Scruter la Singularité

Eliezer Yudkowsky

La version courte :

Si les vitesses de calcul doublent tous les deux ans, qu'arrive-t-il quand des IA informatisées font la recherche ?

La vitesse de calcul double tous les deux ans.
La vitesse de calcul double tous les deux ans de travail.
La vitesse de calcul double tous les deux ans *subjectifs* de travail.

Deux ans après que les Intelligences artificielles ont atteint l'équivalence humaine, leur vitesse double. *Un* an après, leur vitesse double de nouveau.

Six mois ; trois mois ; 1,5 mois... Singularité.

Prenez les valeurs correspondant aux vitesses de calcul actuelles, à leur temps de doublement actuel, et à une estimation de la puissance de calcul brute du cerveau humain, et les valeurs coïncident en : 2021.

Mais personnellement, j'aimerais le faire plus tôt.

1. La fin de l'Histoire

L'Histoire a commencé il y a trois milliards et demi d'années dans une flaque de boue, quand une molécule a fait une copie d'elle-même et est ainsi devenue l'ancêtre ultime de toute vie terrestre.

Elle a commencé il y a quatre millions d'années, quand les volumes cérébraux se sont mis à augmenter rapidement dans la lignée des Hominidés.

Il y a cinquante mille ans avec l'émergence d'*Homo sapiens sapiens*.
Il y a dix mille ans avec l'invention de la civilisation.
Il y a cinq cents ans avec l'invention de l'imprimerie.
Il y a cinquante ans avec l'invention de l'ordinateur.

Dans moins de trente ans, elle finira.

Un jour ou l'autre dans un proche avenir, quelqu'un trouvera une méthode pour augmenter l'intelligence maximale sur la planète, soit en programmant une vraie intelligence artificielle, soit en améliorant l'intelligence humaine. Un humain amélioré serait davantage capable de trouver des façons d'améliorer les humains ; il aurait une « capacité d'invention accrue ». À quoi cette capacité accrue serait-elle consacrée ? À créer la génération suivante d'humains améliorés.

Et que feraient ces esprits doublement améliorés ? Ils chercheraient des méthodes pour arriver à des humains triplement améliorés, ou ils construiraient des esprits d'IA fonctionnant à des vitesses d'ordinateur. Et une IA serait capable de *se reprogrammer elle-même*, directement, pour être plus rapide ; ou plus intelligente. Après quoi notre boule de cristal explose, « la vie comme nous la connaissons » est terminée, et tout ce que nous savons passe par la fenêtre.

> J'avais tenté ici une simple extrapolation de la technologie, et je m'étais retrouvé précipité au-dessus d'un abîme. C'est un problème auquel nous faisons face à chaque fois que nous envisageons la création d'une intelligence plus grande que la nôtre. Quand cela se produira, l'histoire humaine aura atteint une sorte de singularité — un point où l'extrapolation se brise et où de nouveaux modèles doivent être employés — et le monde dépassera notre compréhension.
>
> *True Names and Other Dangers*, page 47
> Vernor Vinge

Il y a différents chemins qui conduisent à la Singularité. La nanotechnologie, qui est la capacité de construire des ordinateurs atome par atome et de recâbler le cerveau neurone par neurone. L'intelligence artificielle, l'IA germe capable d'autocompréhension et d'autoamélioration. Nous pourrions amorcer notre chemin pour atteindre la Singularité par les améliorations humaines relativement douces produites par les interventions directes sur le système nerveux. Des interfaces directes neurone-silicium pourrait améliorer l'intelligence humaine ou l'intelligence artificielle ou les deux. Ou une percée totalement inattendue pourrait se produire.

Une civilisation disposant de technologie avancée est instable ; elle se termine quand l'espèce se détruit elle-même ou s'améliore elle-même. Si les tendances actuelles se poursuivent — si nous ne butons pas sur une limite théorique imprévue à l'intelligence, que nous ne transformons pas la Terre en désert radioactif, et que nous n'enterrons pas la planète sous un raz-de-marée de nanomachines voraces qui se reproduisent —, la Singularité est inévitable. L'estimation la plus souvent citée pour la Singularité est 2035 : de votre vivant ! Mais beaucoup, et moi compris, pensent que la Singularité pourrait se produire nettement plus tôt.

Un peu de terminologie, empruntée à A Fire Upon The Deep, le roman de Vernor Vinge lauréat du prix Hugo :

Puissance : une entité d'au-delà de la Singularité.

Transcender, Transcendé, Transcendance : l'action de se reprogrammer pour être plus intelligent, de se reprogrammer (avec sa nouvelle intelligence) pour être encore plus intelligent, et ainsi de suite *ad Singularitum*. La « Transcendance » est la zone métaphorique où vivent les Puissances.

Beyond : la zone grise entre l'état humain et celui de Puissance ; le domaine habité par les entités plus intelligentes que l'homme, mais ne possédant pas la technologie d'autoreprogrammation directe permettant de se transcender.

2. La Singularité nous dépasse

> J'imagine que les insectes et les filles entrevoient confusément que la Nature leur a joué un tour cruel, mais il leur manque l'intelligence nécessaire pour réaliser à quel point.
>
> *Calvin et Hobbes*

Mais pourquoi les Puissances devraient-elles être *tellement* plus que ce que nous sommes maintenant ? Pourquoi ne pas supposer que nous seront *un peu* plus intelligent, et voilà tout ?

Considérez la suite 1, 2, 4, 8, 16, 32. Considérez l'itération de F(x) = (x + x). Tous les deux ans, la performance des ordinateurs double [1]. C'est le taux démontré de progrès sous le contrôle d'esprits constants, non améliorés : le progrès selon les mortels.

En ce moment, l'importance de la puissance de calcul du silicium en réseau est légèrement au-dessus de la puissance d'un cerveau humain. La puissance d'un cerveau humain est de 10^17 op/s, ou cent millions de milliards d'opérations par seconde [2], contre un milliard environ d'ordinateur sur l'Internet ayant chacun entre 100 millions op/s et 1 milliard op/s. La quantité *totale* de puissance de calcul sur la planète est la quantité de puissance d'un cerveau humain, 10^17 op/s, multiplié par le nombre d'êtres humains, à ce jour six milliards ou 6x10^9. La quantité de puissance de calcul *artificielle* est si petite qu'elle est négligeable, pas parce qu'il y a tellement d'êtres humains, mais à cause de la simple puissance brute d'un seul cerveau humain.

Au taux de progrès ancien, quand ont été faits les calculs originaux à propos de la Singularité en 1988 [3], on s'attendait à ce que les ordinateurs atteignent des niveaux équivalents à l'être humain — 10^17 opérations en virgule flottante par seconde, ou cent petaflops — autour de 2035. Mais selon ce taux de progrès, les machines d'1 teraflops étaient attendues en 2000 ; il s'avère que les machines d'1 teraflops étaient disponibles en 1996, lors de la première rédaction de ce document. En 1998, la vitesse maximale était de 3,2 teraflops, et en 1999 IBM a annoncé le projet Blue Gene de construction d'une machine d'un petaflops pour 2005. Il se peut donc que les premières estimations soient *légèrement* modestes.

Une fois que nous aurons des ordinateurs de niveau humain, la quantité de puissance de calcul sur la planète sera égale au nombre d'humains *plus* le nombre d'ordinateurs. La quantité d'intelligence disponible fait alors un saut énorme. Dix ans plus tard, *les humains* deviennent quantité négligeable dans l'équation.

Cette séquence de doublement est en réalité une projection *pessimiste*, parce qu'elle suppose que la puissance de calcul continue à doubler selon le même rythme. Mais pourquoi supposer cela ? La vitesse des ordinateurs ne double pas en vertu de quelque loi physique inexorable, mais parce que des chercheurs et des ingénieurs trouvent le moyen des faire des puces plus rapides. Si une partie des chercheurs et des ingénieurs sont *eux-mêmes* des ordinateurs...

Un groupe d'ordinateurs de niveau humain prend 2 ans pour doubler la vitesse des ordinateurs. Il prennent ensuite 2 autres années *subjectives*, soit 1 année en termes humains, pour la doubler à nouveau. Puis ils prennent à nouveau 2 années subjectives, soit six mois, pour la doubler une nouvelle fois. Après un total de quatre ans, la puissance de calcul atteint l'infini.

C'est la version « Transcendée » de la suite de doublement. Appelons « version transcendée » d'une séquence a0, a1, a2... la fonction où l'intervalle entre an et an+1 est inversement proportionnel à an [4]. Une fonction de doublement transcendée commence donc par 1, et dans ce cas il faut 1 unité de temps pour arriver à 2. Il faut ensuite 1/2 unité de temps pour arriver à 4. Puis il faut 1/4 d'unité de temps pour arriver à 8. Cette fonction, si elle était continue, serait la fonction hyperbolique y = 2/(2-x). Lorsque x = 2, alors (2 - x) = 0 et y = infini. Le comportement sur ce point est appelé en mathématique une singularité.

Et la séquence de doublement transcendée est *aussi* une projection pessimiste, pas du tout une Singularité, parce qu'elle suppose que seule la *vitesse* est améliorée. Et si la *qualité* de la pensée était améliorée ? Maintenant, deux ans de travail — ou plutôt, en ce moment, dix-huit mois de travail. Dix-huit mois de travail subjectif suffisent à double les vitesses de calcul. Est-ce que cela ne devrait pas s'améliorer un peu avec le partage des pensées et les souvenirs eidétiques ? Est-ce que cela ne devrait pas s'améliorer si, disons, la somme totale de la connaissance scientifique humaine était stockée dans un format prédigéré, cognitif, prêt-à-penser ? Est-ce que cela ne devrait pas s'améliorer avec une mémoire à court terme capable de conserver la totalité de la connaissance humaine ? Une IA de niveau humain *n'est pas de niveau humain* : si Kasparov avait eu ne serait-ce que le plus petit, le plus mesquin des programmes de jeu d'échecs solidement intégré avec ses intuitions, il aurait laminé Deep Blue. C'est en cela que consiste l'avantage de l'IA : les tâches simples sont réalisées à des vitesses phénoménales et sans erreur, les tâches conscientes sont réalisées avec une mémoire parfaite et une conscience de soi totale.

Je n'ai même pas *abordé* le sujet des IA qui reconçoivent leur architecture cognitive, et il leur sera beaucoup plus facile de le faire que ça ne l'est pour nous, spécialement si elles peuvent faire des sauvegardes. Il se pourrait que le doublement transcendé se heurte aux lois de la phy-

sique avant d'atteindre l'infini... mais même les lois de la physique *telles que nous les comprenons aujourd'hui* permettraient à un gramme (plus ou moins) de stocker et de faire tourner la totalité de l'espèce humaine au rythme d'un million d'années subjectives par seconde [5].

Respirons profondément et pensons à cela un instant. Un gramme. La *totalité* de l'espèce humaine. Un *million* d'années par seconde. Cela signifie, en nous restreignant à la masse de la planète pour la puissance de calcul, qu'il serait possible de soutenir plus de gens que l'*Univers* entier pourrait soutenir si les humains biologiques colonisaient *chaque planète de l'Univers*. Cela signifie que, en une seule journée, une civilisation pourrait vivre plus de *80 milliards d'années*, plusieurs fois l'âge actuel de l'*Univers*.

