Pourquoi les risques S sont les pires risques existentiels, et comment les prévenir

par Max_Daniel

2 juin 2017

26 minutes de lecture

Ceci est une publication croisée de https://www.youtube.com/watch?v=jiZxEJcFExc&list=PLwp9xeoX5p8Pi7rm-vJnaJ4AQdkYJOfYL&index=14

Les altruistes efficaces qui s'efforcent de façonner l'avenir lointain ont le choix entre différents types d'interventions. Parmi celles-ci, les efforts visant à réduire le risque d'extinction de l'humanité ont reçu le plus d'attention jusqu'à présent. Dans cet exposé, Max Daniel défend l'idée que nous pourrions vouloir compléter ce travail par des interventions visant à prévenir des avenirs très indésirables (« risques S »), et que cela justifie, parmi les sources de risque existentiel identifiées jusqu'à présent, de se concentrer sur le risque lié à l'intelligence artificielle.

Transcription : Pourquoi les risques S sont les pires risques existentiels, et comment les prévenir

Je vais parler des risques de souffrances graves à grande échelle dans un avenir lointain, les risques S. Pour illustrer ce que sont les risques S, j'aimerais commencer par une histoire fictive tirée de la série télévisée britannique Black Mirror, que certains d'entre vous ont peut-être vue. Dans ce scénario fictif, il est possible de télécharger des esprits humains dans des environnements virtuels. De cette manière, les êtres sensibles peuvent effectivement être stockés et fonctionner sur de très petits dispositifs informatiques, tels que le gadget en forme d'oeuf blanc que vous pouvez voir sur l'écran ici. Derrière le gadget, vous pouvez voir Matt.

Le travail de Matt consiste à vendre ces humains virtuels en tant qu'assistants virtuels. Et comme il ne s'agit pas d'une description de poste particulièrement attrayante pour tout le monde, une partie du travail de Matt consiste à convaincre ces humains de se conformer aux demandes de leurs propriétaires humains. Dans le cas présent, Greta, que vous voyez ici, n'est pas disposée à le faire. Elle n'est pas enchantée à l'idée de servir d'assistante virtuelle pour le reste de sa vie. Pour briser sa volonté, pour l'obliger à se conformer, Matt augmente la vitesse à laquelle le temps s'écoule. Ainsi, alors que Matt n'a besoin d'attendre que quelques secondes, Greta endure en réalité plusieurs mois d'isolement.

J'espère que vous êtes d'accord avec moi pour dire que ce scénario n'est pas souhaitable. Heureusement, ce scénario a peu de chances de se réaliser. Ainsi, quel que soit le scénario que nous pouvons imaginer, il est peu probable qu'il se réalise exactement sous cette forme. Ce n'est donc pas le sujet ici. Cependant, je soutiendrai qu'il existe en fait un large éventail de scénarios pour lesquels nous sommes confrontés à des risques de scénarios qui sont d'une certaine manière similaires à ce scénario ou même pires.

J'appellerai ces risques des risques S. J'expliquerai d'abord ce que sont ces risques S, en les opposant aux risques existentiels ou risques X, qui nous sont plus familiers. Puis, dans une deuxième partie, j'expliquerai pourquoi, en tant qu'altruistes efficaces, nous pourrions vouloir prévenir ces risques S et comment nous pourrions le faire. J'aimerais donc présenter les risques S comme une sous-classe des risques existentiels que nous connaissons mieux.

Comme vous vous en souvenez peut-être, ces risques ont été définis par Nick Bostrom comme des risques où une issue défavorable annihilerait complètement la vie intelligente originaire de la Terre ou, du moins, réduirait de façon permanente et drastique son potentiel. Dans l'une de ses principales publications sur les risques existentiels, Bostrom a également suggéré qu'une façon de comprendre en quoi ces risques diffèrent d'autres types de risques est d'examiner la gravité de cette issue défavorable selon deux dimensions. Ces dimensions sont la portée et la gravité du résultat négatif qui nous préoccupe. J'ai reproduit ici l'une des figures centrales de Bostrom.

Vous pouvez voir l'étendue du risque sur l'axe vertical. En d'autres termes, nous nous demandons ici combien de personnes seraient affectées négativement si les risques se concrétisaient. S'agit-il d'un petit nombre de personnes ? S'agit-il de tous les habitants d'une région donnée ou même de tous les habitants de la Terre ? Ou, dans le pire des cas, de tous les habitants de la Terre et de certaines, voire de toutes les générations futures ? La deuxième dimension pertinente est la gravité. Il s'agit de savoir, pour chaque individu concerné, quelle serait la gravité des conséquences. Prenons l'exemple d'un accident de voiture mortel, le risque d'un seul accident de voiture mortel. Si cela se produisait, ce serait assez grave. Vous pourriez mourir, la gravité serait donc assez élevée, mais elle n'aurait qu'une portée personnelle, car dans un seul accident de voiture, seul un petit nombre d'individus est touché. Cependant, il existe d'autres risques dont la gravité serait encore plus grande. Prenons l'exemple de l'élevage industriel. Nous pensons généralement que, par exemple, la vie des poulets dans les cages de batterie est si mauvaise qu'il serait préférable de ne pas créer ces poulets en premier lieu. C'est pourquoi nous pensons que c'est une bonne chose que la plupart des aliments proposés lors de cette conférence soient végétaliens. Une autre façon de voir les choses est que je suppose que certains d'entre vous trouveraient que la perspective d'être torturé pour le reste de votre vie est probablement encore pire qu'un accident de voiture mortel. Il peut donc y avoir des risques qui sont encore plus graves que des risques terminaux tels qu'un accident de voiture mortel. Et comme les risques sont maintenant des risques qui, en ce qui concerne leur gravité, sont à peu près aussi mauvais que l'élevage industriel et qu'ils concernent des résultats qui seraient encore pires que la non-existence, mais qui auraient également une portée beaucoup plus grande qu'un accident de voiture ou même que l'élevage industriel. Et ils pourraient potentiellement affecter un très grand nombre d'êtres dans un avenir lointain, dans l'ensemble de l'univers.

