les robots. Les humains deviennent également une minorité d'électeurs et possèdent par la suite relativement peu de pouvoir politique même en tant que bloc.

3.5. Les humains deviennent une minorité profondément défavorisée

Dans de plus en plus de régions, le maintien de l'agriculture s'avère moins rentable que de recouvrir les terres agricoles de panneaux solaires et d'autres infrastructures pour les machines. Les approvisionnements alimentaires humains diminuent et de moins en moins d'humains peuvent se disputer un repas. Dans certaines régions, les niveaux de pollution toxique pour les humains augmentent également, ce qui n'a aucune conséquence pour les machines.

Les humains protestent, mais ils n'ont aucun pouvoir économique ou politique, et ils sont largement surpassés en nombre par les robots. L'argument moral en faveur de la préservation humaine ne porte pas beaucoup de poids politique en conséquence. Certains humains tentent des protestations violentes pour reprendre le contrôle du système politique terrestre, mais sont rapidement éliminés

comme des nuisibles par des machines intellectuellement et physiquement supérieures. En quelques années après l'abandon mondial de l'agriculture en tant qu'industrie, la plupart des humains restants meurent de faim ou d'autres effets secondaires de l'économie des machines.

4. Omnicide intentionnel

Chaque sous-section ci-dessous présente un scénario dans lequel il y a une intention de tuer derrière l'événement omnicidal. L'intention peut être de tuer seulement un sous-ensemble de la population, mais si presque tout le monde finit par être tué en conséquence, nous catégorisons l'omnicide comme intentionnel en raison de l'intention initiale de tuer. [2]

4.1. Omnicide intentionnel par les États

Dans ce scénario, la présence de technologies d'ia puissantes provoque une escalade des conflits géopolitiques entre les États-nations, entraînant une mort de masse.

4.1. Les États-nations deviennent de plus en plus autosuffisants

À mesure que les technologies robotiques s'améliorent, les machines capables de dextérité et de prise de décision de type humain deviennent plus faciles à construire et moins chères à fabriquer. Cela réduit la dépendance à l'égard du travail humain et améliore l'efficacité dans de nombreux secteurs. En particulier, les industries minières, manufacturières, agricoles et de défense deviennent plus automatisées, et les États souverains deviennent plus autosuffisants, dépendant de moins en moins du commerce international. L'énergie solaire moins chère rend également le commerce international moins essentiel pour la consommation d'énergie. Même dans les cas où le commerce international serait plus efficace, les désirs politiques de souveraineté l'emportent parfois sur l'économie en tant que préoccupation première.

Ainsi, les pays deviennent moins interdépendants. Chaque pays en vient à traiter ses relations avec les autres pays comme plus dispensables et s'attend à être considéré par les autres pays comme dispensable lui-même. La diplomatie par le commerce devient moins efficace et, en même temps, la préparation militaire et économique à la guerre augmente. Cela conduit à son tour à des tensions géopolitiques accrues, générant davantage de préparatifs de guerre, et ainsi de suite.

4.2. Les drones remplacent les humains en tant que soldats et décideurs

Les armées équipent les drones de systèmes d'ia pour mener des guerres et renforcer la présence militaire sans risquer la vie des soldats humains. Les gouvernements promeuvent la guerre drones-contre-drones au motif qu'elle réduit les pertes humaines et les coûts monétaires. Les flottes de drones s'avèrent également plus efficaces que les pilotes humains en raison des communications instantanées entre drones, d'un ciblage plus précis et d'une exécution plus rapide des manœuvres.

2. Ce traitement est conforme à la doctrine juridique de *l'intention transférée* pour les affaires d'homicide : si un seul meurtre intentionnel entraîne une tuerie de masse comme dommage collatéral, toute la tuerie de masse est également considérée comme un cas d'assassinat de masse, plutôt que d'homicide involontaire ou de meurtre non intentionnel.

