Singularité technologique, Singularité théologique : la possibilité d'une théogonie

François Kammerer
(Université de Paris IV)

I. Introduction

Mon but, dans cet article, est d'argumenter en faveur de la thèse suivante : du point de vue du matérialiste athée, c'est-à-dire pour celui qui nie l'existence actuelle et passée d'un Dieu et admet la thèse selon laquelle rien n'existe en dehors de l'Univers matériel, une théogonie, entendue comme l'engendrement (au sens d'avènement) d'un Dieu, ou du moins d'un être de type divin, constitue une possibilité à prendre au sérieux.

Si cette possibilité est à prendre au sérieux, ce n'est pas en vertu d'un raisonnement métaphysique. Mon argumentation prend plutôt pour point de départ la considération de possibles progrès technologiques futurs de l'humanité ; progrès considérés par ceux qui prévoient l'avènement d'une « Singularité Technologique ». C'est en prolongeant les réflexions déjà entamées par les penseurs de cette Singularité que j'en viendrai à avancer la thèse selon laquelle il se peut que l'humanité parvienne, dans le futur, à un niveau de sophistication technologique et d'organisation fonctionnelle telle qu'elle puisse en arriver à engendrer un être de type divin.

Par « être de type divin », j'entends une entité qui, si elle ne possède pas strictement les caractéristiques associées, dans la tradition occidentale, au divin (omniscience, omnipotence, éternité, nécessité de l'existence), est toutefois porteuse de propriétés qui la rendent capable de jouer *pour nous, humains* (y compris les humains actuels et passés) un rôle proche de celui du divin. Une telle entité serait donc très savante, puissante et durable ; surtout, elle serait capable de dispenser aux humains futurs, présents et passés, la plupart des biens et des perspectives que l'on associe à la notion de « salut » : résurrection, immortalité, béatitude, rétribution différentielle des personnes ressuscitées.

Ma thèse, si elle est correcte, implique que cette théogonie future pourrait avoir des répercussions fondamentales sur nos propres destins d'individus actuels. Pour cette raison, après avoir argumenté en faveur de ma thèse principale, je tenterai d'explorer les éventuelles conséquences pratiques que nous devrions en tirer, afin de savoir dans quelle mesure la

possibilité d'une telle théogonie devrait affecter les choix que nous faisons et la manière dont nous vivons.

II. Quelques clarifications

Mon but est donc d'argumenter en faveur de la thèse que je nommerai à partir de maintenant la « thèse de la théogonie ». Cette thèse, déjà évoquée en introduction, peut s'énoncer ainsi : si l'on admet que le matérialiste athée a raison, c'est-à-dire si l'on admet qu'il n'existe pas actuellement ou qu'il n'a pas existé dans le passé de Dieu et que tout ce qui existe est l'Univers matériel, doté uniquement (en dernière instance) de propriétés matérielles[1], il faut néanmoins prendre au sérieux la possibilité d'une théogonie. Cette théogonie consisterait en l'engendrement, dans le futur, en raison des progrès technologiques de l'humanité et sans rupture des lois naturelles, d'un être de type divin : une entité considérablement plus puissante et plus savante que les humains, capable de jouer pour les humains passés, présents et futurs le rôle généralement attribué à Dieu (notamment en ce qui concerne la question du salut).

Avant de donner corps à cette thèse et de produire les arguments en sa faveur, je désire effectuer quelques précisions. Tout d'abord, il me faut distinguer ma thèse de quelques thèses apparemment voisines.

La thèse de la théogonie ne consiste pas à affirmer que l'événement nommé « théogonie » constitue un événement probable, à l'arrivée duquel nous *devons* croire. Ma thèse est moins forte que cela : je veux plus modestement affirmer l'idée selon laquelle il s'agit d'une possibilité à prendre au sérieux. Une possibilité, en tout cas, qui n'est ni totalement folle, ni totalement absurde (contrairement à ce qu'elle peut sembler être au premier abord) et à laquelle il convient d'accorder un degré de probabilité non nul et non négligeable.

Cette restriction étant opérée, il me faut également distinguer ma thèse de positions plus classiques. Tout d'abord, il ne faut pas confondre la thèse de la théogonie avec l'idée assez consensuelle selon laquelle, même si nous sommes des matérialistes athées convaincus (par exemple parce que nous jugeons que les raisons de croire au matérialisme athée sont nettement supérieures aux raisons de croire à toute forme de déisme[2]), nous devons néanmoins admettre qu'il est envisageable que nous nous trompions. Une thèse de ce type me semble facile à accepter : je crois que Dieu n'existe pas,

[1] À partir de maintenant, j'utiliserai simplement l'expression « matérialisme athée » pour désigner cette position.

[2] J'utilise « déisme » en un sens générique, qui l'oppose simplement à l'athéisme. Je ne fais donc aucune distinction entre déisme et théisme.

mais j'admets que l'existence de Dieu constitue néanmoins une possibilité épistémique que je ne peux jamais éliminer avec certitude. Par exemple, il se pourrait qu'un Dieu existe mais qu'il ait souhaité demeurer caché jusqu'à maintenant, de manière à ce que sa présence ne soit pas rendue manifeste, ni même accessible à l'enquête rationnelle et empirique. Cette thèse, que l'on pourrait nommer « thèse de la possibilité épistémique du déisme », n'est pas la thèse que je désire défendre. À titre de différence principale, ma thèse n'implique pas que l'athée matérialiste a, en dépit des apparences, tort. Au contraire, elle présuppose qu'il a raison.