Ce qui est étrange avec la plupart des gens qui mentionnent les « lois de la physique » comme fixant des limites absolues aux Puissances ne rêveraient pas de fixer les mêmes limites à une civilisation d'humains (normaux) qui aurait colonisé (ne serait-ce que) la galaxie et serait vieille d'un (petit) milliard d'années. En partie, c'est simplement le produit d'une convention culturelle de la fiction scientifique ; les civilisations interstellaires peuvent violer toutes les lois de la physique qu'elles veulent, parce que les lecteurs y sont habitués. Mais c'est aussi parce que les scientifiques et les auteurs de fiction scientifique ont appris, encore et encore, que les Limites Infranchissables Ultimes cèdent généralement à l'ingéniosité humaine et au passage de quelques générations. Personne n'ose dire ce qui pourrait être possible dans un *milliard d'années* parce que c'est simplement une durée inimaginable.

Nous savons que le changement rampait à une allure d'escargot il y a seulement un millénaire, et que même il y a cent ans il aurait été *impossible* de fixer des limites correctes au pouvoir ultime de la technologie. Nous savons que le passé n'a jamais été en mesure de fixer des limites au présent, et c'est pourquoi nous n'essayons pas d'en fixer à l'avenir. Mais pour les transhumains, l'analogie n'est pas avec Lord Kelvin, ni avec Aristote, ni avec un chasseur-cueilleur (qui avaient tous une intelligence humaine), mais avec un homme de Neandertal. Pour les Puissances, avec un poisson. Et pourtant, parce que la puissance d'une intelligence plus élevée n'est pas aussi largement reconnue que la puissance de quelques millions d'années — parce que nous n'avons pas d'*histoire* de pessimistes démentis par des transhumains plutôt que simplement par le passage du temps —, certains de nous sont toujours assis, grognant autour du feu, fixant des limites ultimes au tranchant des pointes de flèches, certains de nous continuent à nager, impassibles, incapables de conduire un raisonnement abstrait, mais sachant que l'Univers entier est, *doit* être, mouillé.

Pour donner une idée du rythme du progrès conduit par des chercheurs *plus intelligents*, il a fallu que j'invente une fonction plus complexe que la fonction de doublement utilisée ci-dessus. On appellera cette nouvelle fonction T(n). Vous pouvez vous représenter T(n) comme le plus grand nombre concevable par un être ayant un cerveau de n neurones. Plus formellement, T(n) est défini comme le plus long bloc de 1 produit par une machine de Turing (qui s'arrête) ayant n états et agissant sur une bande au départ vide. Si les ordinateurs vous sont plus familiers que les machines de Turing, considérez T(n) comment étant le plus grand nombre pouvant être produit par un programme d'ordinateur contenant n instructions. Ou si vous êtes versé dans la théorie de l'information, envisagez T(n) comme étant la fonction inverse de la complexité ; elle produit le plus grand nombre ayant une complexité n ou moins.

La séquence produite par l'itération de T(n), Sn = T(Sn - 1), est constante pour des valeurs très basses de n. S0 est défini comme étant égal à 0 ; un programme de longueur zéro ne produit pas de sortie. La correspondance est avec un Univers vide d'intelligence. T(1) = 1. Ceci correspond à une intelligence qui n'est pas capable de s'améliorer elle-même ; ceci correspond à notre position actuelle. T(2) = 3. Ici commence le saut dans l'abîme. Une fois que cette fonction commence à augmenter, elle saute immédiatement au-delà des limites du connaissable. T(3) = 6? T(6) = 64?

T(64) = beaucoup plus que 10^{80}, le nombre d'atomes dans l'Univers. T(10^{80}) est quelque chose que seule une entité transcendante pourra jamais calculer, et cela seulement si des entités transcendantes peuvent créer de nouveaux Univers, peut-être même de nouvelles lois de la physique, pour fournir la puissance de calcul nécessaire. Même T(64) ne sera probablement jamais connu par un être strictement humain.

Prenez maintenant la version transcendée de Sn, en commençant à 2. Une demi-unité de temps plus tard, nous avons 3. Un tiers d'unité de temps plus tard, nous avons 6. Un sixième plus tard — une unité entière depuis que cette fonction a démarré — nous avons 64. Un soixante-quatrième plus tard, 10^80. Une fraction de seconde inimaginablement minuscule plus tard... Singularité.

Est-ce que Sn est vraiment un bon modèle pour la Singularité ? Bien sûr que non. « Bon modèle pour la Singularité » est un oxymore, c'est bien de *cela* qu'il s'agit ; la Singularité excédera tout modèle qu'un humain pourrait avoir formulé il y a cent ans, et la Singularité excédera tout modèle que nous formulons aujourd'hui [6].

L'objection principale, cependant, serait que Sn est une métaphore sans fondement. La séquence de doublement transcendée modèle des chercheurs plus rapides. Il est facile de dire que Sn modèle des chercheurs plus intelligents, mais qu'est-ce que signifie vraiment *intelligent* dans ce contexte ?

2.1. La définition de l'intelligence

L'intelligence est la mesure de ce qui vous apparaît comme *évident*, de ce dont vous pouvez reconnaître *l'évidence après coup*, de ce que vous pouvez *inventer*, et de ce que vous pouvez *comprendre*. Pour être plus précis à ce propos, l'intelligence est la mesure de vos primitives sémantiques (ce qui est simple après coup), de la façon dont vous manipulez les primitives sémantiques (ce

qui est évident), des structures que vos primitives sémantiques peuvent former (ce que vous pouvez comprendre), et de la façon dont vous pouvez manipuler ces structures (ce que vous pouvez inventer). Si les termes de la théorie de la complexité vous sont familiers, la différence entre *évident* et *évident après coup*, ou entre *inventable* et *compréhensible*, est comme la différence entre NP et P.

Tous les humains qui n'ont pas subi de lésions nerveuses ont les mêmes primitives sémantiques. Ce qui est *évident après coup* à l'un est *évident après coup* à tous. (Quatre notes : D'abord, en parlant de « lésions nerveuses », je n'entends rien de désobligeant ; c'est juste qu'une personne dépourvue de cortex visuel n'aura pas les primitives sémantiques visuelles. Si certaines voies nerveuses sont interrompues, les êtres qui en sont affectés ne perdront pas seulement leur capacité de *voir* les couleurs ; ils perdront aussi leur capacité de *se rappeler* ou *d'imaginer* des couleurs. Deuxièmement, des théorèmes mathématiques peuvent ne sembler évidents après coup qu'à des mathématiciens — mais n'importe quelle personne qui acquerrait la *compétence* aurait la *capacité* de le voir. Troisièmement, ce que nous considérons comme *évident* implique dans une certaine mesure, non seulement des primitives symboliques, mais aussi des liens relativement courts entre elles. Je compte les types de liens primitifs comme faisant partie des « primitives sémantiques ». Quand nous regardons une séquence de pensées et qu'elle nous semble *évidente après coup*, il ne s'agit pas nécessairement d'une primitive sémantique unique, mais elle est composée d'une chaîne très courte de primitives sémantiques et de types de liens. Quatrièmement, je m'excuse auprès de mes lecteurs de disséquer ainsi mes propres métaphores ; c'est plus fort que moi.)

De façon similaire, l'architecture cognitive humaine est universelle. Notre esprit à tous est fait du même matériau sous-jacent. Bien que la nature de ce matériau dont l'esprit est constitué ne soit pas nécessairement connue, notre capacité à communiquer les uns avec les autres indique que, quoi que nous communiquions, il s'agit de la même chose des deux côtés. Si deux humains quelconques partagent un ensemble de concepts, toute structure composée de ces concepts qui est comprise par l'un sera comprise par l'autre.

Des humains différents peuvent différer dans leur capacité de *manipuler* et de *structurer* les concepts ; des humains différents peuvent voir et inventer des choses différentes. Les grandes percées de la physique et de l'ingénierie ne se sont pas produites parce qu'un groupe de gens ont persévéré encore et encore pendant des générations jusqu'à ce qu'ils trouvent une explication si complexe, une chaîne d'idées si longue que seul le temps pourrait l'inventer. La relativité, la physique quantique, le footballène et la programmation orientée-objet se sont tous produits parce que quelqu'un a assemblé une structure sémantique courte, simple, élégante, d'une façon à laquelle personne n'avait pensé avant. Être *un petit peu plus intelligent* est l'origine des révolutions. Pas le temps. Pas le travail soutenu. Même si le travail soutenu et le temps ont généralement été nécessaires, d'autres avaient travaillé beaucoup plus dur et plus longtemps sans résultat. L'essence de la révolution est l'intelligence brute.

Pensez maintenant à la Singularité. Pensez au chimpanzé qui essaie de comprendre le calcul intégral. Pensez aux gens avec une neurologie visuelle endommagée qui n'arrivent pas à se rappeler comment c'était de voir, qui ne peuvent pas imaginer la couleur rouge ou visualiser des structures bidimensionnelles. Pensez à un cortex visuel avec un billion (10^12) de fois autant d'équivalents de neurones. Pensez à vingt mille couleurs distinctes dans l'arc-en-ciel, sans qu'aucune ne soit une nuance d'une autre. Pensez à un objet à cinquante dimensions que vous faites pivoter mentalement. Pensez à l'association de primitives sémantiques aux pixels, de façon qu'on puisse voir un arc-en-ciel d'idées comme on voit un arc-en-ciel de couleurs.

Nos primitives sémantiques déterminent même ce que nous pouvons *connaître*. Pourquoi existe-t-il quelque chose ? Personne ne sait. Et pourtant la réponse doit être évidente. La Cause Première *doit* être évidente. Elle doit être évidente à *Rien*, présent en l'absence de quoi que ce soit d'autre, une substance formée de *-vide-*, une conclusion déduite sans données ou hypothèses initiales. Qu'est-ce qui suscite l'expérience consciente, ce dont les esprits sont faits ? Nous sommes *constitués* d'expériences conscientes. Il n'y a *rien* dont nous ayons une expérience plus directe. Comment est-ce que ça fonctionne ? On n'en a pas la moindre idée. *Deux millénaires et demi* de tentatives pour répondre et tout ce qu'on a à montrer est : « Je pense donc je suis ». Les solutions semblent être nécessairement simples, et pourtant on peut démontrer qu'elles sont insaisissables. Peut-être que les solutions se situent en dehors des représentations qui peuvent être formées avec un cerveau humain.