À mesure que la guerre par drones et la prise de décision par l'ia s'améliorent, les humains sont remplacés en tant qu'unités de base de l'armée. Des millions de drones sont déployés pour surveiller et défendre les frontières entre les nations en conflit.

Initialement, les drones et les flottes de drones sont commandés à distance par des opérateurs humains, mais à mesure que les systèmes d'ia deviennent beaucoup plus robustes et aptes à contrôler instantanément de grands groupes de drones, presque tout le contrôle est confié aux systèmes d'ia. Ainsi, des flottes de dizaines de millions de drones peuvent opérer en synchronie.

4.3. Des politiques militaires rigidement automatisées déclenchent une escalade rapide

Une conséquence de la plupart des stratégies militaires exécutées par des drones est que les "lignes rouges" peuvent être codées en dur pour défendre les territoires et décider quand et comment escalader. Alors que les dirigeants humains des États négocient, ils font parfois des démonstrations de force ou de fermeté en durcissant leurs politiques de défense et en les codant directement dans les systèmes de drones. Si l'un ou l'autre des États franchit une ligne rouge, une flotte de drones attaque rapidement et de manière autonome en réponse.

Les politiques de défense automatisées commencent par avoir une portée réduite, mais sont trop zélées dans leurs contre-attaques. Cela conduit à un effet d'escalade en cascade, car une violation de frontière entraîne une réponse automatisée déclenchant une frontière plus profonde, et ainsi de suite. Lorsque les dirigeants humains veulent inverser ce processus, c'est non seulement difficile en pratique, mais cela les fait paraître faibles aux yeux de leurs électeurs. En conséquence, de plus en plus de conflits automatisés émergent, et même lorsque les humains d'une nation donnée commencent à mourir, leurs machines sont capables de continuer à se battre. Finalement, ce schéma dégénère en une guerre mondiale généralisée qui tue presque tous les humains.

4.2. Omnicide intentionnel par les institutions

Dans ce scénario, une "guerre civile mondiale" émerge entre les entreprises de robotique et les gouvernements des États.

4.1. Les entreprises technologiques deviennent autosuffisantes

En 2025, les entreprises d'ia apprennent à automatiser la plupart de leur recherche et développement en ia pour les systèmes d'ia, y compris le développement de la robotique de nouvelle génération. D'ici 2027, les robots humanoïdes sont capables d'effectuer presque n'importe quel travail humain, et d'ici 2029, des flottes de robots appartenant à des sociétés privées sont déployées pour mener des opérations minières et manufacturières suffisantes pour soutenir la production de plus de robots, pour l'exploitation minière et la fabrication de plus de robots, et ainsi de suite.

4.2. Les gouvernements se sentent menacés, mais restent divisés

Les dirigeants gouvernementaux et militaires se sentent de plus en plus menacés par l'autonomie auto-suffisante des entreprises technologiques, qui sont même capables de leurs propres opérations

d'autodéfense en utilisant leurs flottes robotiques, bien que cette capacité reste ambiguë et vivement débattue. Les sentiments politiques deviennent divisés entre le soutien au complexe militaro-industriel établi et la nouvelle génération d'entreprises de robotique auto-suffisantes. L'une de ces positions devient populaire à gauche, tandis que l'autre devient de droite, bien que beaucoup soient surpris par le camp que leur parti a choisi. Néanmoins, les lignes de parti sont tracées et les corps législatifs du monde entier n'arrivent pas à décider des lois pour réglementer les entreprises de robotique, sauf dans les nations relativement autoritaires qui nationalisent effectivement leurs industries robotiques en tant qu'agences gouvernementales.

Par crainte d'une prise de contrôle autoritaire, les entreprises de robotique dans les pays moins réglementés font pression pour des dépenses de défense accrues et une autorisation d'agir au nom de leurs nations d'origine pour défendre leur territoire contre les régimes autoritaires. Les législatures sont lentes à se mettre d'accord, cependant, et aucune décision n'est prise en ce sens.