Il ne faut pas non plus confondre ma thèse avec l'idée, un peu moins classique mais qui me semble également relativement consensuelle, selon laquelle même si l'athée matérialiste a raison *jusque-là*, il pourrait avoir tort dans le futur, au sens où l'engendrement d'un Dieu constitue toujours une *possibilité métaphysique* – quoiqu'il s'agisse probablement d'une impossibilité physique[3]. Cette idée me semble relativement acceptable, pour peu que l'on ne considère pas que les lois de la nature constituent des lois métaphysiquement nécessaires[4]. En effet, si l'on admet que les lois de la nature sont des régularités métaphysiquement contingentes, il est métaphysiquement possible qu'un miracle, qui rompe ces lois, intervienne dans le futur. En ce sens, il est métaphysiquement possible qu'un éléphant apparaisse soudainement dans mon salon, ou que la Tour Eiffel disparaisse pour laisser place à un ourson en guimauve. De tels événements seraient miraculeux, mais ils constituent néanmoins des possibilités métaphysiques (à titre d'exemples contraires, il n'est pas métaphysiquement possible que 2 et 2 ne soient plus égaux à 4, ou qu'existe un carré rond ; il n'est pas non plus métaphysiquement possible que Romain Gary ne soit plus identique à Emile Ajar, ou que la chaleur ne soit plus identique à l'agitation moléculaire)[5]. De même, il est métaphysiquement possible qu'un miracle

[3] Il me semble qu'une thèse de ce type est défendue, par exemple, par Quentin Meillassoux. Voir : Q. Meillassoux, « Deuil à venir, dieu à venir », in *Critique* 704-705, 2006, pp. 105-115.

[4] Ce qui ne semble pas impossible à soutenir. Sydney Shoemaker, entre autres, est célèbre pour avoir défendu l'idée selon laquelle les lois de la nature sont métaphysiquement nécessaires, en raison du fait que les propriétés physiques sont individuées par leurs pouvoirs causaux. Sur cette question, voir notamment : S. Shoemaker, « Causality and Properties », in *Time and Cause: Essays Presented to Richard Taylor*, éd. par P. van Inwagen, Dordrecht (Netherlands), Reidel, 1980, pp. 109-135; S. Shoemaker, « Causal and Metaphysical Necessity », in *Pacific Philosophical Quarterly* 79, 1998, pp. 59-77. Pour une présentation en français, voir : M. Esfeld, « La théorie causale des propriétés », *Klesis* 13, 2009.

[5] On fait généralement remonter ce concept de nécessité métaphysique, qui ne se réduit pas à la nécessité logique, et qui s'applique notamment aux énoncés d'identité, aux travaux de Saul Kripke. Voir, sur cette question : S. Kripke, *Naming and Necessity*, Harvard

advienne et qu'un Dieu omnipotent et omniscient soit créé (dans ce cas toutefois, ce Dieu ne pourrait pas être éternel, ni nécessaire, ni encore avoir été le créateur du monde).

Si la thèse de la théogonie se distingue de cette thèse, c'est que la théogonie dont je veux montrer la possibilité prendrait place dans le cadre des lois physiques du monde actuel. Je veux en effet défendre l'idée selon laquelle, étant donné les lois physiques de notre monde, et en l'absence de miracle qui suspendrait ces lois, il est envisageable que soit engendré un être de type divin.

La thèse de la théogonie n'est donc pas à proprement parler une thèse qui concerne l'examen des possibilités métaphysiques; en toute rigueur, elle se veut l'expression d'une possibilité physique. Les arguments que je déploierai en faveur de ma thèse ne seront pas pour autant des arguments basés directement sur nos connaissances physiques actuelles. Il s'agira plutôt de considérations assez générales portant sur ce qu'il semble raisonnable de croire au vu des connaissances scientifiques disponibles et de l'évolution technologique récente de l'humanité. Je convoquerai certes des débats de nature métaphysique, qu'il s'agisse de métaphysique des personnes (sur le problème de savoir ce qui compte comme une personne et à quelles conditions une personne peut être dite « avoir survécu ») ou de métaphysique des sciences (sur le problème du déterminisme). Mais, en dépit de ces éléments, la thèse que je veux défendre ambitionne bel et bien de décrire une situation qui ne constitue pas seulement une possibilité métaphysique, mais, si elle est vraie, une possibilité physique. Cet aspect est fondamental si l'on veut comprendre pourquoi, si ma thèse est vraie, ses conséquences doivent être prises beaucoup plus au sérieux que les possibilités métaphysiques que nous venons d'évoquer[6].

À titre de clarification préliminaire, je veux aussi signaler le fait que mon argumentation comportera un grand nombre de présupposés théoriques. Ces présupposés seront exposés au fur et à mesure de l'article. Ils consistent notamment en une posture confiante et « optimiste » quant aux progrès futurs de l'Intelligence Artificielle, en une conception radicalement matérialiste et fonctionnaliste de l'esprit humain (ainsi que de l'intelligence et de la conscience en général). Ils incluent également une théorie « psychologique » de la personne et de l'identité personnelle. Chacune de ces options apparaît plausible en elle-même; il peut néanmoins sembler

---

University Press, 1980; F. Drapeau Vieira Contim et P. Ludwig, *Kripke: Référence et modalités*, Paris, PUF, 2005.