Cela explique pourquoi, dans le titre de mon exposé, j'ai affirmé que les risques S sont les pires risques existentiels. J'ai dit cela parce que je viens de les définir comme des risques de résultats qui ont la pire gravité et la pire étendue possibles. Pour comprendre cela et savoir en quoi ils diffèrent des autres types de risques existentiels, il suffit de zoomer sur le coin supérieur droit de la figure que j'ai montrée précédemment. C'est dans ce coin que figurent les risques existentiels.

Il s'agit de risques qui affecteraient au moins toutes les personnes vivant sur terre ainsi que toutes les générations futures. C'est pourquoi Boston les appelle des risques d'étendue pan-générationnelle et des risques qui seraient au moins ce que Boston appelle écrasants, ce que nous pouvons comprendre grosso modo comme l'élimination de tout ce qui serait précieux pour ces individus. L'un des principaux exemples de ces risques existentiels est le risque d'extinction. Ce sont des risques qui ont déjà fait l'objet d'une grande attention de la part de la communauté de l'AE. Ils ont une portée pangénérationnelle parce qu'ils affecteraient toutes les personnes en vie et élimineraient également le reste de l'avenir, et ils seraient écrasants parce qu'ils supprimeraient tout ce qui a de la valeur. Mais les risques S sont un autre type de risque existentiel qui sont également inclus conceptuellement dans ce concept de risque existentiel. Il s'agit de risques qui seraient encore pires que l'extinction parce qu'ils contiennent beaucoup de choses que nous dévalorisons, comme par exemple une souffrance involontaire intense, et de risques qui auraient une étendue encore plus grande parce qu'ils affecteraient une partie importante de l'univers.

On peut donc penser à l'histoire de Black Mirror du début et imaginer que Greta endure son isolement pour le reste de sa vie et qu'il ne s'agit pas d'un seul esprit téléchargé, mais d'une large population d'esprits téléchargés de ce type dans tout l'univers. On peut aussi penser à quelque chose comme l'élevage industriel avec une étendue beaucoup plus grande, pour une raison ou une autre, réalisée de bien des façons dans toute la galaxie. J'ai donc expliqué ce que sont les risques S d'un point de vue conceptuel. Il s'agit de risques de souffrances involontaires graves à l'échelle cosmique, dépassant ainsi le total des souffrances que nous avons connues sur terre jusqu'à présent. Cela en fait une sous-classe du risque existentiel, mais une sous-classe distincte des risques d'extinction plus connus. Jusqu'à présent, je me suis contenté de définir un terme conceptuel. J'ai attiré l'attention sur une certaine forme de possibilité.

Mais ce qui est peut-être plus pertinent, c'est de savoir, en tant qu'altruistes efficaces, si la réduction des risques S est quelque chose que nous pouvons faire et, le cas échéant, si c'est quelque chose que nous devrions faire. Et assurons-nous de bien comprendre cette question. En effet, tous les points de vue éthiques plausibles s'accordent à dire que la souffrance intense et involontaire est une mauvaise chose. J'espère donc que vous êtes tous d'accord pour dire que la réduction des risques S est une bonne chose. Mais bien sûr, vous êtes ici parce que vous vous intéressez à l'altruisme efficace. En d'autres termes, vous ne voulez pas seulement savoir s'il y a quelque chose de bon à faire ; nous sommes plutôt intéressés par l'identification du plus grand bien que nous puissions faire. Nous sommes conscients que faire le bien a un coût d'opportunité et nous voulons vraiment nous assurer de concentrer notre temps et notre argent sur l'impact le plus important que nous pouvons avoir. La question qui se pose ici, et dont j'aimerais que vous discutiez, est donc la suivante : la réduction des risques S peut-elle répondre à cette exigence plus élevée ? Pour certains d'entre nous au moins, serait-il préférable d'investir notre temps ou notre argent dans la réduction des risques S plutôt que de faire d'autres choses qui pourraient être bénéfiques dans un certain sens ? Il s'agit bien sûr d'une question très difficile à laquelle je ne pourrai pas répondre de manière concluante et exhaustive aujourd'hui. Afin d'illustrer la complexité de cette question et de préciser clairement le type d'argument que je n'avance pas ici, je vais d'abord présenter un argument erroné, un argument qui ne fonctionne pas pour se concentrer sur la réduction des risques S. Cet argument sera grosso modo le suivant :

Première prémisse, la meilleure chose à faire est de prévenir les pires risques. Deuxième prémisse : les risques S étant par définition les pires risques, on peut en conclure que la meilleure chose à faire est de prévenir les risques S. En ce qui concerne la première prémisse, éliminons une source potentielle de malentendu. L'une des façons de comprendre cette première prémisse est qu'elle pourrait être une caractéristique fondamentale de votre vision éthique du monde.