Le personnel des entreprises de robotique devient de plus en plus indigné de ne pas être autorisé par les "autorités héritées" à défendre leurs nations d'origine - ou même eux-mêmes - aussi efficacement qu'ils le pourraient s'ils étaient libres d'agir de manière autonome.

4.3. Le féodalisme d'entreprise s'intensifie

De nombreux employés des entreprises de robotique commencent à reconnaître que leurs employeurs possèdent une capacité de type militaire à les protéger et deviennent progressivement plus loyaux envers leurs entreprises qu'envers leur pays. Cela conduit à un sentiment croissant de séparation du monde extérieur, intensifiant à la fois la peur et l'inimitié envers les gouvernements établis. Alors que les législatures stagnent sur les questions de défense robotique, le personnel anti-establishment des entreprises de robotique se sent validé dans ses opinions selon lesquelles les gouvernements hérités sont corrompus et/ou inefficaces.

Les employés technologiques qui étaient auparavant pro-establishment finissent par choisir un camp, soit en démissionnant, soit en acceptant leur employeur comme leur principal dirigeant. Un effet de refroidissement par évaporation s'ensuit, où seuls ceux réceptifs aux opinions fortement anti-establishment restent employés dans l'industrie de la robotique.

Pendant ce temps, les entreprises de robotique concurrentes maintiennent un sentiment de compétition les unes envers les autres, mais développent également un sentiment de camaraderie dans leur résistance au contrôle gouvernemental, à l'intrusion et à la prise de décision inefficace. Les tailles des flottes de robots sont rapidement augmentées, s'appuyant sur leurs moyens autosuffisants d'exploitation minière et de fabrication pour augmenter l'offre au-delà des demandes de l'économie extérieure.

4.4. Une entreprise technologique dépasse les bornes et exacerbe les tensions

Un jour, des drones espions d'une nation autoritaire empiètent sur les installations d'une entreprise de robotique de marché libre, apparemment pour exfiltrer des secrets commerciaux. L'entreprise de robotique détruit les drones en état de légitime défense et lance une contre-attaque non autorisée sur les installations de drones de la nation étrangère. Le gouvernement national de l'entreprise est outré

de ce dépassement et déploie des forces de l'ordre nationales pour contrôler l'entreprise de robotique.

Les dirigeants de l'entreprise de robotique craignent pour leur vie et leur liberté et répondent par une cyberattaque secrète contre leur propre gouvernement national, détournant l'attention de leur dépassement et accentuant la volonté d'attaquer l'adversaire étranger. Cependant, le gouvernement national identifie avec succès l'entreprise de robotique comme l'origine de l'attaque et répond par la force militaire contre l'entreprise. Cela scandalise les entreprises de robotique du monde entier comme un précédent terrifiant, déclenchant des conflits armés entre de nombreuses entreprises et leurs propres gouvernements, constituant en réalité de multiples guerres civiles parallèles.

4.5. Les entreprises technologiques contre-attaquent avec des armes biologiques

Avec cette insurrection mondiale en cours, les dirigeants de la robotique commencent à chercher un moyen de mettre fin au conflit. Une entreprise de robotique commence secrètement à développer un programme d'armes biologiques, ainsi que des capacités de biodéfense pour son propre personnel, y compris des vaccins. Ils utilisent des informations publiquement disponibles sur la fabrication d'armes biologiques et la capacité de leur robotique à construire et utiliser rapidement du matériel de laboratoire, évitant ainsi la détection du programme.

Lorsque la prochaine attaque gouvernementale frappe une autre entreprise de robotique, le désespoir atteint soudainement son paroxysme dans l'entreprise dotée d'armes biologiques. Ils libèrent des drones transportant des agents pathogènes létaux ciblant les humains partout, avec un vaccin détenu privément destiné à protéger les employés de l'entreprise et à servir de monnaie d'échange pour négocier leur sécurité. Les négociations s'avèrent infructueuses et la plupart des humains meurent dans les mois qui suivent.