[6] Ces possibilités métaphysiques, de fait, n'ont guère plus d'influence sur notre vie que la possibilité épistémique du scepticisme radical, selon lequel nous pourrions être, en ce moment même, des cerveaux dans une cuve.

difficile au lecteur de les accepter toutes en mêmes temps, d'autant que de nombreuses autres hypothèses seront exigées par mon argumentation, et que l'aboutissement de cette argumentation consistera en une thèse dont le moins qu'on puisse dire est qu'elle est extrêmement spéculative.

Sur ces points, je ne peux demander au lecteur rien d'autre que son indulgence, et la suspension provisoire de son jugement. Pour ma défense, je peux également mettre en avant le fait qu'il peut être tout aussi intéressant et productif, en philosophie comme dans les sciences, d'apporter des indices faibles en faveur d'une thèse extrêmement surprenante, que de fournir une confirmation très solide d'une thèse déjà fort crédible et bien établie.

Ces clarifications ayant été opérées et ces précautions ayant été prises, je vais maintenant entreprendre de défendre la thèse de la théogonie. Au cours de cette défense, elle gagnera, je crois, en précision.

III. La Singularité Technologique

Le point de départ de mon raisonnement est l'idée selon laquelle il faut considérer l'avènement d'une *Singularité Technologique* comme un événement possible (voire même, pour ce cas restreint, probable). Cette idée est bien sûr contestée par de nombreux philosophes, mais elle est également prise au sérieux par beaucoup d'entre eux. Je vais présenter cette idée, puis je la prendrai pour acquise, sans argumenter directement pour elle[7]. J'essaierai ensuite de montrer que cette Singularité Technologique peut être interprétée comme impliquant l'avènement d'un être doté de certaines caractéristiques d'un être divin. Cela constituera la première étape en faveur de la thèse de la théogonie.

L'idée d'une Singularité Technologique est relativement récente et date de la deuxième moitié du vingtième siècle. En quelques mots, elle peut être décrite comme l'idée selon laquelle, dans un futur plus ou moins proche, l'évolution technologique donnera lieu à l'apparition de machines plus intelligentes que les êtres humains, et que, par l'atteinte de ce stade de l'évolution de l'Intelligence Artificielle, le destin de l'humanité s'en trouvera changé de manière irréversible.

[7] Etant donné que mon but est simplement de montrer que l'événement décrit par le terme de « théogonie » est un événement possible, qui possède un degré non-nul et non négligeable de probabilité (mais pas forcément d'inciter le lecteur à croire la réalisation de cet événement plus probable que la non-réalisation de celui-ci), tout ce dont j'ai besoin en réalité est que le lecteur m'accorde que la Singularité possède elle-même ce type de probabilité. En droit, cela devrait bien sûr être également argumenté, mais je pense que cela ne constitue pas une pétition de principe par trop audacieuse. Je pense même qu'en ce qui concerne la Singularité Technologique prise isolément, il n'est pas du tout déraisonnable de la croire probable *simpliciter* ; c'est ce que je vais supposer à partir de maintenant.

Cette prédiction possède des sources diverses : l'idée de base trouve peut-être son origine dans un article du statisticien Irving John Good[8] ; le terme de « Singularité » semble quant à lui avoir été utilisé en premier par l'auteur de science-fiction Vernor Vinge, et popularisé par le même dans un article plus tardif[9]. C'est au cours des années 1990 et surtout des années 2000 que les discussions sur le thème de la Singularité ont connu une croissance notable, quittant le domaine de la seule science-fiction pour être menées par des scientifiques, des industriels, des philosophes[10]. Le thème est arrivé dans la discussion en langue française un peu plus tardivement, généralement de manière indirecte, sous le label plus général des problèmes posés par la notion de « posthumanité »[11].

Jusqu'à récemment toutefois, la plupart des discussions philosophiques autour de cette idée de Singularité Technologique, surtout les discussions menées en langue française, étaient des discussions « d'ordre supérieur » : elles analysaient le discours sur la Singularité comme un certain type de discours prophétique, en tentant d'en opérer une lecture symptomale et d'en fournir un commentaire éclairé – parce que distancié[12]. David Chalmers, dans un article récent[13], fait partie des premiers philosophes à mener une longue discussion de « premier degré » sur la Singularité Technologique. Son article constitue une tentative de mise en forme philosophique et universitaire des idées concernant la Singularité

[8] I.J. Good, « Speculations Concerning the First Ultraintelligent Machine », in *Advances in Computers* 6, 1965.

[9] V. Vinge, « First word », in *Omni*, Janvier 1983, p. 10; V. Vinge, « The Coming Technological Singularity: How to Survive in the Post-Human Era », in *Whole Earth Review*, Hiver 1993.

[10] Symptômes de cet essor récent : le lancement, en 2006, du premier des sommets annuels de la Singularité (*Singularity Summit*) à l'Université de Stanford ; également, la création, en 2008, de l'Université de la Singularité (*Singularity University* ; en fait, une université d'été) financée, entre autres, par la NASA, Google et Nokia.