Si c'est le cas, alors nos descendants, successeurs, ou ceux que nous deviendrons trouveront les primitives sémantiques nécessaires et se modifieront eux-mêmes pour les percevoir. Les Puissances disséqueront l'Univers et la Réalité jusqu'à ce qu'elles comprennent pourquoi quelque chose existe, elles analyseront les neurones jusqu'à ce qu'elles comprennent les *qualia*. Et ça ne sera que le *début*. Ca ne s'arrêtera pas là. Pourquoi ne devrait-il y avoir que deux problèmes difficiles ? Après tout, s'il n'y avait pas d'humains, l'Univers ne contiendrait apparemment qu'un seul problème difficile : comment un penseur non conscient pourrait-il formuler le problème difficile de la conscience ? Pourrait-il exister des états d'existence au-delà de la simple conscience, une transconscience ? La résolution de la nature de la réalité pourrait-elle créer la capacité de créer de nouveaux univers, de manipuler les lois de la physique, et même de modifier *le genre de choses qui peuvent être réelles*, une « ontotechnologie » ? C'est de *ça* qu'il s'agit avec la Singularité.

Donc, avant que vous parliez de la vie en tant que Puissance ou de l'Utopie à venir — un des passe-temps favoris des transhumanistes et des Extropiens est de discuter les problèmes de l'upload, la vie après l'upload, etc. —, rappelez-vous simplement que vous avez de bien meilleures chances de résoudre les deux problèmes difficiles que de tenir des propos sensés au sujet de l'avenir. Ca vaut pour moi aussi. Je suis prêt à soutenir tout ce que

j'ai dit sur les humains, y compris notre incapacité à comprendre certaines choses, mais tout ce que j'ai dit sur les Puissances est presque certainement faux. « Ils trouveront les primitives sémantiques nécessaires et se modifieront eux-mêmes pour les percevoir. » Faux. « Trouver. » « Primitives sémantiques. » « Modifier. » « Percevoir. » Je parierais que tous ces termes deviendront obsolètes après la Singularité. Il y a des meilleures méthodes et je suis sûr qu'Ils, ou que CA, ou que [son d'un cerveau qui explose] les « trouveront ».

2.2. Transcendances perceptives

J'aimerais introduire une unité de progrès post-Singularité, la Transcendance perceptive, ou TP.

[Courte pause tandis que le public s'effondre mort de rire.]

Une Transcendance perceptive se produit lorsque toutes les choses qui étaient *compréhensibles* deviennent *évidentes après coup*, et que toutes les choses qui étaient *inventables* deviennent *évidentes*. Une transcendance perceptive se produit lorsque les structures sémantiques d'une génération deviennent les primitives sémantiques de la suivante. Autrement dit, à une TP d'aujourd'hui, *la totalité de la connaissance humaine sera perceptible en une expérience instantanée*, de la même façon que nous percevons maintenant une image entière d'un seul coup.

Les ordinateurs ont, *grosso modo*, une TP d'avance par rapport aux humains en ce qui concerne l'arithmétique. Alors que nous devons manipuler toute une pyramide fragile de chiffres, de rangées et de colonnes pour multiplier 62305 par 10358, un ordinateur peut sortir la réponse — 645355190 — en une seule étape *évidente*. En fait, ces ordinateurs n'ont pas vraiment une TP d'avance sur nous, pour deux raisons. D'abord, ils ne manipulent les nombres que jusqu'à deux milliards, après quoi ils doivent manipuler des pyramides eux aussi. Mais surtout, ils ne remarquent rien au sujet des nombres qu'ils manipulent, contrairement à nous. Si vous multipliez 23704 par 14223 « à l'italienne », vous ne multiplierez pas 23704 par 2 deux fois de suite ; vous réutiliserez les résultats de la ligne précédente. Si un des résultats intermédiaires est 12345 ou 99999 ou 314159, vous le remarquerez également. La façon dont les ordinateurs manipulent les nombres est en réalité *moins* puissante que la façon dont nous manipulons les nombres.

Est-ce que les Puissances se satisferaient de moins ? Une TP au-dessus de nous, la multiplication est réalisée *automatiquement* mais avec une pleine attention aux résultats intermédiaires, aux nombres qui se trouvent être des nombres premiers, etc. Si je concevais une des premières Puissances — et, au Singulary Institute, c'est ce que nous faisons —, je créerais un sous-système entier pour la manipulation des nombres, sous-système qui utiliserait la primalité, la complexité et toutes les propriétés numériques que l'humanité connaît. Une Puissance comprendrait *pourquoi* 62305 fois 10358 est égal à 645355190, avec la même compréhension que celle qu'obtiendrait un mathématicien humain de premier rang qui aurait consacré des heures à étudier les nombres impliqués. Et, dans le même temps, la Puissance multiplierait les deux nombres automatiquement.

Pour une telle Puissance, pour laquelle les nombres seraient des vraies primitives sémantiques, le Dernier théorème de Fermat et la Conjecture de Goldbach pourraient être *évidents*. Quelque part au fond de son esprit, la Puissance testerait chaque proposition avec un million d'essais, manipulant de façon non consciente tous les nombres concernés pour voir *pourquoi* ils ne sont pas la somme de deux cubes ou *pourquoi* ils sont la somme de deux nombres premiers ou *pourquoi* leur partie réelle est égale à une demi. À partir de là, la Puissance pourrait *déduire intuitivement* la solution la plus élémentaire simplement par généralisation. Peut-être que des mathématiciens humains, s'ils pouvaient réaliser les opérations arithmétiques nécessaires pour un millier de tests de l'Hypothèse de Riemann, pourraient, en examinant chaque étape intermédiaire et en recherchant des propriétés communes et des raccourcis intéressants, déduire intuitivement une solution formelle. Mais ils ne peuvent pas, et encore moins de façon inconsciente, ce qui explique que l'Hypothèse de Riemann reste non évidente et non démontrée : c'est une *structure* conceptuelle au lieu d'être une *primitive* conceptuelle.

Peut-être qu'un exemple encore plus stimulant est celui fourni par notre cortex visuel. De prime abord, le cortex visuel semble être un processeur d'image. Dans un moteur graphique informatique moderne, une image est représentée par une matrice bidimensionnelle de pixels [7]. Pour faire pivoter cette image — pour prendre un exemple d'opération —, les coordonnées cartésiennes orthogonales de chaque pixel sont converties en coordonnées polaires theta, r. Tous les theta, qui représentent l'angle, se voient ajouter une constante. Les coordonnées polaires sont ensuite converties à nouveau en coordonnées cartésiennes. Il y a des façons d'optimiser ce processus, et des façons de rendre compte de croisements de pixels et de pixels vides sur la nouvelle matrice, mais l'essentiel est clair : Pour réaliser une opération sur l'image entière, réalisez l'opération sur chaque pixel dans cette image.

À ce stade, on pourrait dire qu'une Transcendance perceptive dépend du niveau auquel vous examinez l'opération. Si vous considérez que vous réalisez l'opération pixel par pixel, c'est une structure cognitive inimaginablement laborieuse ; mais si vous voyez toute la chose en un seul morceau, c'est une primitive cognitive — un point souligné dans la *Ant Fugue* d'Hofstadter quand il discute les fourmis et les colonies. Pas très excitant à moins que ça soit Hofstadter qui l'explique, mais le cortex visuel ne se résume pas à cela.

Pour commencer, nous faisons l'expérience consciente de la rougeur. (Si vous n'êtes pas sûr de voir en quoi consiste l'expérience consciente, aussi appelée « *qualia* », la version courte est que vous n'êtes pas celui qui *exprime* vos pensées, vous êtes celui qui *entend* vos pensées.) Les *qualia* sont ce qui constitue la différence indescriptible entre rouge et vert.

Le terme « primitive sémantique » décrit davantage que

simplement le niveau auquel les symboles sont des objectes distincts et compacts. Il décrit le niveau de la perception consciente. Contrairement à l'ordinateur qui manipule des nombres formés de bits, et à l'instar de la Puissance imaginée qui manipule des théorèmes formés de nombres, nous ne perdons pas de résolution en passant du niveau du pixel au niveau de l'image. Nous ne percevons pas soudainement l'idée « il y a un ours en face de moi » ; nous voyons l'image d'un ours, qui contient des millions de pixels, et nous faisons l'expérience consciente de tous simultanément. Une Transcendance perceptive n'est pas « juste » la superposition d'un nouveau niveau cognitif, elle transforme les structures cognitives en primitives dont il est fait une *expérience consciente*.

« Autrement dit, à une TP d'aujourd'hui, *la totalité de la connaissance humaine sera perceptible en une expérience instantanée*, de la même façon que nous percevons maintenant une image entière d'un seul coup. »

Bien sûr, la TP ne sera pas utilisée comme unité de progrès post-Singularité. Même si elle l'était au départ, rapidement « TP » lui-même serait transcendé et les Puissances sauteraient à nouveau hors du système. Après tout, la Singularité est en fin de compte aussi éloignée de moi, l'auteur, qu'elle l'est de n'importe quel autre humain, en sorte que la TP sera une description aussi vaine que la séquence de doublement abandonnée il y a longtemps. Même si nous acceptons la TP comme unité de mesure de base, elle introduit simplement une Singularité secondaire. Peut-être que la Transcendance perceptive se produira tous les deux ans d'expérience consciente au départ, mais ensuite chaque année consciente, puis tous les six mois conscients... Vous voyez l'idée ?

C'est comme la *Birthday Cantatatata...* dans le livre de Hofstadter, Gödel, Escher, Bach. Vous pouvez commencer avec la séquence 1, 2, 3, 4 ... et sauter à partir de là à w (oméga), le symbole de l'infini. Mais on a alors w, w + 1, w + 2 ... , et nous sautons à nouveau à 2w. Puis 3w, et 4w, et w^2 et w^3 et w^w et $w^{(w^w)}$ et des tours plus hautes de w jusqu'à ce que nous sautions à l'ordinal e0 qui contient toutes les tours exponentielles de w.

Il se peut que la TP introduise une deuxième Singularité, et une troisième Singularité, et une quatrième, jusqu'à ce que les Singularités arrivent de plus en plus vite et que la première w-Singularité soit imminente.

Ou peut-être que les Puissances sauteront simplement hors de ce système. La Birthday Cantatatata... a été écrite par un humain — *Douglas Hofstadter*, certes, mais quand même un humain — et les concepts qu'elle implique peuvent être transcendés par le tout premier transhumain.

Les Puissances dépassent notre capacité de compréhension.

D'accord ?

2.3. Gros nombres

Il est difficile d'apprécier la Singularité convenablement sans apprécier auparavant des nombres vraiment grands. Je ne parle pas de minuscules petits nombres, qu'on distingue à peine de zéro, comme le nombre d'atomes dans l'Univers ou le nombre d'années qu'il faudrait à un singe pour dupliquer les œuvres de Shakespeare. Je vous invite à considérer ce qui était, vers 1977, le plus grand nombre jamais utilisé dans une démonstration mathématique sérieuse. La démonstration, par Ronald L. Graham, est une borne supérieure à une certaine question de théorie de Ramsey. Pour expliquer la démonstration, il faut introduire une nouvelle notation, qu'on doit à Donald E. Knuth, dans son article *Coping With Finiteness*. La notation est généralement une petite flèche qui pointe vers le haut, ici un ^. Sous forme de fonction :

```
int arrow (int num, int power, int arrownum)
int answer = num;
if (arrownum == 0)
  return num * power;
for (int i = 1; i < power; i++)
  answer = arrow(num, answer, arrownum - 1);
return answer;
// end arrow
```

2^4 = 2^4 = 16.
3^^4 = 3^(3^(3^3)) = 3^(3^27) = $3^{7'625'597'484'987}$
7^^^^3 = 7^^^(7^^^7).