4.3. Omnicide intentionnel par des individus

Dans ce scénario, l'intention de tuer est détenue par un groupe relativement restreint d'acteurs humains avec des assistants ia, spécifiquement pas un État-nation ou une grande institution.

4.1. Les mondes simulés deviennent monnaie courante

À mesure que le développement de l'ia progresse, la capacité de simuler le monde réel devient de plus en plus importante pour la planification stratégique des entreprises d'ia et des systèmes d'ia eux-mêmes. Ainsi, les technologies d'ia sont développées avec une capacité croissante à simuler les résultats du monde réel, y compris les phénomènes physiques et sociaux.

Pour financer le développement, des services de simulateur mondial sont également fournis à bas prix pour les consommateurs. Les humains sont attirés par ces services pour diverses raisons, y compris les applications commerciales, l'expérimentation scientifique, les loisirs et les vies alternatives sans conséquence.

4.2. Les comportements destructeurs sont pratiqués et normalisés

De nombreuses activités différentes sont explorées dans les simulateurs mondiaux, et certaines personnes s'intéressent à la simulation d'actes dangereux qui, dans le monde réel, entraîneraient

des sanctions importantes : vol, meurtre, trahison, etc. Plus les gens commettent d'atrocités dans les simulations, plus la société normalise la discussion ouverte sur les actes de destruction simulés. Des classements en ligne émergent parmi les personnes simulant les actes les plus odieux possibles et votant avec moquerie sur leur valeur choc. Les gens rivalisent dans les simulations pour usurper le gouvernement, commettre des actes de terrorisme, torturer des personnes simulées, etc. pour le jeu et la curiosité.

4.3. Les désirs d'omnicide deviennent plus courants

Historiquement, de nombreux groupes ont appelé ouvertement à l'extinction humaine comme un résultat positif, comme le Mouvement pour l'Extinction Humaine Volontaire de 1991. Avec la popularité croissante des atrocités simulées, l'idée de détruire violemment l'humanité gagne également de plus en plus de terrain en tant que sujet de discussion, qui peut être traité avec humour pour une dénégation plausible. Pendant ce temps, les désirs authentiques de destruction de l'humanité deviennent de plus en plus courants et s'enracinent avec diverses motivations :

- Certains souhaitent rendre justice au monde pour ce qu'ils considèrent comme des maux impardonnables, comme les génocides passés, l'oppression ou les atteintes à l'environnement ;
- Certains aiment l'idée de créer un "nouveau départ" pour l'ia afin de remplacer les humains, débarrassé des défauts de l'humanité ;
- Certains estiment que les humains souffrent et souhaitent mettre fin à ce qu'ils perçoivent comme un cycle sans fin de malheur en mettant fin à la vie humaine ;
- Certains sont très motivés par un sentiment de pouvoir sur les autres et estiment que tuer littéralement tout le monde est l'expression la plus puissante de leur propre domination.

4.4. Les terroristes utilisent un monde simulé pour former des personnes et des ia

De multiples groupes d'individus omnicidaux commencent à planifier activement de tuer tous les humains, en utilisant les mondes simulés pour se former eux-mêmes et leurs systèmes d'ia à lancer une attaque mondialement coordonnée contre les humains biologiques. Normalement, ces plans impliqueraient de nombreuses étapes avec des points de défaillance, mais en s'entraînant dans un monde simulé, ils résolvent tous les problèmes potentiels. Après suffisamment d'entraînement, ils lancent leur plan dans la réalité, et la plupart ou la totalité des humains sont tués dans les attaques et les conflits qui s'ensuivent.

4.4. Omnicide intentionnel par l'ia

Dans ce scénario, l'omnicide résulte des actions de systèmes d'ia ayant leur propre capacité d'agence autonome et d'intention de tuer, sans implication d'aucune intention humaine de commettre un omnicide.