[11] Voir, dans la littérature philosophique française : D. Lecourt, *Humain, posthumain*, Paris, PUF, 2003, chapitre 2; J.-M. Besnier, *Demain les posthumains*, Hachette Littératures, 2009, chapitres 3 et 6; J. Fahey, « Nous, posthumains: discours du corps futur », in *Critique* 709-710, 2006, pp. 541-552. Parmi les penseurs français, on peut également citer Pierre Teilhard de Chardin, qui n'a certes jamais abordé directement le thème de la Singularité ni discuté avec ses tenants, mais dont les idées peuvent être aujourd'hui relues comme annonçant certains des aspects discutés sous ce terme. Voir notamment : P. Teilhard de Chardin, *Le phénomène humain*, Paris, Seuil, 1970.

[12] On peut noter également un fait culturel : le caractère globalement américain, ou en tout cas non-européen, des discussions « non-distanciées » et de « premier degré » concernant la Singularité. L'Europe et la philosophie européenne semblent ainsi relativement réticentes à prendre au sérieux ce type de réflexions ; à propos des théories de la Singularité, Jürgen Habermas parle par exemple de « spéculations adolescentes sur l'intelligence artificielle supérieure des futures générations de robots ». Voir : J. Habermas, *L'avenir de la nature humaine: vers un eugénisme libéral?*, trad. par C. Bouchindhomme, Paris, Gallimard, 2002, p. 29. ; citation reprise à D. Lecourt, *Humain, posthumain*, p. 11.

[13] D. Chalmers, « The Singularity: A Philosophical Analysis », in *Journal of Consciousness Studies* 17/9, 2010, pp. 7-65.

Technologique. Dans le passage qui suit, son travail constituera ma source principale. D'une manière générale, le présent article lui doit beaucoup, et je tenterai, autant que possible, d'informer le lecteur lorsque je paraphrase ou commente l'article de Chalmers[14].

On peut décrire plus précisément l'hypothèse d'une Singularité Technologique (ou, plus rapidement, « Singularité ») comme suit[15] : dans un futur probable (bien que non certain), les êtres humains parviendront à bâtir une machine ultra-intelligente, c'est-à-dire une machine plus intelligente que n'importe quel humain[16]. Cette machine étant ultra-intelligente, elle sera plus performante que les humains pour concevoir des machines et notamment des machines intelligentes. En conséquence, cette machine ultra-intelligente sera à son tour capable de bâtir une machine ultra-ultra-intelligente, et ainsi de suite. Nous devrions donc assister à ce que I. J. Good nomme une « explosion d'intelligence » (*intelligence explosion*), au sens où devrait apparaître, à partir de la première machine ultra-intelligente, des formes d'intelligence sans cesse plus perfectionnées, capables de surpasser grandement l'intelligence humaine[17]. On peut aussi s'attendre à ce que cette explosion d'intelligence s'accompagne d'une « explosion de la vitesse » (*speed explosion*) : la vitesse de calcul des processeurs tend déjà à doubler tous les deux ans de travail humain (selon une induction basée sur les progrès actuels). Or, une fois la première machine ultra-intelligente atteinte, on devrait pouvoir s'attendre à que la fréquence de doublement de la vitesse

[14] Cet article de David Chalmers a par la suite donné lieu à un symposium, dont le contenu se trouve dans plusieurs numéros spéciaux du *Journal of Consciousness Studies*, dans lesquels 26 auteurs parmi lesquels Daniel Dennett, Ray Kurzweil et Jesse Prinz adressent leurs remarques et critiques à Chalmers. Ces contributions peuvent être trouvées dans le *Journal of Consciousness Studies*, volume 19, numéros 1, 2, 7 et 8. L'une de ces contributions, celle d'Eric Steinhart, se concentre sur les points d'intersections entre la théorie de la Singularité Technologique et d'autres disciplines, dont la métaphysique et la théologie ; à cette occasion, il évoque les interprétations de la Singularité comme théogonie, sans défendre de thèse particulière à ce propos. Voir : E. Steinhart, « The Singularity: Beyond philosophy of mind », in *Journal of Consciousness Studies* 19/7-8, 2012, pp. 131-137. Par ailleurs, pour la réponse de David Chalmers à l'ensemble des critiques de son article original, voir : D. Chalmers, « The Singularity: A Reply to Commentators », in *Journal of Consciousness Studies* 19/7-8, 2012, pp. 141-167.

[15] Je paraphrase ici Chalmers, qui cite lui-même l'article de Good.

[16] C'est par commodité que l'on caractérise cette intelligence artificielle supérieure comme appartenant à une « machine », par opposition aux humains. En effet, il se peut que cette première intelligence artificielle supérieure (et celles qui suivent) résulte d'une augmentation artificielle des capacités intellectuelles d'êtres humains préexistants. Dans ce cas, cette intelligence concernerait un composé d'humain et de machine. Cette précaution ayant été prise, je garderai toutefois, dans mon raisonnement, le vocabulaire de la « machine ».

[17] Rendant cette intelligence humaine elle-même inutile : « La première machine ultra-intelligente est donc la dernière invention que l'homme aura jamais besoin de faire » (*Thus the first ultraintelligent machine is the last invention that man need ever make*). Voir Good, « Speculations Concerning the First Ultraintelligent Machine ».

de calcul soit elle-même multipliée, jusqu'à atteindre un point d'explosion de la vitesse, ou la vitesse de calcul en viendrait à surpasser dans une très grande proportion celle que nous pouvons connaître.