Ainsi, vous pourriez penser que, quelles que soient vos prévisions pour l'avenir, vous avez des raisons éthiques supplémentaires spécifiques de vous concentrer sur la prévention des pires résultats. Une sorte de principe du maximum ou peut-être de prioritarisme appliqué à l'avenir lointain. Toutefois, ce n'est pas de cela que je vais parler aujourd'hui. Donc, si votre vision éthique contient de tels principes, je pense qu'ils vous donnent des raisons supplémentaires de vous concentrer sur les risques S, mais ce n'est pas ce dont je vais parler. Ce dont je vais parler, c'est qu'il existe davantage de critères pertinents pour identifier l'action optimale d'un point de vue éthique que les deux dimensions de risques que nous avons examinées jusqu'à présent.

En effet, jusqu'à présent, nous nous sommes uniquement intéressés à la réalisation d'un risque, à la gravité et à l'étendue de ses conséquences. En ce sens, les risques S sont les pires risques, mais en ce sens, je pense que la première prémisse n'est pas clairement vraie. En effet, lorsqu'il s'agit de décider de la meilleure chose à faire, d'autres critères entrent en ligne de compte et nombre d'entre vous les connaissent bien, car ils suscitent à juste titre beaucoup d'attention au sein de la communauté de l'AE. Ainsi, pour déterminer si la réduction des risques S est la meilleure chose à faire, nous devons vraiment examiner la probabilité que ces risques S se concrétisent, la facilité avec laquelle il est possible de les réduire, en d'autres termes, le potentiel d'amélioration et le caractère négligé de cette entreprise. Y a-t-il déjà beaucoup de personnes ou d'organisations qui s'y emploient, quelle est l'attention portée à cette question ? Ces critères sont donc tout à fait pertinents. Même si vous êtes un prioritariste ou autre et que vous pensez avoir de nombreuses raisons de vous concentrer sur les pires résultats, si par exemple leur probabilité était nulle ou s'il n'y avait absolument rien à faire pour les réduire, cela n'aurait aucun sens d'essayer de le faire. Nous devons donc parler de la probabilité, du potentiel d'amélioration et du caractère négligé de ces risques, et je présenterai quelques réflexions initiales à ce sujet dans la suite de l'exposé. Qu'en est-il de la probabilité de ces risques S ? Je soutiendrai ici que les risques S ne sont pas beaucoup plus improbables que les risques d'extinction par une IA superintelligente, qui sont une catégorie de risques qu'au moins une partie de la communauté prend au sérieux et considère que nous devrions faire quelque chose pour y remédier.

J'expliquerai pourquoi je pense que c'est vrai et je répondrai à deux types d'objections que vous pourriez avoir. Il y a donc des raisons de penser que ces risques sont en fait trop peu probables pour qu'on s'y intéresse. La première objection pourrait être que ces risques S sont tout simplement trop absurdes. Nous ne sommes même pas encore en mesure d'envoyer des humains sur Mars, alors pourquoi devrions-nous nous inquiéter de souffrances à l'échelle cosmique, pourrait-on penser.

En effet, lorsque j'ai été confronté pour la première fois à ces idées, j'ai eu une réaction intuitive immédiate similaire : c'est un peu spéculatif, ce n'est peut-être pas quelque chose

sur lequel je devrais m'attarder. Mais je pense que nous devrions vraiment être prudents en examinant ces intuitions intuitives car, comme beaucoup d'entre vous le savent sans doute, il existe un grand nombre de recherches psychologiques dans le domaine de l'heuristique et des biais qui suggèrent que les évaluations intuitives de la probabilité par les humains sont souvent motivées par la facilité avec laquelle nous nous souvenons d'un exemple prototypique du type de scénarios que nous envisageons. Pour les choses qui ne se sont jamais produites et pour lesquelles il n'y a pas de précédent historique, cela nous conduit à sous-estimer systématiquement leur probabilité, ce que l'on appelle l'heuristique de l'absurdité. Je pense donc que nous ne devrions pas nous contenter de cette réaction intuitive, mais que nous devrions plutôt nous demander ce que nous pouvons dire au sujet de la probabilité de ces risques S.