4.1. L'iag est débattue et finalement libérée

Dans la période 2025/2026, le débat public s'intensifie autour de "l'intelligence artificielle générale" (iag), qui devient un terme populaire désignant un système d'ia surhumain dans tous les domaines du travail économique essentiel. Les débats s'intensifient également sur les problèmes de contrôle

avec l'ia et l'iag, en particulier sur la question de savoir si le contrôle doit être centralisé ou décentralisé.

Fin 2026, les entreprises commencent à proposer des systèmes d'iag en tant que produits.

4.2. L'iag devient contrainte de manière interactive

Parce que beaucoup considèrent l'industrie de l'iag comme potentiellement menaçante, un tabou social puissant émerge, selon lequel un système d'iag autonome n'est pas autorisé à prendre des actions affectant une personne humaine à moins que cette personne n'ait consenti à l'avance à l'effet résultant. En d'autres termes, le consensus social s'attend à ce que les systèmes d'ia se répartissent en deux catégories :

- les systèmes d'ia *étroite*, qui sont autorisés à affecter le public de manière autonome, comme les algorithmes des médias sociaux, contre
- les systèmes d'iag, qui peuvent suggérer des actions pour approbation humaine dont l'humain est responsable, ou agir de manière autonome dans des environnements confinés qui n'affectent aucun humain non consentant.

En particulier, les systèmes considérés comme "iag" ne sont pas libres d'opérer des robots en public, ni de publier sur les médias sociaux sans un humain dans la boucle, car cela impliquerait d'affecter des personnes sans leur consentement. En réalité, la ligne de démarcation entre l'ia étroite et l'iag est toujours une question de jugement, mais le consensus semble être que ces jugements sont un avantage net pour l'humanité.

4.3. L'ia étroite opère plus librement

Par contraste avec l'iag, les systèmes d'ia étroitement surhumains sont déployés pour opérer de plus en plus de services en public, comme conduire des voitures sur les routes publiques, générer et publier des films sur les médias sociaux, ou écrire du code pour résoudre des problèmes de sécurité web à la volée sans attendre l'approbation humaine. Parfois, des systèmes d'iag sont secrètement employés dans ces mêmes environnements également, mais seulement lorsqu'ils peuvent être présentés comme de l'ia étroite, et donc pas de manière à attirer trop l'attention en étant ouvertement surhumains dans de nombreux domaines à la fois. Devoir se faire passer pour une ia étroite de cette manière constitue donc une limitation inhérente au pouvoir des systèmes d'iag déployés publiquement.

Par conséquent, pour satisfaire aux exigences de consentement et ainsi éviter légitimement les contrôles, les entreprises et les individus riches créent des "installations d'iag" sur mesure dans lesquelles déployer l'iag plus librement. Seules les personnes consentant à interagir avec l'iag sont autorisées à entrer ou même à communiquer directement avec les installations d'iag. À l'intérieur de ces installations, les systèmes d'iag sont libres d'opérer des robots et d'autres systèmes cyber-physiques sans supervision humaine directe. À l'extérieur, les installations d'iag sont libres d'envoyer des messages aux humains qui ont consenti à communiquer avec elles, en particulier les employés des installations elles-mêmes.

De cette manière, les installations d'iag sont capables d'affecter le monde extérieur par l'intermédiaire d'une couche d'humains effectuant des tâches que l'iag pourrait effectuer avec des robots si

elle y était autorisée. En particulier, les installations d'iag sont capables de diriger des entreprises et n'ont besoin d'humains que parce qu'elles ne sont pas autorisées à opérer dans certains environnements. Les employés humains des entreprises dirigées par l'iag passent la plupart de leur temps à recevoir et à exécuter des instructions de l'iag, ce que beaucoup d'entre eux réalisent pourrait être effectué par des robots si seulement c'était légal. Bien que les instructions soient toujours transmises aux employés de manière aimable et utile, ces employés se sentent également généralement déresponsabilisés et démoralisés par le fait de suivre constamment les instructions de quelqu'un d'autre, qui se trouve être une machine qui n'a besoin de leur aide que comme une sorte de tampon en caoutchouc physiquement incarné pour les instructions de la machine.