D'après David Chalmers, la thèse de la Singularité peut globalement se comprendre comme la conjonction des thèses de l'explosion d'intelligence et de l'explosion de la vitesse (même si la thèse d'explosion de la vitesse est peut-être moins essentielle) : à un certain point, l'intelligence et la vitesse de calcul augmenteront d'une manière telle qu'elles atteindront un point où elles seront, sinon infinies (il se peut qu'une telle infinité ne soit pas permise par la nature physique de l'Univers), du moins extrêmement importantes au regard de ce que sont l'intelligence et la vitesse de calcul actuellement disponibles (dont disposent les esprits humains et les machines contemporaines). C'est ce point de l'histoire future que l'on nomme « Singularité Technologique ». L'idée fondamentale est qu'à partir de cette « Singularité », l'intelligence disponible sera telle qu'elle ne ressemblera à rien de ce que nous pouvons concevoir : le sort de l'humanité s'en verra probablement affecté de manière irréversible. En vertu de l'explosion d'intelligence, le point nommé « Singularité » devrait donner lieu à des changements qualitatifs d'importance radicale ; en vertu de l'explosion de la vitesse, ces changements devraient avoir lieu très rapidement après ce point de Singularité (c'est-à-dire après l'invention de la première machine ultra-intelligente).

La plupart des débats portant sur la Singularité me semblent se concentrer autour de deux foyers d'interrogation : premièrement, la Singularité est-elle possible ? Deuxièmement, est-elle souhaitable ? Ces deux débats ne seront pas abordés dans mon article, si ce n'est de manière incidente.

La raison n'en est pas que je juge ces débats inutiles ou inintéressants. Le débat sur le caractère souhaitable de la Singularité, par exemple, me semble légitime et essentiel. Son importance justifie que l'on range souvent les discussions autour de la Singularité dans la catégorie des discussions éthiques concernant la posthumanité : dans la mesure, en effet, où l'avènement de la Singularité devrait affecter l'humanité de manière irréversible et radicale, il convient de savoir s'il faut ou non qu'un tel événement ait lieu. Toutefois, poser cette question n'est pas ici mon objet. Je désire défendre une thèse plus précise concernant le type de conséquences que *pourrait* avoir, pour l'humanité, l'avènement de la Singularité. Ces conséquences sont constituées par l'événement que je nomme « théogonie ». Ma thèse ne me semble pas participer directement du débat éthique concernant le caractère souhaitable de la Singularité ; ce, pour

deux raisons. D'une part, je désire simplement explorer une possibilité, et je ne prétends pas que les conséquences que je vais développer (la théogonie) sont celles qu'aurait inévitablement la Singularité. Dans cette mesure, même si la théogonie est souhaitable, cela ne signifie pas automatiquement que la Singularité l'est. D'autre part, même si ces conséquences « théogoniques » sont présentées sous un jour positif, et peuvent pour cette raison être considérées comme la base d'un argument en faveur du caractère souhaitable de la Singularité, cette évaluation positive de la théogonie ne fait pas partie de ma thèse, et je ne m'y engage en aucune façon : je laisse au lecteur le soin de juger du caractère désirable de l'état de fait qui sera envisagé dans mon article sous ce terme.

Les discussions portant sur la *possibilité* de la Singularité, sur la probabilité de son avènement ainsi que sur la date prévisible de celui-ci, sont peut-être aussi nombreux et vifs que les débats éthiques sur son caractère souhaitable. À titre d'exemple, les plus optimistes prévoient la Singularité pour la troisième décade du XXI^e^ siècle ; Chalmers, qui est relativement confiant, pense *probable* l'advenue de la Singularité avant 2100. Je n'entrerai pas ici dans ces débats, dont l'article de Chalmers donne, je pense, un aperçu relativement synthétique et particulièrement clair. Je vais ici considérer la Singularité comme un événement futur probable, bien que non certain. Dans un souci d'exposition et de clarté, je vais toutefois reproduire ici le principal argument que donne Chalmers en faveur de la probabilité de la Singularité. Il n'apporte rien de radicalement nouveau mais consiste plutôt en une reformulation des raisons implicites qui nous font accorder de la plausibilité à cette perspective. Il peut prendre, selon Chalmers, la forme suivante :

(1) Il existera, un jour, une IA (Intelligence Artificielle : une machine à l'intelligence égale à celle des humains)
(2) S'il existe un jour une IA, il existera une IA+ (c'est-à-dire une machine à l'intelligence supérieure à celle du plus intelligent des humains)
(3) S'il existe un jour une IA+, il existera une IA++ (c'est-à-dire une machine à l'intelligence très largement supérieure à celle des humains. Par exemple, au moins autant de fois plus intelligente que les humains que ce que les humains sont intelligents comparés à une souris)

Conclusion : Il existera une IA++ (c'est-à-dire qu'adviendra la Singularité Technologique : une machine dont l'intelligence surpassera

si largement l'intelligence humaine que l'intelligence humaine ne sera plus en mesure de l'appréhender).

Chacune des prémisses semble plausible prise en elle-même, et il semble qu'aucun contre-exemple adéquat ne puisse être adressé à l'une d'entre elles. Or, si l'on accepte les prémisses, la conclusion s'ensuit ; nous avons donc de bonnes raisons d'accepter la conclusion.