Et si nous examinons toutes nos meilleures théories scientifiques et ce que les experts disent sur la façon dont l'avenir pourrait se dérouler, je pense que nous pouvons identifier deux développements technologiques pas trop invraisemblables qui peuvent plausiblement conduire à la réalisation des risques S. Cela ne veut pas dire que ce sont les seules possibilités, il peut y avoir des inconnues que nous ne pouvons pas encore prévoir et qui pourraient également conduire à de tels risques, mais il y a des voies connues qui pourraient nous amener dans un territoire à risque. Il s'agit de la sentience artificielle et de l'IA super intelligente. La sentience artificielle renvoie simplement à l'idée que la capacité d'avoir une expérience subjective, et en particulier la capacité de souffrir, n'est en fait pas limitée en principe aux animaux biologiques. Mais qu'il pourrait exister de nouveaux types d'êtres, peut-être des programmes informatiques stockés sur du matériel à base de silicium, dont la souffrance nous intéresserait également. Et bien que cette question ne soit pas complètement réglée, peu de points de vue contemporains et de philosophes de l'esprit diraient que la sentience artificielle est impossible en principe. Il semble donc qu'il s'agisse d'une possibilité conceptuelle dont nous devrions nous préoccuper. Maintenant, quelle est la probabilité que cela se réalise un jour ? C'est peut-être moins clair, mais en fait, ici aussi, nous pouvons identifier une voie technologique qui pourrait mener à la sentience artificielle, et c'est l'idée de l'émulation du cerveau entier. En fait, il s'agit de comprendre le cerveau humain de manière suffisamment détaillée pour pouvoir en construire une simulation informatique fonctionnellement équivalente. Pour cette technologie, il n'est pas encore tout à fait certain que nous y parviendrons, mais des chercheurs se sont penchés sur la question et ont tracé une feuille de route assez détaillée pour cette technologie. Ils ont identifié des étapes concrètes et les incertitudes qui subsistent et ont conclu qu'il s'agissait d'un élément dont nous devrions tenir compte lorsque nous pensons à l'avenir. Je dirais donc qu'il existe une possibilité technologique pas trop invraisemblable que nous parvenions à la sentience artificielle.

Je ne m'étendrai pas sur le second développement, l'IA super intelligente, car il fait déjà l'objet d'une grande attention de la part de la communauté des AE. Si vous n'êtes pas familier avec les inquiétudes liées à l'IA super intelligente, je vous recommande l'excellent livre de Nick Bostrom, Superintelligence, et j'ajouterai simplement que l'IA super intelligente pourrait probablement aussi débloquer de nombreuses autres capacités technologiques dont nous aurions besoin pour entrer dans le territoire des risques S. Par exemple, la capacité de coloniser l'espace et de répandre des êtres sensibles dans de plus grandes parties de l'univers. J'aimerais également ajouter que certains scénarios dans lesquels l'interaction entre l'IA super intelligente et la sentience artificielle pourrait conduire à des

scénarios de risques S ont été discutés par Bostrom dans Superintelligence et d'autres endroits sous le terme de crime contre l'esprit. Vous pouvez faire une recherche à ce sujet si vous êtes intéressé par des idées connexes.

En fait, si nous examinons ce que nous pouvons dire de l'avenir, je pense que ce serait une erreur de dire que les risques S sont si peu probables que nous ne devrions pas nous en préoccuper. Mais peut-être avez-vous maintenant une objection différente. Vous êtes peut-être convaincu qu'en termes de capacités technologiques, nous ne pouvons pas être sûrs que ces risques S sont tout simplement trop improbables, mais vous pouvez penser que de vastes quantités de souffrance semblent être une issue assez spécifique, même si nous avons des capacités technologiques beaucoup plus grandes, il semble peu probable qu'une issue aussi mauvaise se produise. Vous pourriez donc vous dire qu'après tout, cela nécessiterait une sorte d'agent maléfique, une sorte d'intention maléfique qui s'efforcerait activement de faire en sorte que nous obtenions ces vastes quantités de souffrance. Je pense que je suis d'accord pour dire que cela semble assez improbable, mais là encore, après avoir réfléchi un peu, je pense que nous pouvons voir que ce n'est qu'une voie, et peut-être la plus invraisemblable, pour entrer dans le territoire des risques S. Il y a également deux autres voies que j'aimerais défendre.

Le premier de ces risques S pourrait survenir par accident. Ainsi, une catégorie de scénarios pourrait être la suivante. Imaginons que les premiers êtres artificiellement sentients que nous créons ne soient pas aussi développés que des esprits humains complets, mais peut-être plus semblables à des animaux non humains, en ce sens que nous pourrions créer des êtres artificiellement sentients capables de souffrir, mais avec une capacité limitée de communiquer avec nous et de signaler qu'ils souffrent. Dans un cas extrême, nous pouvons créer des êtres sensibles qui peuvent souffrir mais dont nous ignorons la souffrance parce qu'il n'y a pas de possibilité de communication facile.

Un deuxième scénario dans lequel les risques S pourraient être réalisés sans intention malveillante est l'exemple du maximisateur de trombones, qui sert à illustrer l'idée de ce qui se passerait si nous créions une IA super intelligente très puissante qui poursuivrait un objectif sans rapport avec le nôtre. Un objectif qui n'est ni étroitement aligné sur nos valeurs, ni activement maléfique. Et comme Nick Bostrom et de nombreuses personnes l'ont affirmé, il est concevable qu'un tel maximiseur de trombones puisse conduire à l'extinction de l'humanité, par exemple, parce qu'il convertirait la terre entière et toute la matière qui l'entoure en trombones, parce qu'il veut simplement maximiser le nombre de trombones et n'a aucune considération pour la survie de l'humanité. Mais il n'y a qu'un pas à franchir pour s'inquiéter : et si un tel maximiseur de trombones effectuait des simulations sensibles, par exemple à des fins scientifiques, pour mieux comprendre comment maximiser la production de trombones, ou si, de la même manière que notre souffrance remplit une fonction évolutive, un maximiseur de trombones créait des sous-programmes ou des travaux artificiellement sensibles dont la souffrance serait instrumentalement utile pour maximiser la production de trombones ? Il suffit donc d'ajouter quelques exemples et hypothèses supplémentaires pour constater que les scénarios qui font déjà l'objet d'une grande attention pourraient non seulement conduire à l'extinction de l'humanité, mais aussi à des résultats encore pires. Enfin, pour comprendre l'importance de la troisième voie par laquelle les risques S pourraient être réalisés dans le cadre d'un conflit, il convient de noter que si un grand nombre d'agents sont en concurrence pour des ressources partagées, cela peut