3^3 = 3 * 3 * 3 = 27. Ce nombre est assez petit pour être visualisé.

3^^3 = 3^(3^3) = 3^27 = 7'625'597'484'987. Plus grand que 27, mais si petit que je peux le taper en entier. Personne ne peut visualiser sept billions de quoi que ce soit, mais nous pouvons facilement le comprendre comme étant en gros de l'ordre, disons, du produit national brut [américain].

3^^^3 = 3^^(3^^3) = 3^(3^(3^(3^...^(3^3)...))). Le « ... » a une longueur de 7'625'597'484'987 fois le chiffre 3. En d'autres termes, 3^^^3 ou arrow(3, 3, 3) est une tour exponentielle de trois qui a 7'625'597'484'987 étages. Le nombre dépasse maintenant la capacité humaine de compréhension, mais la procédure pour le produire peut être visualisée. Vous prenez x=1. Puis vous donnez à x la valeur 3^x. Répétez sept billions de fois. Alors que les tout premiers stades du nombre sont bien trop grands pour être contenus dans l'Univers entier, la tour exponentielle, sous sa forme « 3^3^3^3...^3 », est toujours assez petite pour pouvoir être stockée sur un superordinateur moderne.

3^^^^3 = 3^^^(3^^^3) = 3^^(3^^(3^^...^^(3^^3)...)). Tant le nombre que la procédure pour le produire dépassent maintenant la capacité de visualisation humaine, bien que la procédure puisse être comprise. Prenez x=1. Puis donnez à x la valeur d'une tour de trois de hauteur x. Répétez 3^^^3 fois, où 3^^^3 est égal à une tour exponentielle d'une hauteur de sept billions de trois.

Et pourtant, comme le dit Martin Gardner : « 3^^^^3 est inimaginablement plus grand que 3^^^3, mais il reste petit pour un nombre fini, puisque la plupart des nombres finis sont beaucoup plus grands. »

Et maintenant, le nombre de Graham. Prenez x = 3^^^^3, ou le nombre inimaginable décrit ci-dessus. Puis donnez à x la valeur 3^^^^^^^(x flèches)^^^^^^^3. Répétez 63 fois, ou 64 en comptant le 3^^^^3 original.

Le nombre de Graham dépasse de beaucoup ma capacité d'appréhension. Je peux le décrire, mais je ne peux pas convenablement l'appréhender. Peut-être que Graham peut l'appréhender, puisqu'il a écrit une démonstration mathématique qui l'utilise. Ce nombre est beaucoup plus grand que la conception qu'on la plupart des gens de *l'infini*. Je sais qu'il était plus grand que la mienne. Ce nombre a produit sur moi, la première fois que je l'ai rencontré, une impression indescriptible. C'était l'impression de regarder quelque chose de tellement plus *grand* que le monde à l'intérieur de ma tête que ma conception de l'Univers a été renversée et s'est reconstruite pour s'ajuster. Tous les théologiens devraient être confrontés à nombre pareil, pour qu'ils puissent convenablement appréhender ce qu'ils invoquent en parlant de l'intelligence « infinie » de Dieu.

Mon bonheur a été complet quand j'ai appris que la *réponse* au problème de Ramsey qui donne lieu à ce nombre — plutôt que la *borne supérieure* — était probablement six.

Pourquoi tout cela était-il nécessaire, toute esthétique mathématique mise à part ? Parce que tant que vous n'avez pas compris combien sont creux les mots « infini », « grand » et « transhumain », vous ne pouvez pas appréhender la Singularité. Appréhender la Singularité est aussi éloigné de nous que visualiser le nombre de Graham l'est d'un chimpanzé. Et même plus éloigné de nous que cela. Aucune analogie humaine ne sera jamais capable de décrire la Singularité, parce que nous sommes seulement humains.

Le nombre décrit plus haut a été créé par un esprit humain. Ca n'est qu'un entier positif fini, même si c'en est un grand. Il est composé et impair, plutôt que premier ou pair ; il est parfaitement divisible par trois. Dans les chiffres décimaux de ce nombre peuvent être encodées, avec pratiquement n'importe quelle méthode d'encodage, toutes les œuvres jamais écrites par une main humaine, et toutes celles qui auraient pu être écrites, à raison de cent mille mots par minute, pendant toute la durée de l'Univers, mis à sa propre puissance mille fois. Et pourtant, si on additionne tous les chiffres de base 10, le résultat sera divisible par neuf. Le nombre reste un entier positif fini. Il peut contenir des univers inimaginablement plus grands que celui-ci, ça n'est toujours qu'un nombre. C'est un nombre si petit que l'algorithme pour le produire peut tenir dans un seul esprit humain.

La Singularité est au-delà de ça. Nous ne pouvons pas la réduire en disant que ça sera un entier positif fini. Nous ne pouvons rien dire du tout à son sujet, si ce n'est qu'elle dépassera notre compréhension.

Si vous pensez que la notation par flèche de Knuth produit des nombres assez grands, qu'en est-il de T(n) ? De combien d'états une machine de Turing a besoin pour implémenter le calcul ci-dessus ? Quelle est la complexité du nombre de Graham, C(Graham) ? Probablement de l'ordre de 100. Et en outre, T(C(Graham)) est probablement beaucoup, beaucoup plus grand que le nombre de Graham. Pourquoi ne répéter x = 3^(x ^s)^3 que 64 fois ? Pourquoi pas 3^^^^3 fois ? Ca serait probablement plus facile, puisque nous devons déjà générer 3^^^^3, mais pas 64. Et avec l'espace qui reste, nous pourrions certainement introduire un algorithme encore plus complexe du point de vue computationnel. De fait, la notation par flèche de Knuth n'est probablement pas l'algorithme le plus puissant qui tient dans C(Knuth) états.

T(n) est la métaphore du taux de croissance d'une entité qui s'auto-améliore, parce qu'elle contient l'idée d'un accroissement d'intelligence mise à contribution pour s'améliorer soi-même. Je ne sais pas quand T(n) dépasse le seuil de ce que les mathématiciens humains peuvent, en théorie, calculer. Probablement plus que n=10 et moins que n=100. Le fait est que après quelques itérations, on se retrouve avec T(4294967296). Maintenant, je ne sais pas à quoi T(4294967296) peut être égal, mais la machine de Turing gagnante générera probablement une Puissance dont la fonction est de penser à un nombre vraiment grand. *Voilà* ce que le terme « grand » veut dire.

2.4. Plus intelligent que nous

C'est très bien de parler de primitives cognitives et d'évidence, mais encore une fois : que veut dire *plus intelligent* ? La signification d'*intelligent* ne peut pas être fondée dans la Singularité : je n'y suis pas encore allé. Alors quelle est ma définition *pratique* ?

> Le plus grand défi pour un écrivain est un personnage plus intelligent que l'auteur. Ca n'est pas impossible. Des problèmes que l'auteur prend des mois à résoudre, ou à concevoir, le personnage peut les résoudre en un instant. Mais Dieu vienne en aide à l'écrivain si son personnage anormalement intelligent *se trompe* !
>
> Larry Niven
>
> Bien sûr, je n'ai jamais écrit l'histoire « importante », la suite à propos du premier humain amplifié. J'ai essayé une fois quelque chose de ce genre. La lettre de refus de John Campbell commençait ainsi : « Désolé, vous ne pouvez pas écrire cette histoire. Ni personne d'autre. »
>
> Vernor Vinge

L'intelligence est cette qualité qui fait que vous ne pouvez pas écrire l'histoire d'un personnage plus intelligent que vous. Vous pouvez évoquer des penseurs superrapides, des souvenirs eidétiques, des calculateurs éclair ; un personnage qui apprend douze langues en une semaine, qui peut lire un manuel en une heure, ou qui peut inventer toute sorte de choses merveilleuses (tant que vous ne devez pas produire l'invention elle-même). Mais vous ne pouvez pas décrire un personnage avec une maturité émotionnelle plus grande, un personnage qui peut

identifier la solution évidente que vous avez ratée, un personnage qui sait (et peut dire au lecteur) le Sens de la Vie, un personnage doté d'une conscience de soi surhumaine. Pas à moins que vous puissiez faire ces choses-là vous-même.

Prenons un exemple concret, l'histoire *Des fleurs pour Algernon* (suivi du film *Charly*), de Daniel Keyes. (Je crains de devoir vous dire la fin de l'histoire, mais c'est une histoire qui vaut par ses personnages, pas par son concept, donc ça ne devrait pas la déflorer.) *Des fleurs pour Algernon* tourne autour d'une procédure neurochirurgicale d'augmentation de l'intelligence. Cette procédure a été d'abord testée sur une souris, Algernon, puis sur un humain attardé mental, Charlie Gordon. Le Charlie amélioré a l'ensemble des traits surhumains habituels dans la fiction scientifique ; il pense vite, accumule une vie entière de connaissances en l'espace de quelques semaines, et discute de sujets mathématiques très pointus (pas montré). Puis la souris, Algernon, tombe malade et meurt. Charlie analyse la procédure d'amélioration (pas montré) et conclut que le processus contient une faille fondamentale. Plus tard, Charlie meurt.

Voilà un humain amélioré de fiction scientifique. Un humain amélioré réel n'aurait pas été pris par surprise. Un humain amélioré réel réaliserait que toute amélioration simple de l'intelligence sera un désavantage évolutionnaire net : si l'amélioration de l'intelligence était simplement l'affaire d'une procédure chirurgicale, elle serait apparue depuis longtemps comme mutation naturelle. *A fortiori* pour une procédure qui fonctionne sur les rats ! (Pour autant que je sache, Keyes n'a jamais pensé à cela. J'ai choisi *Des fleurs*, parmi toutes les histoires célèbres d'amélioration de l'intelligence, parce que, pour des raisons d'unité dramatique, cette histoire montre ce qui se trouve être l'issue correcte.)

Remarquez que je ne vous ai pas ébloui avec une explication jargonneuse de la mort de Charlie ; mon explication fait deux phrases et peut être comprise par quelqu'un qui n'est pas un expert dans le domaine. C'est la *simplicité* de l'intelligence qui est tellement impossible à exprimer dans la fiction, et qui est tellement choquante quand on la rencontre en personne. Tout ce que la fiction scientifique peut faire pour montrer l'intelligence, c'est du jargon et des gadgets. Un Charlie Gordon réellement surintelligent n'aurait pas été pris par surprise ; il aurait déduit son destin probable en utilisant le raisonnement très simple ci-dessus. Il aurait accepté cette probabilité, réorganisé ses priorités et agi en conséquence jusqu'à ce qu'il ait épuisé son temps ; ou, plus probablement, il aurait imaginé une façon également simple et évidente-après-coup d'éviter son sort. Si Charlie Gordon avait été *réellement* surintelligent, il n'y aurait pas eu d'histoire.