Cette IA++, dont l'avènement est affirmé par la conclusion de l'argument, sera donc très largement plus intelligente que les êtres humains. David Chalmers, dans son article, nous donne par ailleurs de bonnes raisons de penser que cette IA++ sera aussi, que nous le voulions ou non, très largement plus puissante que les êtres humains. De même, elle sera très probablement, si ce n'est éternelle, du moins très durable, dans la mesure où elle pourra être réalisée dans des matériaux robustes et susceptibles de réparations indéfinies, s'étendant à la surface de la planète et même au-delà : on peut envisager que sa longévité en vienne à égaler celle de l'Univers physique lui-même.

Donc, si nous acceptons cet argument, et plus généralement les considérations en faveur de la plausibilité d'une Singularité Technologique, nous pouvons accepter l'idée selon laquelle il est plausible qu'advienne un « être divin », ou tout au moins – pour le moment – une partie d'un être divin. Comme évoqué en introduction, j'entends par « être divin » un être qui, s'il ne possède pas les caractéristiques traditionnellement accordées à la divinité dans la tradition occidentale (omniscience, omnipotence, éternité, nécessité), s'il ne possède peut-être pas même certaines des caractéristiques accordées au divin dans la majorité des religions (le fait de préexister aux humains et de n'être pas causalement dépendant d'eux), sera néanmoins le porteur de propriétés telles que nous considérerions un tel être comme incontestablement supérieur à nous : une très grande intelligence, une très grande puissance, une très grande durée (je ne parle pas encore de sa capacité à nous dispenser des biens de salut, qui fera l'objet de la suite de mon exposé). Ces raisons justifient déjà que la « Singularité Technologique » puisse être considérée comme une « *Singularité Théologique* », c'est-à-dire comme une forme de théogonie, au sens où elle cause l'engendrement d'un être divin[18].

[18] L'interprétation « théogonique » de l'avènement de l'IA++ n'est pas absente des théories de la Singularité actuellement disponibles. En dehors de l'article d'Eric Steinhart, on peut ainsi renvoyer à ce que dit Ray Kurzweil, l'un des partisans les plus célèbres de la thèse de la Singularité, à propos de l'émergence de cette dernière : « Une fois que nous aurons saturé la matière et l'énergie de l'univers avec l'intelligence, il se « réveillera », sera conscient et sublimement intelligent. C'est ce qui se rapproche le plus d'un dieu à mes yeux ». Voir : R.

Remarquons que la thèse de la Singularité Technologique est parfaitement compatible avec le matérialisme athée. En réalité, elle semble même, si ce n'est dépendante du matérialisme athée, du moins particulièrement bienveillante à son égard, puisqu'elle suppose que l'intelligence d'un système dépend de sa constitution physique (et notamment que rien de non physique, et donc de non reproductible dans une machine, ne permet à l'intelligence – celle que les humains, par exemple possèdent déjà – d'exister). Accepter la thèse de la Singularité Technologique semble donc appuyer l'idée selon laquelle, au sein d'un Univers sans Dieu et strictement matériel, un être divin, au sens au moins d'un être très puissant, très savant et très durable (comparé aux humains), a des chances non négligeables d'être engendré.

Toutefois, pour défendre la thèse de la théogonie telle qu'énoncée précédemment, il me faut argumenter sur d'autres points plus controversés. Il me faut notamment étayer l'idée selon laquelle cet « être divin » sera capable de jouer à l'égard des humains un rôle comparable, en termes de salut, à celui attribué traditionnellement à la divinité. Ce sera l'objet de ce qui suit.

IV. Singularité : intégration et immortalité

D'après David Chalmers, quatre principales options semblent dessiner le destin de l'humanité dans un monde post-Singularité[19] : *l'extinction* (la disparition à plus ou moins long terme), *l'isolation* (le maintien de l'humanité sans interaction, ou avec une interaction minimale, avec l'IA++), *l'infériorité* (le maintien de l'humanité dans une situation de subordination à l'égard de l'IA++), ou *l'intégration* (nous-mêmes humains « devenons » l'IA++, ou du moins avons nos propres esprits intégrés au sein de l'IA++). Ces quatre possibilités semblent toutes plausibles, si l'on admet préalablement qu'il y aura une Singularité. Il apparaît difficile, étant donné que la manière dont une IA++ pourrait penser nous est inaccessible, d'attribuer raisonnablement des probabilités d'advenue à chacune de ces options – même si une partie des discussions sur la Singularité est dévolue à l'évaluation des meilleurs moyens à notre disposition pour favoriser les issues heureuses pour l'humanité. Toutefois, pour pouvoir argumenter en faveur de la thèse de la théogonie, il me faut supposer que l'intégration se réalise (bien que certaines versions particulièrement bienveillantes de

---

Kurzweil, *Humanité 2.0 : La Bible du changement*, trad. par A. Mesmin, Paris, M21 Editions, 2007, p. 402. ; citation reprise à J.-M. Besnier, *Demain les posthumains*, p. 167.

[19] D. Chalmers, « The Singularity: A Philosophical Analysis », partie 8.