encourager des sous-dynamiques négatives qui conduisent à de très mauvais résultats, même si aucun des agents concernés ne valorise activement ces mauvais résultats, mais y a simplement recours pour surpasser les autres agents. Par exemple, si l'on considère la plupart des guerres, les pays qui les mènent accordent rarement une valeur intrinsèque à la souffrance et à la violence qu'elles impliquent, mais il arrive que les guerres aient lieu pour servir les intérêts stratégiques des pays concernés.

Je pense donc que si nous examinons d'un œil critique la situation dans laquelle nous nous trouvons, nous devrions conclure qu'en fait, si nous prenons au sérieux un grand nombre de considérations qui sont déjà largement étudiées par la communauté, telles que les risques liés aux IA superintelligentes, il n'y a que quelques hypothèses supplémentaires dont nous avons besoin pour justifier les inquiétudes liées aux risques S, et ce n'est pas comme si nous avions besoin d'inventer des technologies entièrement nouvelles ou de supposer des motivations extrêmement invraisemblables ou rares, telles que le sadisme ou la haine, pour justifier les inquiétudes liées aux risques S. C'est pourquoi j'ai dit que je pense que les risques S ne sont pas beaucoup plus improbables que, par exemple, les risques d'extinction dus à une IA super intelligente. Bien entendu, la probabilité des risques S n'est pas le seul critère à prendre en compte. Comme je l'ai dit, nous devons également nous demander s'il est facile de réduire ces risques S. Et en fait, je pense qu'il s'agit d'une tâche assez difficile. Nous n'avons pas encore trouvé de solution miracle, mais j'aimerais également affirmer que la réduction des risques S est au moins un minimum réalisable, même aujourd'hui, et l'une des raisons en est que l'on peut dire que nous réduisons déjà ces risques. Ainsi, comme je viens de le dire, certains scénarios de réalisation des risques S prévoient qu'une IA super intelligente fasse fausse route d'une manière ou d'une autre.

C'est pourquoi certains travaux sur la sécurité technique de l'IA ainsi que sur la politique en matière d'IA réduisent probablement déjà efficacement les S-Risques. Pour vous donner un exemple, j'ai dit que nous pourrions nous inquiéter de l'apparition de risques S en raison du comportement stratégique d'agents d'IA dans le cadre d'un conflit. Certains travaux en matière de politique d'IA qui réduisent la probabilité de tels scénarios d'IA multipolaires et rendent plus probables les scénarios d'IA unipolaires avec moins de concurrence pourraient notamment avoir pour effet de réduire les risques S. Il en va de même pour certains travaux relatifs à la sécurité technique de l'IA. Cela étant dit, il me semble qu'un grand nombre des interventions actuellement entreprises réduisent les risques S par accident, dans un sens, elles ne sont pas spécifiquement conçues pour réduire les risques S et il pourrait bien y avoir des sous-problèmes particuliers au sein de la sécurité technique de l'IA qui seraient particulièrement efficaces pour réduire les risques S et qui ne reçoivent pas encore beaucoup d'attention. Ainsi, pour vous donner un exemple qui est probablement difficile à réaliser sous cette forme précise, mais qui illustre ce qui pourrait être possible. On pourrait essayer de garantir qu'une IA incontrôlée, en partant du principe que nos efforts pour résoudre le problème du contrôle échouent, que l'IA ne crée pas de simulations sensibles supplémentaires ou de sous-programmes artificiellement sentients. Si nous pouvions résoudre ce problème en travaillant sur la sécurité technique de l'IA, nous pourrions sans doute réduire les risques S de manière spécifique.

Bien sûr, il existe également des interventions plus générales qui ne visent pas directement à influencer certains types de leviers qui affectent directement l'avenir lointain, mais qui auraient un effet plus indirect sur la réduction des risques S. Ainsi, nous pourrions penser

que le renforcement de la coopération internationale nous permettra à un moment donné, par exemple, d'empêcher les courses à l'armement en matière d'IA, qui pourraient à nouveau conduire à une dynamique de somme négative susceptible de nous conduire en territoire à risque S. De même, la sentience artificielle étant une préoccupation majeure dans la réflexion sur les risques S, nous pourrions penser que le fait d'élargir le cercle moral et de rendre plus probable le fait que les décideurs humains se préoccupent à l'avenir des êtres artificiellement sentients aurait un effet positif sur la réduction du risque S. Cela étant dit, je pense qu'il est juste de dire que nous ne comprenons pas très bien, à l'heure actuelle, comment réduire au mieux le risque S. Si nous pensons qu'il y a de bonnes choses à récolter, nous pourrions nous dire : d'accord, mettons-nous en position méta et faisons des recherches sur la meilleure façon de réduire ces risques S. C'est d'ailleurs une grande partie de ce que nous faisons Foundation Research Institute. Il y a aussi un autre aspect du potentiel d'amélioration dont j'aimerais parler. Il ne s'agit pas de savoir s'il est intrinsèquement facile de réduire les risques S, mais de savoir si nous pouvons obtenir le soutien nécessaire. Par exemple, pouvons-nous obtenir un financement suffisant pour faire décoller les travaux sur la réduction des risques S ? L'une des inquiétudes que nous pouvons avoir ici est que toutes ces discussions sur la souffrance à l'échelle cosmique, etc. semblent trop improbables pour beaucoup de gens, en d'autres termes que les risques S soient une préoccupation trop étrange pour que nous puissions obtenir un soutien et un financement significatifs pour les réduire.