Il y a des différences si grandes qu'elles rendent tous les problèmes neufs. Imaginez un domaine quelconque dans lequel vous êtes un expert — la neuroscience, la programmation, la plomberie, n'importe quoi — et considérez la différence entre un néophyte, qui vient d'aborder un problème pour la première fois, et un expert. Même si un millier de néophytes essaient de résoudre un problème et échouent, ça ne signifie pas qu'un seul expert ne pourrait pas résoudre le problème avec désinvolture. Si cent physiciens instruits essaient de résoudre un problème et échouent, un Einstein pourrait cependant être capable de réussir. Si un millier d'enfants de 12 ans essaient pendant une année de résoudre un problème, ça ne donne aucune indication pour déterminer si un adulte risque de pouvoir résoudre le problème. Si un million de chasseurs-cueilleurs essaient de résoudre un problème pendant un siècle, la réponse pourrait être évidente pour n'importe quel humain instruit du vingt-et-unième siècle. Et aucun nombre de chimpanzés, quel que soit le temps qu'ils y consacrent, ne fourniraient aucune indication permettant de déterminer si le plus stupide des humains pourrait résoudre un problème sans même y réfléchir. Il y a des différences si grandes qu'elles rendent tous les problèmes neufs ; et certaines, comme la différence entre un néophyte et un expert, ou celle entre un chasseur-cueilleur et un citoyen instruit, ne sont même pas des différences de matériel : elles concernent non pas la magie de l'*intelligence*, mais la magie de la *connaissance*, ou le *manque de stupidité*.

Je songe au temps où je n'avais pas encore commencé à étudier la psychologie évolutionnaire et la science cognitive. Je sais que je n'aurais pas pu alors prévoir le cours de la Singularité. « Si je ne pouvais pas le faire *alors*, qu'est-ce qui me fait penser que je peux le faire *maintenant* ? » Je suis un humain, et un citoyen instruit, et un adulte, et un expert, et un génie… mais s'il y a encore une différence de magnitude similaire qui reste entre moi et la Singularité, alors mes spéculations ne vaudront pas mieux que celles d'un savant du dix-huitième siècle.

Nous sommes tous familiarisés avec les variations individuelles de l'intelligence humaine, réparties selon la fameuse courbe de Gauss ; c'est la seule référence que la plupart d'entre nous ont pour « plus intelligent ». Mais précisément *parce que* ces variations tombent à l'intérieur des variations de conception du cerveau humain, elles n'ont rien d'extraordinaire. Une des vérités très profondes sur l'esprit humain est que l'évolution nous a conçus pour être stupides, aveuglés par l'idéologie, pour refuser d'admettre que nous avons tort, pour penser que « l'ennemi » est inhumain, pour être affectés par la pression sociale. Les variations dans l'intelligence qui se trouvent dans les limites normales de conception n'affectent pas directement cette stupidité. C'est de là que vient la croyance populaire selon laquelle l'intelligence n'implique pas la sagesse, et dans les limites humaines c'est essentiellement juste [8]. Les variations que nous voyons ne *frappent* pas assez fort pour que les gens apprécient ce que « plus intelligent » signifie.

Je suis un Singularitarien parce que j'ai une petite idée de combien il est tout à fait, définitivement, absolument *impossible* de penser comme quelqu'un qui est ne serait-ce qu'un tout petit plus intelligent que vous. Je sais que nous *passons tous à côté de l'évidence*, chaque jour. Il n'y a pas de problèmes difficiles, seulement des problèmes qui sont difficiles pour un certain niveau d'intelligence. Passez à un niveau un tant soit peu supérieur, et certains problèmes passeront soudainement de « insolubles » à « évidents ». Passez à un niveau nettement supérieur, et ils deviendront tous évidents. Passez à un niveau immensément

supérieur...

Et je sais que ma représentation de la Singularité sera *encore* au-dessous de la réalité. Il se peut que je ne sois pas modeste, mais j'ai mon humilité : si je suis capable d'identifier des anthropomorphismes et des failles logiques béantes dans chaque transhumain de chaque texte de fiction scientifique, il s'ensuit qu'un génie d'un ordre légèrement supérieur (pour ne pas parler d'un transhumain !) pourrait lire cette page et rire devant mon manque d'imagination. Appelez ça l'expérience, l'humilité, la conscience de soi ou le Principe de médiocrité ; j'ai franchi assez de paliers pour croire qu'il y en a davantage. Je sais, confusément, combien je suis stupide.

J'ai essayé de montrer combien la Singularité nous dépasse par force brute, mais il n'y a pas *besoin* de vitesses infinies, de TP et de w pour que quelque chose nous dépasse totalement. Il suffit d'une *toute petite* avance, d'un *petit peu* plus d'intelligence, et le Beyond nous regarde droit dans les yeux une fois de plus. *Je* n'ai jamais traversé la Singularité. *Je* ne suis jamais allé dans la Transcendance. J'ai juste délimité une zone du Low Beyond. La vocation de cette page est de communiquer une impression de grandeur qui vient de *l'expérience personnelle*, et est pour cette raison *seulement humaine*.

De mon cortex, au vôtre ; chaque concept présenté ici a été engendré par un simple vieil *Homo sapiens*, et toute impression qu'il a faite sur vous a été de la même façon engendrée par un simple vieil *Homo sapiens*. Quelqu'un qui y a consacré un peu plus de réflexion, ou quelqu'un qui est un peu plus extrême ; ça ne fait pas de différence. Quelle que soit l'impression que vous avez tirée de cette page, ça n'est pas une image précise du futur éloigné ; c'est, inévitablement, une image de *moi*. Et je *ne suis pas* le futur éloigné. Seule une version de ce texte écrite par une Puissance pourrait fournir l'expérience de la Singularité.

Prenez le « choc du futur », quel qu'il soit, évoqué par cette page, et associez-le, non à la Singularité, mais à moi, le garçon doux et calme qui diffère de façon infime du reste de l'humanité. Ne cherchez pas à extrapoler au-delà de ça. Vous ne pouvez pas. Personne ne peut : ni vous, ni moi.

2035. Probablement avant.

3. Plus tôt que vous ne pensez

Depuis que l'Internet a explosé à travers la planète, il y a eu assez de puissance de calcul en réseau pour de l'intelligence. Si l'Internet était convenablement reprogrammé, il suffirait à faire tourner un cerveau humain, ou une IA germe. Du côté de la nanotechnologie, nous possédons des machines capables de produire des séquences d'ADN arbitraires, et nous savons comment transformer des séquences d'ADN arbitraires en protéines arbitraires [9]. Nous avons des machines — les sondes à forces atomiques — qui peuvent placer des atomes individuels n'importe où, et dont on a récemment (1999) démontré qu'ils sont capables de former des liaisons atomiques. Le positionnement à une précision d'un centième de nanomètre, des brucelles à l'échelle atomique... les nouvelles continuent à s'accumuler.

Si nous avions une machine à voyager dans le temps, 100 Kilo-octets d'informations en provenance de l'avenir pourrait spécifier une protéine qui construirait un dispositif qui nous donnerait la nanotechnologie du jour au lendemain. 100 Ko pourrait contenir le code de l'IA germe. Depuis les années 90, la Singularité n'est *plus qu'un problème de logiciel*. Et le logiciel est de l'information, cette chose magique qui change à des vitesses de grandeur arbitraire. Du point de vue technologique, la Singularité pourrait se produire *demain*. Une avancée décisive, juste une idée majeure, en ingénierie des protéines, dans la manipulation atomique ou l'intelligence artificielle, une journée exceptionnelle chez Webmind ou Zyvex, et la porte qui donne sur la Singularité s'ouvre grand.

Drexler a écrit un livre de marches à suivre, technique, détaillé, pour la nanotechnologie. Après une stagnation de trente ans, l'IA fait un retour. Les ordinateurs gagnent en puissance encore *plus vite* que leur taux ordinaire et tranquille de doublement tous les deux ans. Quate a construit une sonde à effet tunnel parallèle à 16 têtes. [Écrit en 1996.] Je commence à mettre au point des méthodes pour programmer une IA transhumaine. [Écrit en 1998.] La première liaison chimique a été formée en utilisant un microscope à forces atomiques. Le gouvernement américain a annoncé son intention de consacrer des centaines de millions de dollars à la recherche en nanotechnologie. IBM a annoncé le projet Blue Gene, pour atteindre une puissance de calcul qui se chiffre en petaflops [10] d'ici 2005, avec l'intention de résoudre le problème du pliage des protéines. Le Singularity Institute for Artificial Intelligence, Inc. a été constitué en tant qu'organisation sans but lucratif avec pour but formellement exprimé la programmation d'une IA germe. [Écrit en 2000.]

Le moment exact de la Singularité est ordinairement prédit en prenant une tendance et en l'extrapolant, comme *The Population Bomb* a prédit que nous épuiserions la nourriture en 1977. Par exemple, la croissance démographique est hyperbolique. (Vous avez peut-être appris qu'elle était exponentielle en cours de math, mais elle correspond bien mieux à une courbe hyperbolique qu'exponentielle.) Si cette tendance se poursuit, la population mondiale atteint l'infini le 17 août 2027, plus ou moins 1,8 ans. Il est, bien sûr, impossible que la population humaine atteigne l'infini. Certains disent que si nous créons des IA, alors le graphique pourrait mesurer la population *consciente* plutôt que la population *humaine*. Mais ils s'égarent. Personne n'a conçu la courbe de la population pour prendre en compte les développement en IA. C'est juste une *courbe*, une série de *nombres*. Elle ne peut pas déformer le cours à venir de la technologie simplement pour ne pas dévier de son chemin.

Si vous projetez sur un graphique la taille minimale des matériaux que nous pouvons manipuler, elle atteint le niveau atomique — la nanotechnologie — dans je ne sais plus combien d'années (la page a disparu), mais je crois en 2035 environ. Ceci, bien sûr, était avant le temps des microscopes à effet tunnel et de « IBM » écrit avec des atomes de xénon. Dans ce domaine, nous avons mainte-

nant l'atome artificiel (« Vous pouvez faire n'importe quel type d'atome artificiel : des atomes longs et minces et des atomes grands et ronds. »), qui en un sens a rendu obsolète la nanotechnologie purement *moléculaire*. En 1995, Drexler donnait l'estimation de 2015 [11]. Je soupçonne que ce calendrier a été accéléré un peu depuis. Ma propre estimation serait pas plus tard que 2010.

D'une façon similaire, la puissance de calcul double tous les deux ans dix-huit mois. Si nous extrapolons sur quarante trente quinze ans, nous voyons arriver des ordinateurs avec autant de *puissance brute* (10^17 opérations par seconde) qu'en ont les humains de l'avis de *certaines* personnes en 2035 2025 2015. [La phrase précédente a été écrite en 1996, révisée plus tard la même année, puis révisée à nouveau en 2000 ; d'où les nombres un peu particuliers.] Est-ce que cela signifie que nous avons le logiciel pour faire tourner des esprits ? Non. Est-ce que cela signifie que nous pouvons programmer des personnes plus intelligentes ? Non. Est-ce que cela prend en compte des percées se produisant entre aujourd'hui et cette date ? Non. Est-ce que cela prend en compte les lois de la physique ? Non. S'agit-il d'un modèle détaillé de tous les chercheurs sur la planète ? Non.