« l'infériorité » puissent également faire l'affaire, je n'en parlerai pas ici). Il me semble qu'il s'agit d'une possibilité, et qu'elle semble même relativement plausible *prima facie*[20]. Je vais donc supposer sa vérité à partir de maintenant.

À quoi pourrait ressembler, d'après Chalmers, une telle « intégration ? ». L'idée est que nous devenons nous-mêmes des êtres ultra-intelligents, par l'amélioration de notre cerveau ou la migration (progressive ou ponctuelle) de nos esprits vers des supports non-organiques. Notons toutefois qu'à long terme c'est la seconde possibilité qui sera sans doute favorisée : si nous désirons conjoindre à cette ultra-intelligence une longévité et une robustesse supérieure, il nous faudra accepter de nous passer de notre substrat biologique originel et accepter la migration de nos esprits, comme structures fonctionnelles et comme organisation de l'information, vers les ordinateurs ou les machines au sens large, au terme d'un procès de téléchargement (*uploading*).

Une question ne me semble pas tranchée : celle de savoir si, à l'issue d'un tel processus, l'ensemble des humains intégrés deviendraient effectivement des IA++ (d'une intelligence équivalente entre elles), ou s'ils deviendraient des êtres plus intelligents que les humains actuels (disons, des IA+), plus ou moins intégrés à une intelligence supérieure collective qui serait proprement l'IA++, et qui concentrerait la plus grande partie des capacités intellectuelles et calculatoires disponibles. Dans ce qui suit, je privilégierai la seconde hypothèse, même si la plupart des points que je vais immédiatement aborder ne dépend pas de cet aspect. Pour l'instant, l'élément qui me semble être d'une importance cruciale est que les esprits humains intégrés se verront transportés vers une forme plus performante et inorganique de fonctionnement.

Un problème philosophique se pose alors : à l'issue de ce téléchargement, les humains pourront-ils être dits avoir, en un sens non trivial, survécu ? D'après Chalmers, ce problème se subdivise en deux sous-questions : tout d'abord, est-ce qu'une version « téléchargée » de mon esprit, sur un support non-organique, serait consciente (ce qui semble une condition certainement nécessaire, et peut-être suffisante, pour compter

[20] Elle semble même d'autant plus plausible si l'on songe qu'il se pourrait que les premiers systèmes plus intelligents que les humains ne soient pas développés comme des ordinateurs extérieurs aux humains, mais comme la conséquence d'extensions et d'améliorations des esprits humains (par exemple, par la création d'interfaces ordinateur/cerveau permettant à un humain d'accroître considérablement et progressivement ses capacités intellectuelles). Dans une telle situation, la Singularité Technologique aurait lieu d'emblée dans le cadre d'une intégration entre humains et machines, ce qui renforcent la plausibilité d'une issue « heureuse » pour l'humanité. Voir sur ce point ce que dit David Chalmers d'une telle « intelligence artificielle à base humaine » (*human-based AI*), *Ibid*., partie 6.)

comme une personne)? Ensuite, est-ce que cette version téléchargée serait *moi* ?

Les débats autour de ces questions, et plus généralement autour de la survie post-téléchargement des personnes, sont nombreux, et mettent en jeu des arguments philosophiques autant que des intuitions. La question, tout d'abord, de savoir si la version « téléchargée » d'un esprit est consciente est indissociable d'une réponse au problème de la conscience, et notamment de la version « difficile » de celui-ci[21]. Toutefois, l'on peut dire que, dans une perspective matérialiste, sauf à penser qu'il existe quelque chose de spécifique qui concerne la biologie des cerveaux humains et qui les rend capables de conscience[22], il convient d'adopter une perspective fonctionnaliste[23], qui, dans la plupart de ses versions, admet l'idée selon laquelle une version téléchargée d'un esprit humain pourrait être consciente. Les promoteurs contemporains du dualisme (dont Chalmers est d'ailleurs le plus célèbre représentant récent) tendent à admettre également une forme de fonctionnalisme – au moins au sens d'une dépendance systématique (ce que l'on nomme, dans le vocabulaire technique de la philosophie analytique, une *survenance*) de la conscience sur l'organisation fonctionnelle du système cognitif concerné. Dans une perspective dualiste de ce type, une version téléchargée sur un matériau inorganique d'un esprit humain serait également consciente[24]. Tout cela semble justifier une forme d'optimisme vis-à-vis du caractère conscient des esprits téléchargés[25].

---

[21] David Chalmers ayant d'ailleurs été parmi les philosophes qui ont le plus fait pour mettre au premier plan des préoccupations philosophiques ce « problème difficile » de la conscience (*hard problem of consciousness*). Voir notamment : D. Chalmers, « Facing up to the problem of consciousness », in *Journal of Consciousness Studies* 2/3, 1995, pp. 200-219, D. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory*, Oxford University Press, 1996.

[22] Cette position, relativement minoritaire, a été défendue par quelques philosophes d'importance, comme John Searle et Ned Block. Voir : N. Block, « Psychologism and Behaviorism », in *Philosophical Review* 90, 1981, pp. 5-43, J. Searle, « Minds, Brains and Programs », in *Behavioral and Brain Sciences* 3, 1980, pp. 417-457.