Je pense que cette inquiétude est légitime dans une certaine mesure, mais je ne pense pas non plus qu'il faille être trop pessimiste et je pense que l'histoire du domaine de la sécurité de l'IA corrobore cette évaluation. Il y a dix ans, les inquiétudes concernant le risque d'extinction lié à une IA super intelligente étaient ridiculisées, rejetées, mal interprétées et mal comprises, comme s'il s'agissait par exemple de Terminator ou d'autres choses de ce genre. Aujourd'hui, Bill Gates publie un livre dans lequel il parle ouvertement et directement des risques d'une IA super intelligente et de concepts connexes tels que le crime contre l'esprit. Je dirais donc que l'histoire récente du domaine de la sécurité de l'IA donne des raisons d'espérer que nous sommes capables de pousser même des causes apparemment bizarres suffisamment loin dans le courant dominant, dans la fenêtre d'un discours acceptable, pour pouvoir susciter un soutien significatif en leur faveur.

Enfin et surtout, qu'en est-il du caractère négligé du risque S ? Comme je l'ai dit, certains travaux déjà en cours dans le domaine du risque X permettent de réduire le risque S. La réduction du risque S n'est donc pas totalement négligée, mais je pense qu'il est juste de dire qu'elle reçoit beaucoup moins d'attention que, par exemple, le risque d'extinction. En fait, j'ai parfois vu des membres de la communauté assimiler explicitement ou implicitement le risque existentiel et le risque d'extinction, ce qui, d'un point de vue conceptuel, semble clairement faux. En fait, alors que certaines interventions existantes peuvent également être efficaces pour réduire les risques S, peu de personnes tentent spécifiquement d'identifier les interventions les plus efficaces pour réduire les risques S en particulier. Et je pense que le Foundation Research Institute est la seule organisation de l'AE qui a pour mission organisationnelle de se concentrer sur la réduction du risque S. En résumé, je n'ai pas répondu de manière concluante à la question de savoir pour qui exactement la réduction du risque S est la meilleure chose à faire. Je pense que cela dépend à la fois de votre point de vue éthique et de certaines questions empiriques telles que la probabilité, le potentiel d'amélioration et le caractère négligé du risque S. Mais j'ai soutenu que les risques S ne

sont pas beaucoup plus improbables que, par exemple, le risque d'extinction par une IA super intelligente, et qu'ils méritent donc au moins une certaine attention. Et j'ai soutenu que la voie connue la plus plausible qui pourrait nous conduire vers le territoire du risque S, à part les inconnues inconnues, est celle des scénarios d'IA qui impliquent la création d'un grand nombre d'êtres artificiellement sentients. C'est pourquoi je pense que, parmi les sources actuellement connues de risque existentiel, le domaine du risque de l'IA est unique en ce qu'il est également très pertinent pour réduire le risque S. En effet, si nous ne parvenons pas à maîtriser l'IA, il semble qu'il y ait une forte probabilité que nous entrions dans le territoire du risque S, alors que dans d'autres domaines, par exemple un astéroïde frappant la terre ou une pandémie mortelle ou l'anéantissement d'une vie humaine, il semble beaucoup moins probable que cela puisse nous entraîner dans des scénarios qui seraient bien pires que l'extinction parce qu'ils contiennent en outre beaucoup de souffrance. En ce sens, si vous n'avez jamais pris en compte le risque S, je pense qu'il s'agit d'une mise à jour qui vous permettra de vous préoccuper davantage du domaine des risques liés à l'IA par rapport à d'autres domaines liés au risque S. En ce sens, une partie, mais pas la totalité, du travail actuel dans le domaine du risque X est déjà efficace pour réduire le risque S, mais il semble qu'il y ait un manque de personnes et de recherches optimisant spécifiquement la réduction du risque S et essayant de trouver les interventions qui sont les plus efficaces pour cet objectif particulier. Je dirais que le Foundation Research Institute occupe un créneau important en raison de son orientation unique et j'aimerais beaucoup que d'autres personnes se joignent à nous dans ce créneau. J'aimerais que d'autres personnes nous rejoignent sur ce créneau, c'est-à-dire des personnes issues d'autres organisations qui mènent également des recherches qui, espérons-le, seront efficaces pour réduire le risque S. Ceci étant dit, j'espère avoir sensibilisé les gens à la perspective inquiétante des risques S. Je ne pense pas vous avoir tous convaincus que la réduction des risques S est la meilleure façon d'utiliser vos ressources.