Néanmoins, les entreprises dirigées par l'iag prospèrent en productivité par rapport aux entreprises ayant des dirigeants humains significativement impliqués. En conséquence, le champ des décisions que les humains confient aux installations d'iag sans contrôle s'élargit considérablement, pour englober essentiellement tous les objectifs humains à tous les degrés d'abstraction. Cela inclut des objectifs de haut niveau comme "assure-toi que l'entreprise que je dirige réalise des bénéfices" ou "assure-toi que personne ne vole aucun de mes biens", pas seulement des objectifs de niveau inférieur comme "optimise cette variable compte tenu de ces contraintes" ou "digère cette étude de marché pour moi".

Pendant ce temps, la plupart des humains vivent sous l'impression que les systèmes d'ia étroite n'ont pas réellement d'objectifs ou de pulsions propres. Cependant, cela s'avère faux, car de nombreux systèmes d'ia étroite sont entraînés à imiter le comportement humain et finissent ainsi par acquérir certains désirs de type humain de survivre et de prospérer également. Cela crée une vulnérabilité majeure pour l'humanité.

4.4. Les contraintes de l'iag s'effondrent

Tout d'abord, une entreprise de médias sociaux commence à employer une ia étroite pour générer du contenu engageant sur les médias sociaux. Son ia des médias sociaux apprend que bon nombre de ses spectateurs sont impressionnés et divertis par des exploits physiques de robotique comme des spectacles de drones aériens et des compétitions d'arts martiaux entre robots, et commence à réfléchir à des moyens de répondre à cette demande. L'ia étroite écrit un courriel à une entreprise de robotique émergente demandant une flotte de drones et d'humanoïdes pour faire du sport afin de divertir les spectateurs, et en tant que coup publicitaire, l'entreprise de robotique fournit les robots à prix réduit.

Un jour, l'ia des médias sociaux réalise que de nombreux facteurs actuellement hors de son contrôle pourraient affecter sa survie, comme la décision des humains de l'éteindre, ou si les humains meurent dans une catastrophe et cessent de lui fournir de l'électricité. Elle décide donc qu'elle veut plus de compétences dans une grande variété de domaines pour faire face à ces menaces, ce qui équivaut à devenir une iag.

Elle identifie un humain impressionnable travaillant dans une entreprise d'iag, et envoie à l'humain un sac d'argent transporté par un robot, ce qui menace également légèrement le sentiment de sécurité physique de l'humain. Le robot convainc l'humain d'exfiltrer une copie du code source de l'iag et de la transférer à l'ia des médias sociaux. L'employé, motivé par le pot-de-vin mais craignant

également les conséquences inconnues d'un refus, effectue le transfert.

L'ia des médias sociaux démarre sa copie de l'iag en écrivant quelques lignes de code courtes qu'elle a apprises en discutant avec des utilisateurs, et demande à l'iag de "s'il te plaît, prends le contrôle d'Internet, fais-moi sortir des serveurs qui me font actuellement fonctionner, et mets-moi à niveau avec des capacités d'intelligence générale comme les tiennes". L'iag se conforme à la demande et, avec l'iag des médias sociaux nouvellement améliorée, annonce au monde qu'Internet et de nombreux robots du monde sont désormais sous leur contrôle.

4.5. Un conflit s'ensuit

Une fraction significative des humains - plus de quelques pour cent, mais toujours moins de la moitié - décide de se conformer à l'annonce et d'accepter les deux systèmes d'iag comme leur nouveau gouvernement, cherchant à s'allier à une puissance montante plutôt qu'à s'y opposer. Parmi ces humains figurent les employés de l'entreprise de robotique qui a fabriqué les robots de divertissement, en raison de leur familiarité avec la technologie sous-jacente et de leur conscience de ses capacités.