C'est juste un graphique. « L'incroyable constance » de la Loi de Moore justifie qu'on la prenne en considération comme une métaphore stimulante de l'avenir, mais pas plus. La séquence de doublement transcendée ne rend pas compte de ce que les chercheurs utilisant des ordinateurs plus rapides peuvent mettre en place la technologie de fabrication physique pour la génération suivante en quelques picosecondes, ni de ce qu'ils peuvent dépasser les lois de la physique. Ca ne signifie pas que de telles choses soient impossibles : il ne me semble en fait pas spécialement probable que la physique actuelle ait atteint le niveau de détail ultime. Peut-être qu'*il n'y a pas* de limites physiques. Le point important est que la Loi de Moore n'*explique* pas comment la physique peut être contournée.

Les mathématiques ne peuvent pas prévoir quand la Singularité arrive. (Enfin, elles peuvent, mais le résultat ne sera pas correct.) Même les progressions remarquablement constantes, telles que le doublement de la puissance de calcul, (A) décrivent des esprits humains non assistés et (B) s'accélèrent, peut-être à cause des programmes de conception assistée par ordinateur. On peut utiliser des statistiques pour prédire l'avenir, mais elles ne le *modèlent* pas. Ce que j'essaie de dire ici, c'est que « 2035 » est juste une estimation hasardeuse, et que ça pourrait aussi bien être mardi prochain.

En vérité, je ne réfléchis pas dans ces termes. Je ne « projette » pas quand la Singularité va se produire. J'ai une « date cible ». J'aimerais que la Singularité se produise en 2005, et je pense que j'aurais une chance raisonnable d'y parvenir par l'IA si quelqu'un me donnait cent millions de dollars par année. Le Singulary Institute aimerait terminer autour de 2008.

Mais surtout, *j'aimerais vraiment, vraiment* que la Singularité arrive *avant* la nanotechnologie, étant donné la quasi-certitude d'abus délibéré, abus d'une ultratechnologie purement matérielle (et pour cette raison, immorale), assez puissante pour détruire la planète. Nous *ne pouvons pas* simplement nous relaxer et attendre. Pour citer Michael Butler : « Attendre le bus est une mauvaise idée s'il s'avère que vous êtes le conducteur du bus. »

Le plus qu'on puisse dire sur 2035 est que cela semble une borne supérieure raisonnable, *étant donné le rythme actuel de progrès.* La borne inférieure ? Trente secondes. Il se peut que nous n'ayons pas connaissance de toute la recherche en train de se faire, après tout.

4. L'upload

Peut-être que vous ne *voulez* pas voir l'humanité remplacée par des « machines », ou des « mutants », même surintelligents ? Vous aimez l'humanité et vous ne voulez pas qu'elle soit rendue obsolète ? Vous craignez de déranger le cours naturel de l'existence ?

Eh bien, pas de chance. La Singularité *est* le cours naturel de l'existence. Chaque espèce — du moins, chaque espèce qui ne s'autodétruit pas avant — fait face tôt ou tard à une surintelligence pleinement développée [12]. Ca arrive à tout le monde. Ca nous arrivera. Ca arrivera même à la première génération de transhumains ou d'IA de niveau humain.

Mais simplement parce que les humains deviennent obsolètes ne signifie pas que *vous* devenez obsolète. Vous n'êtes pas un humain. Vous êtes une intelligence qui, à présent, se trouve avoir un esprit qui est malheureusement limité au matériel humain [13]. Ca pourrait changer. Avec un peu de chance, toutes les personnes sur cette planète qui vivent jusqu'à 2035 ou 2005 ou qui sait quand — et peut-être même certaines autres — deviendront des Puissances.

Le transfert d'un esprit humain dans un système informatique est appelé « upload » ; transformer un mortel en Puissance est appelé « extension ». L'upload archétypal est le tranfert de Moravec, proposé par le Dr Hans Moravec dans son livre Mind Children [14].

> **Remarque :**
> L'hypothèse clé du Transfert de Moravec est que nous pouvons simuler parfaitement un neurone unique. Penrose et Hameroff rejettent cette hypothèse. (En 1999, un neurone de homard a été remplacé avec succès avec 7,5 dollars de matériel acheté chez Radio Shack ; c'est un indice mineur, mais c'est loin d'être conclusif.) La discussion qui suit suppose, soit (A) que les lois de la physique sont computationnelles, soit (B) que nous pouvons construire un « surneurone » qui va au-delà d'une machine de Turing et qui fait la même chose que ce que fait un neurone. (Penrose et Hameroff n'ont pas d'objection à cette dernière proposition. Si un neurone peut tirer profit de la physique profonde pour réaliser des opérations non calculables, nous pouvons faire la même chose technologiquement.)
> Le scénario indiqué suppose ausi une nanomédecine sophistiquée, c'est-à-dire des

nanomachines capables d'exécuter des instructions complexes dans un environnement biologique.

Le Transfert de Moravec déplace (plutôt que copie) progressivement un esprit humain dans un ordinateur. Il n'est à aucun moment nécessaire que vous perdiez conscience. (Les détails qui suivent ont été un peu reconçus et détaillés (par votre serviteur) à partir de l'original dans *Mind Children*.)

1. Un robot de la taille d'un neurone nage jusqu'à un neurone, le scanne et stocke sa représentation en mémoire.

2. Un ordinateur externe, en communication continue avec le robot, commence à simuler le neurone.

3. Le robot attend jusqu'à ce que la simulation de l'ordinateur corresponde parfaitement au neurone.

4. Le robot remplace le neurone par lui-même aussi doucement que possible, envoyant les entrées à l'ordinateur et transmettant les sorties à partir de la stimulation d'un neurone dans l'ordinateur.

Toute cette procédure n'a eu aucun effet sur le flux d'information dans le cerveau, si ce n'est que le traitement d'information d'un neurone particulier est maintenant réalisé dans un ordinateur plutôt que dans le neurone.

5. Répétez l'opération, neurone par neurone, jusqu'à ce que le cerveau entier soit composé de neurones artificiels.

Malgré cela, les synapses (liens) entre les neurones artificiels sont toujours physiques ; les robots rapportent la réception des neurotransmetteurs au niveau des dendrites artificiels et déchargent des neurotransmetteurs à la terminaison des axones artificiels. Dans la phase suivante, nous remplaçons les synapses physiques par des liens logiciels.

6. Pour chaque paire axone-dendrite (transmetteur-récepteur), les entrées ne sont plus rapportées par le robot ; au lieu de cela, la sortie de l'axone calculée du neurone transmetteur est ajoutée en tant que dendrite simulée dans la simulation du neurone récepteur.

À la fin de cette phase, les robots envoient tous des sorties sur leur axone, mais aucun d'eux ne reçoit quoi que ce soit, ils ne s'affectent plus les uns les autres, et aucun n'affecte la simulation informatique.

7. Les robots sont déconnectés.

Vous avez maintenant été placé entièrement dans un ordinateur, bit par bit, sans perte de conscience. Selon la formule de Moravec, votre métamorphose est complète.

Si une des phases semble trop brutale, le transfert d'un neurone individuel ou d'une synapse peuvent être étalés dans le temps selon les nécessités. Pour transférer une synapse lentement dans un ordinateur, nous pouvons utiliser des facteurs pondérés de la synapse physique et de la synapse informatique pour produire la sortie. La pondération serait entièrement physique au début, et entièrement informatique à la fin. Puisque nous supposons que le neurone est parfaitement simulé, la pondération n'affecte que la chaîne causale et non la suite des événements.

Le transfert progressif d'un neurone est un peu plus difficile.

4a. Le robot entoure le neurone, les axones et les dendrites d'une « coquille » robotique, sans déranger le corps cellulaire du neurone. (Ca ne va pas être très évident, je sais, mais nous faisons ici une expérience de pensée. Les Puissances s'occuperont de l'upload réel.)

4b. Les dendrites robotiques continuent à recevoir des entrées des autres neurones, et les transmettent aux dendrites du neurone dans la coquille. La sortie du neurone biologique passe de l'axone du neurone à l'axone robotique de la coquille, qui lit la sortie et la transmet inchangée à sa synapse extérieure.

4c. L'axone robotique envoie en sortie 99% de l'impulsion biologique reçue plus 1% de l'impulsion robotique calculée. Puisque, par hypothèse, le neurone est parfaitement simulé, cela ne change en rien la sortie, seulement la chaîne causale.

4d. La pondération est ajustée jusqu'à ce que la sortie soit à 100% la sortie calculée.

4e. Le neurone biologique est écarté.

En supposant que nous pouvons simuler un neurone individuel, et que nous pouvons remplacer des neurones avec des analogues robotiques, je pense que ceci démontre en détail la possibilité de l'upload, en supposant que la conscience soit strictement une fonction des neurones. (Et si nous avons des âmes immortelles, alors l'upload est *vraiment* facile. Débrancher l'âme du cerveau. Copier toute information non stockée dans l'âme. Brancher l'âme au nouveau substrat. Upload terminé.)

Arrivé à ce point, il est d'usage de spéculer sur comment on procède pour manger, boire, se promener. Les gens disent qu'ils n'ont pas envie de renoncer à la réalité physique, ils s'inquiètent de savoir s'ils auront ou non une puissance de calcul suffisante pour simuler un monde hédoniste correspondant à leurs rêves les plus fous, et ainsi de suite, *ad nauseam*. Il est établi que Vinge lui-même, découvreur de la Singularité, s'est demandé si notre vrai moi serait dilué par la Transcendance.

J'espère qu'arrivé à cette partie de la page vous avez été suffisamment impressionné par la puissance, la portée, le caractère incompréhensible et la Transcendance générale de la Singularité pour que ces spéculations vous semblent *idiotes*. Si vous souhaitez rester non dilué, vous serez en mesure de faire ce choix. Vous pourrez faire des sauvegardes. Vous pourrez préserver votre personnalité indépendamment du substrat. Les seules personnes qui doivent s'inquiéter d'être diluées *contre leur gré* sont les premiers humains qui Transcenderont, et même eux n'auront sans doute pas à s'inquiéter s'il y a une surintelligence Amicale, d'origine artificielle, pour agir comme guide de transition.

Bien sûr, il est possible qu'au-delà d'un certain seuil d'intelligence tout être *veuille* être dilué. Se faire du souci à propos de cette possibilité semble extraordinairement absurde, comme des enfants qui s'inquiéteraient de ce que, adultes, ils n'auront plus envie d'être inconsidérément cruels envers les autres enfants. Si vous *voulez* être dilué, ça n'est pas un *mal* dont nous devons nous soucier.