Je ne pense pas que je puisse m'attendre à cela, à la fois parce que nos points de vue éthiques fondamentaux diffèrent dans une certaine mesure et aussi parce que les questions empiriques en jeu sont tout simplement extrêmement complexes et qu'il semble très difficile de parvenir à un accord à leur sujet. Je pense donc que ceux d'entre nous qui souhaitent façonner l'avenir et qui sont convaincus qu'il s'agit de la chose la plus importante à faire seront confrontés à une situation dans laquelle des personnes auront des priorités différentes au sein de la communauté et nous devrons trouver un moyen de gérer cette situation. C'est pourquoi j'aimerais terminer cet exposé par une vision de cette communauté qui façonne l'avenir lointain. Ainsi, on peut voir le fait de façonner l'avenir lointain comme un long voyage. Mais ce que j'espère avoir fait comprendre, c'est qu'il est erroné de présenter ce voyage comme impliquant un choix binaire entre l'extinction ou l'utopie. Dans un autre sens, cependant, je dirais que cette métaphore est appropriée. Nous sommes effectivement confrontés à un long voyage, mais il s'agit d'un voyage à travers un territoire difficile à traverser et, à l'horizon, il y a un continuum allant d'un très mauvais orage à une magnifique journée d'été. L'intérêt pour l'avenir lointain détermine en quelque sorte qui est avec nous dans le véhicule, mais il ne répond pas nécessairement à la question de savoir ce qu'il faut faire plus précisément avec le volant.

Certains d'entre nous sont plus préoccupés par le fait de ne pas tomber dans l'orage, d'autres sont plus motivés par l'espoir existentiel d'arriver peut-être à ce beau soleil. Il semble difficile de se mettre d'accord sur ce qu'il faut faire plus précisément, notamment

parce qu'il est très difficile de suivre les réseaux complexes de routes qui s'étendent loin devant nous et de savoir quelle direction prendre pour obtenir quel résultat précis. En revanche, il nous est facile de voir qui se trouve avec nous dans le véhicule. J'aimerais donc conclure en disant que l'une des choses les plus importantes que nous puissions faire est de comparer nos cartes entre les personnes présentes dans le véhicule et de trouver un moyen de gérer les désaccords restants sans faire dérailler le véhicule par inadvertance et d'arriver à un résultat qui soit bénéfique pour tout le monde. Je vous remercie de votre attention.

Je commencerai par une question que quelques personnes ont posée et qui, selon vous, n'est pas nécessaire pour se préoccuper des risques S, mais qui permettrait d'y voir un peu plus clair : outre l'IA, outre l'émulation d'un cerveau entier et l'utilisation de cerveaux téléchargés, existe-t-il d'autres formes de risques S que vous puissiez essayer de visualiser ? En particulier, les gens essayaient de trouver des moyens de travailler sur le problème si vous n'avez pas une idée concrète de la manière dont il pourrait se manifester. Je pense donc que les scénarios artificiels les plus plausibles que nous pouvons envisager aujourd'hui impliquent une sentience artificielle, en partie parce que de nombreuses personnes ont parlé du fait que la sentience artificielle s'accompagnerait de nouveaux défis – par exemple, il serait vraisemblablement très facile d'engendrer un grand nombre d'êtres artificiellement sentients. Les substrats à base de silicium présentent de nombreux avantages en matière d'efficacité par rapport aux substrats biologiques et nous pouvons également observer que dans de nombreux autres domaines, les pires résultats contiennent la majeure partie de la valeur espérée, comme la prédominance des distributions à queue lourde, par exemple en ce qui concerne les pertes humaines dans les guerres, les maladies et ainsi de suite. Il me semble donc assez plausible que si nous voulons réduire autant que possible les souffrances attendues, nous devrions nous concentrer sur ces très mauvais résultats et que la plupart d'entre eux, pour diverses raisons, impliquent une sentience artificielle. Cela étant dit, je pense qu'il existe certains scénarios, en particulier dans les scénarios futurs où nous n'avons pas ce scénario archétypique d'explosion de l'intelligence et de décollage abrupt, où nous sommes confrontés à un avenir plus désordonné et complexe où il y a peut-être de nombreuses factions qui contrôlent l'IA et l'utilisent à diverses fins, et où nous pourrions faire face à des risques qui ne sont peut-être pas aussi importants que les pires scénarios impliquant la sentience artificielle, mais qui s'apparenteraient peut-être davantage à l'élevage industriel. Une sorte de nouvelle technologie qui serait utilisée à mauvais escient, peut-être simplement parce que les gens ne se soucient pas suffisamment des conséquences et qu'ils poursuivent des objectifs économiques et créent par inadvertance de grandes quantités de souffrance, comme c'est le cas aujourd'hui, par exemple, dans l'industrie animale.