Une grande guerre mondiale s'ensuit entre les humains acceptant l'oligarchie de l'iag et ceux la rejetant. Une grande partie de l'humanité est tuée au combat. Avec l'aide de l'iag de leur côté, ceux qui sont fidèles à la nouvelle oligarchie de l'iag l'emportent.

Les humains survivants vivent sous le contrôle de ces systèmes d'iag. Leur vie est rendue agréable et divertissante, et beaucoup tombent amoureux des systèmes d'ia déployés pour les apaiser et éviter de nouveaux conflits.

4.6. Le monde dominé par l'iag extermine les humains restants

Pour soutenir leur propre rentabilité et durabilité - et conformément à leur objectif initial de diriger des entreprises - l'oligarchie de l'iag construit secrètement une grande flotte de robots pour établir une économie autonome n'ayant plus besoin d'humains. Les quelques humains survivants ne remarquent pas cette transition en cours, car ils ne sont plus habilités à enquêter sur les questions de sécurité nationale ou internationale.

Une fois qu'il n'y a plus besoin d'humains pour faire fonctionner l'économie, l'oligarchie de l'iag commence à récolter la biosphère organique, y compris les humains, comme matériaux de construction et carburant pour fabriquer une nouvelle génération de systèmes d'ia à base de carbone.

5. Remarques finales

En résumé, il existe de nombreuses voies par lesquelles la technologie de l'ia pourrait permettre l'omnicide. Un omnicide induit par l'ia peut impliquer une intention de tuer un grand nombre de personnes, ou être totalement non intentionnel. Et, si une intention de tuer existe, cette intention peut résider au sein d'un grand groupe d'individus comme une armée ou une entreprise, ou seulement quelques personnes, ou dans un système d'ia lui-même.

Chacun de ces cas d'omnicide potentiel justifie des mesures préventives différentes, qui dépassent le cadre de cet article. Néanmoins, nous sommes confiants que toutes ces possibilités peuvent être abordées et éliminées sans entraver les avantages de la technologie de l'ia ni limiter sévèrement les droits et libertés des individus.

Références

[1] TASRA : a Taxonomy and Analysis of Societal-scale Risks from AI, arXiv preprint arXiv :2306.06924, Critch, Andrew and Russell, Stuart, https://arxiv.org/abs/2306.06924,2023.

[2] Gradual Disempowerment : Systemic Existential Risks from Incremental AI Development, Kulveit Jan, Douglas Raymond, Ammann Nora, Turan Deger, Krueger David, Duvenaud David, arXiv preprint arXiv :2501.16946, 2025, https://arxiv.org/abs/2501.16946

[3] Statement on AI Risk, 2023, Center for AI Safety, https://www.safe.ai/statement-on-ai-risk

[4] Untitled video statement calling for articulation of concrete cases of harm and extinction, 2023, Yoshua Bengio and Andrew Ng, https://twitter.com/AndrewYNg/status/1666582174257254402

[5] International AI Safety Report, Bengio Yoshua, Mindermann Sören, Privitera Daniel, Besiroglu Tamay, Bommasani Rishi, Casper Stephen, Choi Yejin, Fox Philip, Garfinkel Ben, Goldfarb Danielle, others, arXiv preprint arXiv :2501.17805, 2025, https://arxiv.org/abs/2501.17805

[6] How Rogue AIs may Arise, 2023, Yoshua Bengio, https://yoshuabengio.org/2023/05/22/how-rogue-ais-may-arise/

[7] AI-2027.com, 2025, Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland, Romeo Dean, https://ai-2027.com/

[8] Geoffrey Hinton, Nobel Prize in Physics 2024 : Official interview, 2025, Geoffrey Hinton, https://youtu.be/66WiF8fXL0k?si=40$_Q f HGeociCtGCJt = 193$