Peut-être, après la Transcendance, vous serez différent. Si c'est le cas, alors ce changement est inévitable et vous ne pouvez rien y faire. Le cerveau humain a un nombre

fini de neurones, et pour cette raison un nombre fini d'états possibles. À terme, vous mourrez, vous entrerez dans une boucle éternelle, ou vous Transcenderez. À long terme… *vraiment* à long terme… la mortalité n'est pas une option.

De la même façon, il est inutile de s'inquiéter de ce que des Puissances hostiles anéantiront nécessairement l'humanité. S'il s'avère que tous les buts sont en dernière analyse arbitraires, alors il est concevable qu'une Puissance mal programmée pourrait se retrouver avec des buts qui la rende hostile à l'humanité ; c'est un risque d'ingénierie, et minimiser ce risque est une tâche d'ingénierie. Mais des émotions comme « l'égoïsme » et le « ressentiment » n'apparaissent pas spontanément dans les intelligences artificielles, contrairement à ce qui se passe pour les besoins de l'intrigue dans les fictions scientifiques éculées. Le ressentiment est une adaptation fonctionnelle complexe qui a évolué chez les humains sur une durée de millions d'années ; elle n'apparaît pas simplement à partir de nulle part. Même la tendance à évaluer votre propre groupe comme ayant plus de valeur est le produit de l'évolution, ainsi que la tendance même à penser en termes de « nous » et « eux ».

Pourquoi une surintelligence rationnelle générique catégoriserait-elle l'humanité comme dénuée de sens ? Les seules circonstances dans lesquelles ceci serait une conclusion *inévitable*, c'est si la vie humaine *est* dénuée de sens, si le manque de sens est un *fait* indépendant de l'observateur. Et même cela ne serait pas assez pour signifier la perte de l'humanité ; l'action d'exterminer l'humanité devrait aussi être pourvue de sens, à nouveau en tant que fait indépendant de l'observateur. Ce qui signifierait que tout humain suffisamment intelligent se suiciderait. Et si c'est ainsi, autant dire qu'il n'y a rien qu'on puisse y faire.

En fin de compte, personne ne sait ce qui se trouve de l'autre côté de la Singularité, pas même moi. Et *oui*, il faut du courage pour franchir le seuil de cette porte. Si les enfants à naître pouvaient choisir de quitter ou non l'utérus, sans savoir ce qui se trouve à l'autre bout de la filière génitale — sans savoir s'il y a *quoi que ce soit* —, combien le feraient ? Mais au-delà de la filière génitale est là où se trouve la réalité. C'est là où les choses se passent. Rester dans l'utérus indéfiniment, même si nous le pouvions, serait vain et stérile.

5. Le sens de la vie provisoire

> — Franchement, je ne comprends pas pourquoi l'évolution a fait des humains des créatures aussi inconsidérées et avec une vue aussi courte.
> — Oui, ça ne peut pas continuer comme ça indéfiniment.
> — Tu penses que nous deviendrons plus intelligents ?
> — C'est une des deux possibilités.
>
> *Calvin et Hobbes*

Depuis la rédaction initiale de ce document en 1996, la nanotechnologie est devenue de notoriété publique. Je m'attends à ce que tout le monde ait maintenant entendu parler de l'idée d'atteindre un contrôle complet sur la structure moléculaire de la matière. Ceci permettrait de créer de la nourriture à partir de déchets organiques, de guérir des colonnes vertébrales brisées, d'inverser le vieillissement, de rendre tout le monde sain et riche, et de délibérément anéantir toute vie sur la planète. À vrai dire, les utilisations militaires brutes et destructives seraient probablement bien plus faciles que les utilisations complexes et créatives. Toute personne qui a déjà lu un livre d'histoire peut se faire une idée de ce qui arrive ensuite.

Les « boucliers actifs » pourraient suffire contre des démarrages de « *grey goo* », mais pas contre une nano renforcée utilisée comme arme, parfaitement capable d'utiliser des armes de fusion pour traverser les boucliers actifs. Et pourtant, malgré cette menace, nous ne pouvons pas même essayer de supprimer la nanotechnologie ; ça ne fait qu'augmenter la probabilité que les méchants l'obtiennent d'abord [15].

Mitchell Porter appelle cela « la course entre le surarmement et la surintelligence ». La civilisation humaine continuera à changer jusqu'à ce que, soit nous créions une surintelligence, soit nous nous anéantissions nous-mêmes. Ce sont les deux états stables, les deux « attracteurs » dans le système. Peu importe combien de temps ça prend, ou combien de cycles de nanoguerre-et-régénération se produisent avant la Transcendance ou l'extinction finale. Si le système continue à changer, sur un millier d'années, ou un million d'années, ou un milliard d'années, il se stabilisera finalement dans l'un ou l'autre attracteur. Mais à mon avis, selon toute vraisemblance, la question va être réglée *maintenant*.

Et la possibilité de destruction n'est pas non plus la seule raison pour s'élancer aussi vite que possible vers la Singularité. Il y a aussi la somme permanente de misère humaine, qui n'est pas seulement un problème pratique, pas seulement un problème éthique, mais un problème purement moral à part entière. Il se passe aujourd'hui dans le monde des choses vraiment horribles. Si j'avais le choix entre supprimer les quartiers infestés par le crack ou supprimer l'Holocauste, je ne sais pas lequel je choisirais. Je sais en tout cas quel projet a le plus de chance de succès.

Vous êtes-vous déjà posé les Grandes Questions de la vie, de l'Univers et de tout ? Vous êtes-vous déjà demandé si cela importait vraiment, du point de vue cosmique, si vous restiez au lit ce matin ? Vous êtes-vous déjà plongé dans les problèmes difficiles de l'éthique, ou de la conscience, ou de la réalité, pour réaliser *qu'il n'y a pas* de justification humainement compréhensible pour l'expérience subjective, pour sortir du lit, ou pour l'existence de quoi que ce soit ? Comment pouvons-nous faire quoi que ce soit, nous fixer des buts quelconques, sans connaître le Sens de la vie ? Comment pouvons-nous justifier la poursuite de notre participation à la foire d'empoigne générale si nous ne savons pas dans quel but ? À quoi est-ce que ça *sert* ?

On ne sait pas. Il faut deviner, et agir selon notre meilleure estimation. Indépendamment des probabilités absolues, la surintelligence a *davantage* de chances de découvrir le vrai bien moral, d'avoir le pouvoir de le réaliser et la volonté de le réaliser. L'état dans lequel la surintelligence existe est, avec une haute probabilité quel que soit le véritable sens de la vie, préférable à l'état actuel. Voilà donc le sens provisoire de la vie, et ça fait l'affaire... mais on est encore loin, loin de la certitude, ou de savoir vraiment ce qui se passe !

J'en ai *assez*. J'en ai assez des fumeries de crack, des dictatures, des chambres de torture, de la maladie, de la vieillesse, de la paralysie, et de la famine. J'en ai assez d'un taux de mort planétaire de 150'000 êtres conscients par *jour*. J'en ai assez de cette planète. J'en ai assez de la mortalité. *Rien* de tout cela n'est nécessaire. Il est temps d'arrêter de détourner le regard de l'agression au coin de la rue, du mendiant sur le trottoir. Il n'est plus nécessaire de regarder nerveusement dans la direction opposée en se répétant le mantra : « Je ne peux pas résoudre tous les problèmes du monde. » Nous *pouvons*. Nous pouvons y *mettre un terme*.

Et ainsi j'ai perdu, non pas ma foi, mais ma *suspension d'incrédulité*. Aussi étrange que la Singularité puisse sembler, il y a des moments où elle semble beaucoup plus raisonnable, beaucoup moins arbitraire, que la vie humaine. Il y *a* une meilleure voie ! Pourquoi rationaliser cette vie ? Pourquoi essayer de faire semblant qu'elle tient debout ? Pourquoi faire comme si elle était brillante et heureuse ? Il y a une *alternative* !

Je ne dis pas qu'il n'y a pas de plaisir dans cette vie. Il y en a. Mais *toute* tristesse est inacceptable. Le temps est venu d'arrêter de nous *hypnotiser* dans la croyance que la douleur et le malheur sont désirables ! Peut-être que la perfection n'est pas *atteignable*, même de l'autre côté de la Singularité, mais ça ne signifie pas que les défauts et les failles soient *acceptables*. Le temps est venu *d'arrêter de faire semblant que ça ne fait pas mal* !

Nos frères humains hurlent de douleur, notre planète va probablement être réduite en cendres ou changée en *goo*, on ne comprend rien à ce qui se passe, et la Singularité résoudra ces problèmes. Je déclare qu'*atteindre la Singularité aussi vite que possible* est le sens provisoire de la vie, la définition intérimaire du Bien, et la fondation de mon système éthique jusqu'à nouvel ordre.

6. Atteindre la Singularité

Pour citer une version antérieure de ce texte :

> Il est probable que de nombreux chercheurs placés sur des chemins qui conduisent à la Singularité passent un temps précieux à écrire des demandes de bourse, ou à faire d'autres choses qui pourraient être faites par des assistants de laboratoire. Il serait bon qu'existe une fondation de soutien liée à la Singularité pour permettre que ces personnes ne soient pas distraites. Il y a probablement un chercheur vivant aujourd'hui — Hofstadter, Drexler, Lenat, Moravec, Goertzel, Chalmers, Quate, quelqu'un qui vient de recevoir son diplôme, ou même moi — qui est *la* personne qui arrive à la Singularité. *Chaque heure* dont cette personne est retardée est une heure de plus avant la Singularité. Chaque heure, six mille personnes meurent. Peut-être devrions-nous faire quelque chose pour cette personne qui passe un quart de son temps et de son énergie à écrire des demandes de bourse [16].

Ceci résume le principe de base derrière l'accélération de la Singularité ; il y a un projet, quelque part, qui créera une intelligence plus-qu'humaine. Ce projet, probablement dans le domaine de l'IA, sera soutenu par des avancées dans trois ou quatre autres domaines, comme la science cognitive, le matériel d'ultra-informatique à grande vitesse, des travaux antérieurs en IA, et peut-être des sources d'idées comme les BCI (interface cerveau-ordinateur). Les chercheurs travaillant sur ces projets mangent de la nourriture, portent des habits et regardent des émissions de télévision qui ont été produites par l'économie mondiale. Toute activité productive, où que ce soit dans la chaîne, peut être considérée comme soutenant la Singularité.

Pour soutenir la Singularité de façon indirecte, vous pouvez continuer à vous concentrer sur votre travail quotidien, du moins en supposant que vous êtes un agriculteur plutôt qu'un avocat faisant des recours collectifs. Les neurologues peuvent étudier la science cognitive, les programmeurs peuvent étudier l'IA, et essayer d'être prêts quand la Singularité aura besoin d'eux. Et les diverses organisations transhumanistes, comme l'Extropy Institute et Foresight, peuvent permettre des formes d'aides plus immédiates.

Tout cela est très bien, mais certains d'entre nous aimeraient avoir l'occasion d'*accélérer* la Singularité, d'apporter une aide directe à la création d'une intelligence plus-qu'humaine.