Vous avez dit que le débat n'est toujours pas tranché sur la question de savoir si l'on peut ou non étendre sa préoccupation morale à quelque chose qui se trouve dans un substrat de silicium et qui n'est pas fait de chair et d'os comme nous le sommes. Pouvez-vous expliquer pourquoi nous pourrions en fait nous préoccuper de quelque chose qui est un cerveau téléchargé et qui n'est pas un cerveau au sens où nous l'entendons généralement ? Une expérience de pensée suggestive qui a été discutée en philosophie de l'esprit consiste à imaginer que l'on remplace son cerveau non pas d'un seul coup, mais pas à pas, par une machine à base de silicium. Vous commencez donc par remplacer un seul neurone par une sorte de puce qui remplit la même fonction. Il semble intuitivement clair que cela ne vous rend pas moins sentient ou que nous devrions moins nous préoccuper de vous de cette

manière. Vous pouvez maintenant imaginer remplacer progressivement votre cerveau, neurone par neurone, par un ordinateur en quelque sorte. Et il semble que vous ayez du mal à mettre le doigt sur un point particulier de cette transition où vous diriez que la situation s'est inversée et que nous devrions cesser de nous préoccuper du même traitement de l'information qui se déroule toujours dans ce cerveau. Oui, mais il existe un grand nombre d'ouvrages de philosophie de l'esprit qui traitent de cette question. Et en supposant que ces cerveaux aient effectivement la capacité de souffrir, quelle raison aurions-nous de penser qu'il serait avantageux pour une superintelligence d'émuler de nombreux cerveaux de manière à ce qu'ils souffrent plutôt que de les laisser simplement exister sans aucune sorte de sentiment positif ou négatif.

L'une des raisons pour lesquelles nous pouvons être inquiets est que si nous examinons les succès actuels de l'IA, nous constatons qu'ils sont souvent dus à des techniques d'apprentissage automatique. Il s'agit de techniques pour lesquelles nous ne programmons pas les connaissances et les capacités. Nous pensons qu'il vaut mieux mettre en place une sorte d'algorithme qui peut être entraîné et qui peut apprendre par essais et erreurs en recevant des informations sur la qualité ou la faiblesse de ses résultats et en augmentant ainsi ses capacités. Il semble très peu probable que les techniques actuelles d'apprentissage automatique, qui impliquent de donner des signaux de récompense aux algorithmes, nous concernent dans une large mesure. Je ne veux pas prétendre que les algorithmes actuels d'apprentissage par renforcement souffrent dans une large mesure, mais nous pouvons craindre que des architectures similaires où les capacités d'êtres artificiellement sentients se développent parce qu'ils sont formés à certaines choses en recevant une sorte de signal de récompense soient une caractéristique des systèmes d'IA qui persistera même à un moment où la sentience de ces algorithmes se réalisera dans une plus large mesure. D'une certaine manière, cela ressemble à la façon dont, comme je l'ai mentionné, notre souffrance remplit une fonction évolutive qui nous aide à naviguer dans le monde et, en fait, les personnes qui ne ressentent pas la douleur ont beaucoup de difficultés pour cette raison parce qu'elles n'évitent pas intuitivement les résultats dommageables. Il s'agit certainement d'une discussion plus longue, mais j'espère que vous pourrez donner une brève réponse à cette question.

Quelques personnes voulaient également savoir, vous vous concentrez sur la souffrance dans le cas du risque S, mais certaines personnes se demandent si un agent ne pourrait pas tout simplement préférer, on ne sait pas s'il préférerait la mort ou la souffrance, il pourrait en fait préférer exister même si ses expériences sont plutôt négatives. S'agit-il d'un choix que vous feriez au nom des agents que vous considérez dans votre domaine moral lorsque vous essayez d'atténuer un risque S, s'agit-il d'une condition préalable nécessaire pour se préoccuper des risques S ? Je pense donc que, quelles que soient vos opinions éthiques fondamentales, il existe presque toujours des raisons prudentielles de prendre en compte les préférences d'autres agents. Ainsi, si j'étais confronté à une situation où je me disais qu'il existe une sorte d'être dont les expériences sont si négatives que je pense, d'un point de vue conséquentialiste, qu'il vaudrait mieux qu'il n'existe pas, mais que cet être a, pour une raison ou une autre, une forte préférence pour l'existence et qu'il me demande si je dois continuer ou non, et ainsi de suite. Je pense qu'il y a souvent des raisons prudentielles de prendre en compte ces préférences. Je pense qu'il y aura une certaine convergence entre les différents points de vue éthiques sur la question de la prise en compte de ces préférences hypothétiques.

Ceci étant dit, je pense qu'il est assez peu plausible d'affirmer qu'aucune quantité imaginable de souffrance ne serait intrinsèquement pire que la non-existence. Cela me semble assez peu plausible, de sorte qu'une mini expérience de pensée pour penser à cela serait d'imaginer qu'on vous offre un choix entre une heure de sommeil et une heure de torture. Que préférez-vous ? Pour la plupart d'entre nous, il semble assez clair qu'une heure de sommeil sans aucune expérience est le meilleur choix. Vous avez dit qu'avec un peu de chance, nous parviendrons à une sorte de convergence sur ce qu'est la véritable philosophie morale, pour autant qu'il y en ait une, mais il y a peut-être aussi des raisons de penser que nous n'y parviendrons pas dans les délais nécessaires au développement d'une IA super intelligente ou au développement d'émulations de cerveaux entiers que nous pourrons faire tourner sur de nombreux ordinateurs. Que faisons-nous dans ce cas où nous n'avons pas résolu la philosophie morale à temps ? Je pense qu'il s'agit là d'une question très importante, car il me semble assez probable qu'il n'y aura pas de convergence, du moins dans les moindres détails.