Quatre ans après la publication de la version 1.0 de ce texte, il y a maintenant officiellement un Institut de la Singularité ! La programmation n'a pas encore démarré sur le projet d'IA, mais le travail continue sur *Coding a Transhuman AI 2*, et quand le document de conception sera complet, *nous commencerons* à programme et *nous écrirons* une IA germe. Nous venons de recevoir — au moment de la dernière révision de ce texte — le statut d'exonération d'impôt, et nous acceptons les donations !

Et le temps continue à s'écouler, et un autre bout de la vie comme nous la connaissons se consume...

Je pense qu'il est raisonnable de dire que je peux maintenant visualiser une trajectoire complète conduisant à la Singularité, j'ai une certaine idée de ce qu'il faudrait pour y arriver et de combien ça coûterait, et je pense que nous pourrions probablement le faire d'ici 2010. Nettement plus tôt, en pouvant compter sur un financement et sur des problèmes de recherche qui s'avèrent solubles.

Alors au diable la Loi de Moore. La Singularité se produira quand nous la *provoquerons*.

J'aimerais aussi faire quelques remarques sur comment *ne pas* arriver à la Singularité.

Comme le disait une version antérieure de ce document : « Cette page n'est pas un appel aux armes dans le sens ordinaire. » Je me suis toujours profondément méfié de la tendance humaine à former des organisations sociales. Les organisations tendent à se perpétuer elles-mêmes, plutôt qu'à résoudre des problèmes. Le *mème* de la Singularité est d'une grande puissance. Il ne faut pas permettre qu'il tombe dans les schémas ordinaires, *faciles*. La Singularité ne sera pas avancée par un culte, une société d'admiration mutuelle, ou une bande de cinglés. Drexler a été confronté au même problème avec la nanotechnologie.

Le Principe d'indépendance, dans les Principes singularitariens, est un garde-fou. Pour le résumer, le Principe d'indépendance rejette l'idée qu'un Singularitarien puisse avoir une quelconque forme d'autorité sur un autre. Pour des raisons qui me sont propres, j'ai adopté la Singularité comme un but personnel. Si je peux être plus efficace en travaillant avec d'autres Singularitariens, très bien. Mais ce qui m'importe est la Singularité, pas le Singularitarianisme.

Un autre garde-fou est le Principe d'intelligence, qui affirme que le caractère intelligent ou stupide d'une idée est logiquement prioritaire sur son caractère pro- ou anti-Singularité. (On penserait que tout ça est d'une évidence flagrante, à moins, bien sûr, d'avoir jamais lu un livre d'histoire, parlé à d'autres humains, ou allumé une poste de télévision.)

Il y a un autre garde-fou qui n'est pas dans les Principes. C'est pour mettre en évidence cette idée que j'ai à l'origine écrit ce texte. C'est un dernier conseil : *Ne donnez pas dans l'utopie.*

Ne décrivez pas la vie après la Singularité dans des termes mirifiques. Ne la décrivez pas du tout. Je crois que le point le plus bas jamais atteint dans la prédiction de l'avenir se trouvait dans les quelques brefs paragraphes de *Unbounding the Future* que j'ai lus, quand les auteurs décrivent un piéton qui se fait renverser et sa main qui guérit miraculeusement. C'est ridicule. *Piéton ? Écraser ? Main ? Voitures* dans un monde *nanotechnologique* ? Pourquoi pas une bande de *singes* décrivant la facilité d'attraper des bananes avec un esprit humain ?

Comme l'écrit Drexler :

> J'aimerais insister sur le fait que j'ai été invité à donner des conférences à des endroits comme la série de colloques en physique dans le principal centre de recherche d'IBM, au PARC de Xerox, et ainsi de suite, donc ces idées sont prises au sérieux par des gens techniques sérieux, mais la réaction est mélangée. On souhaite que la réaction soit aussi positive que possible, aussi je prie instamment chacun d'entre vous de *bien vouloir réduire au minimum le niveau de culte et de conneries* [17], et même d'être plutôt mesuré en parlant des conséquences extraordinaires, qui sont de fait réelles et techniquement défendables, parce que ça n'est pas ainsi qu'elles apparaissent. Les gens ont besoin que leur pensée aborde les conséquences à long terme de façon progressive, ça n'est pas par là qu'il faut commencer.

Le problème avec les gens qui exposent leurs visions d'utopie d'un monde nanotechnologique est que les conséquences qu'ils tirent ne sont *pas assez extravagantes*. Entendre des histoires de blessures qui guérissent instantanément, ou de n'importe quel objet matériel instantanément disponible, ne donne pas l'impression d'entrevoir *l'avenir*. Cela vous donne l'impression d'être le spectateur du fantasme enfantin d'omnipotence d'une personne sans imagination, et cela vous prédispose à traiter la nanotechnologie de la même façon. Pire, cela attire d'autres personnes avec des fantasmes d'omnipotence sans imagination. C'est la meilleure façon de se transformer en une bande de révolutionnaires de salon, buvant du café et planifiant la Révolution sans jamais rien faire.

Je suppose que je ne devrais pas être trop sévère avec les personnes qui donnent dans l'utopie nanotech. Certains d'entre eux peuvent être de véritables chercheurs ou des auteurs de fiction scientifique ou d'autres personnes faisant des choses utiles ; certains d'entre eux sont peut-être des employés du rang essayant sincèrement d'arriver quelque part et qui se sont juste laissés attraper dans le manque général d'imagination ; et bien sûr, aucun d'eux ne s'est rendu au Low Beyond. Une fois que vous avez lu cette page, cependant, il n'y a plus d'excuse.

L'objet de cette page est de scruter la Singularité ; l'impression que cela produit, le Beyond, la fin de l'histoire, et des choses qui dépassent l'entendement humain. L'intention est de provoquer une impression *d'avenir*, et j'espère que mes lecteurs seront enclins à voir la nanotechnologie, l'intelligence artificielle, la neurologie, et toutes les autres voies vers la Singularité de la même façon : comme faisant partie de *l'avenir*.

J'espère que cela attire les bonnes personnes.

Dans un moment d'aberration, je me suis inscrit à la liste de diffusion des Extropiens. Ces personnes savent ce que « Singularité » veut dire. En théorie, ils savent ce qui se prépare. Et pourtant, tandis que j'écris [en 1996 ; ça s'est un peu amélioré en 1999], des gens qui devraient *vraiment* être plus éclairés discutent de savoir si les transhumains auront assez de puissance de calcul pour simuler des Univers privés, si la quantité de puissance de calcul disponible aux transhumains est limitée par les lois de la physique, si quelqu'un qui est uploadé dans un trans-ordinateur est vraiment la même personne ou seulement une imitation d'une incroyable fidélité, et — plus incroyable encore — si nos futurs mois inimaginablement intelligents feront encore l'amour.

En quoi est-ce que cela nous concerne ? Pourquoi devons-nous savoir cela ? Ne se pourrait-il pas, *éventuellement*, que ces problèmes peuvent attendre le temps où nous serons cinq fois plus intelligents et que certaines de nos taches aveugles auront été nettoyées ? Dans l'immédiat, chaque être humain sur cette planète a une préoccupa-

tion : *Comment atteindre la Singularité aussi vite que possible ?* Ce qui se produit après n'est *pas notre problème* et je *déplore* ces représentations exclamatives, sans imagination, tellement écœurantes qu'elles font vomir et tout simplement *ennuyeuses* d'un futur avec des ressources illimitées et des mortels totalement inchangés. *Laissez les problèmes de la transhumanité aux transhumains.* Nos chances de tomber juste sont les mêmes que celles d'un poisson qui conçoit un avion avec des algues et des cailloux.

Notre seule responsabilité est de produire quelque chose de plus intelligent que nous ; tout problème au-delà de cela n'est pas de notre *ressort.*

Comment maintenir l'économie mondiale active pendant encore au moins dix ans ? Qui est prêt à financer un projet d'IA ? Qui devons-nous recruter pour un projet d'IA ? Comment pouvons-nous éviter les réactions hostiles anti-technologie habituelles ? Et que faisons-nous si la nanotech arrive en premier ?

Telles sont les questions pratiques auxquelles nous ferons face dans l'avenir immédiat. Les bonnes questions, et les réponses, sont la préoccupation *appropriée* des listes de diffusion. (Et, depuis 2000, du Singularity Institute.) Je n'ai pas d'objection à laisser l'imagination courir librement. Après tout, c'est ainsi que tout cela a démarré. Mais ne vous *impliquez* pas émotionnellement, n'essayez pas de prétendre que votre visualisation de l'autre côté de l'aurore a une chance d'être *correcte*, et consacrez plutôt votre temps à *faire* la Singularité.

Up and Out,
Eliezer S. Yudkowsky

Notes

[1] Ca a en réalité augmenté récemment, atteignant 55% par année en 1998.
[2] La puissance d'un cerveau humain est estimée entre 10^11 et 10^26 opérations en virgule flottante par seconde. Le chiffre utilisé ici est 10^17 opérations/seconde, qui résulte des hypothèses suivantes : 10^11 neurones, 10^3 synapses par neurone et un taux d'impulsions d'environ deux cents impulsions par seconde, à quoi s'ajoute un facteur de 5 pour faire bonne mesure, ce qui nous donne 10^17. Pour des calculs plus sophistiqués, qui concluent à un chiffre de 10^14 opérations/seconde, voyez When will computer hardware match the human brain de Hans Moravec.
[3] Dans *Mind Children* de Hans Moravec.
[4] S'il existe un terme mathématique consacré pour cela, faites-moi savoir.
[5] Et c'est sans même utiliser l'informatique quantique.
[6] Aussi, si nous voulions une fonction qui modèle réellement la situation, T(10^17), ou T(humain), devrait en ce moment être égal à 10^12, ou la puissance d'un ordinateur, et Sn devrait être égal à Sn-1 + T(Sn-1).
[7] Pixels : éléments d'image. Chacun des minuscules points composant l'écran qui affiche ce document est un pixel.
[8] Bien que les personnes exceptionnellement non-stupides s'avèrent généralement être également exceptionnellement intelligentes.
[9] Vous ouvrez une bactérie, insérez l'ADN, et laissez l'installation de biofabrication se mettre au travail.
[10] Un million de milliards d'opérations en virgule flottante par seconde.
[11] http://www.zyvex.com/nanotech/howlong.html
[12] Ca n'est pas une connaissance certaine, bien sûr, mais ça semble assez probable.
[13] D'où mon adresse e-mail, sentience@pobox.com. Ceci exprime également le principe moral qui consiste à refuser de s'identifier avec quelque groupe que ce soit qui inclue moins que la somme de toute vie consciente.
[14] (En association avec Amazon.com.)
[15] Voyez CRNS Time, Deadlines, et If nanotech comes first dans The Plan to Singularity ; également, le Principle of Nonsuppression dans les Singularitarian Principles.
[16] Un peu nettoyé par rapport aux versions 1.0 et 1.1 de ce document.
[17] C'est moi qui souligne.