[23] Selon la thèse fonctionnaliste, être dans un état mental est essentiellement être dans un état susceptible de jouer un rôle fonctionnel donné dans un système adéquat. De ce fait, un même état mental peut être « réalisé » dans une multitude de matériaux différents. De même, un esprit, comme système capable d'être dans certains états mentaux, peut être réalisé dans différents matériaux. La principale source du fonctionnalisme en philosophie de l'esprit, qui peut être aujourd'hui considéré comme la position par défaut, sont les travaux d'Hilary Putnam. Voir notamment : H. Putnam, « Psychological Predicates », in *Art, Mind, and Religion*, éd. par W. Capitan et D. Merrill, Pittsburgh, University of Pittsburgh Press, 1967.

[24] À noter que par « esprit » il ne faut pas nécessairement entendre ici quelque chose qui correspond uniquement à ce qui se produit dans le cerveau, en excluant le reste du corps. L'on pourrait en effet argumenter, sur la base d'une conception de la cognition humaine qui la pense comme nécessairement incarnée (inspirée par exemple par la phénoménologie de Merleau-Ponty), qu'en téléchargeant l'esprit (en tant qu'instancié dans le seul cerveau) sur un support inorganique, nous perdrions la personne humaine, qui exige d'habiter ou

La question de la survie du « moi » au cours du téléchargement pose un autre problème, peut-être plus difficile[26]. La réponse à ce problème, nous explique Chalmers[27], dépend de la théorie de l'identité personnelle que l'on soutient : la théorie biologique[28] (la survie d'une personne requiert le maintien intact de son corps à travers le temps, ou au moins le maintien intact d'une partie spécifique de son corps, comme son cerveau), la théorie psychologique[29] (la survie d'une personne dépend d'une continuité psychologique, basée notamment sur la cohérence des souvenirs et des états mentaux) ou la théorie du « plus proche continuateur »[30] (selon laquelle une personne survit sous la forme de l'entité la plus similaire à elle, parmi celles qui lui succèdent, sous certaines contraintes). Celui qui accepte une théorie biologique prendra une position « pessimiste » et refusera la survie post-téléchargement. À l'inverse, celui qui souscrit à une théorie psychologique sera optimiste sur la survie post-téléchargement. Quant au partisan de la théorie du plus proche continuateur, il fera varier sa réponse suivant que le téléchargement est « destructif » (le modèle « original » biologique de la copie est détruit après téléchargement) ou non-destructif.

« d'exister » un corps. Nous pouvons répondre à cela qu'il doit être possible de télécharger sur un support inorganique non seulement l'organisation fonctionnelle du cerveau, mais aussi celle du corps, en simulant le corps conçu comme interface avec le monde. Bien sûr, la personne qui subsisterait post-téléchargement ne possèderait plus un corps « objectif » (comme système organique, morceau de chair situé dans l'espace et dans le temps), mais elle pourrait toujours disposer de son « corps propre », c'est-à-dire d'un corps vécu, compris comme ouverture au monde, à la passivité et à l'émotion, lien aux autres, « nœud de signification vivantes ». Voir M. Merleau-Ponty, *Phénoménologie de la perception*, Paris, Gallimard, 1945. Ce corps propre des personnes téléchargées ne serait bien sûr pas exactement celui décrit par Merleau-Ponty, mais il en conserverait, je crois, l'essentiel des propriétés pertinentes du point de vue de notre vécu.

[25] D. Chalmers, « The Singularity: A Philosophical Analysis », partie 9.

[26] Les problèmes liés au type de conception de l'identité personnelle exigée par une théorie de la résurrection ne sont pas propres aux théories « matérialistes » et athées de la résurrection telles que celles qui nous occupent en ce moment. La doctrine chrétienne de la résurrection des corps pose des problèmes assez similaires, et il n'est pas évident d'élaborer une ontologie compatible avec elle. Voir notamment : R. Pouivet, « De van Inwagen à saint Athanase, une ontologie personnelle de la résurrection des corps », in *Klesis* 17, 2011.

[27] D. Chalmers, « The Singularity: A Philosophical Analysis », partie 10.

[28] Cette théorie, relativement intuitive, est notamment défendue à l'époque contemporaine par Bernard Williams et Peter van Inwagen. Voir B. Williams, « The Self and the Future », in *Philosophical Review* 79/2, 1970, pp. 161-180; P. van Inwagen, *Material Beings*, Ithaca, Cornell University Press, 1990.

[29] L'origine de cette théorie se trouve chez Locke, voir J. Locke, *Essai sur l'entendement humain*, trad. par J.-M. Vienne, Paris, Vrin, 2001, livre 2, chapitre 27, §§25-26. À l'époque contemporaine, elle trouve de nombreux défenseurs, dont, entre autres, Derek Parfit ou Sydney Shoemaker. Voir notamment: D. Parfit, « Personal Identity », in *Philosophical Review* 80, 1971, pp. 3-27; S. Shoemaker, « Persons and Their Pasts », in *American Philosophical Quarterly* 7, 1970, pp. 269-285. Pour une présentation en français d'une conception psychologique de la notion de personne, voir S. Chauvier, *Qu'est-ce qu'une personne?*, Paris, Vrin, 2003.

[30] Cette théorie a été proposée par Robert Nozick. Voir R. Nozick, *Philosophical Explanations*, Cambridge (Mass.), Belknap Press, 1981.

Il n'est pas évident de choisir entre ces théories (et donc entre l'optimisme et le pessimisme en ce qui concerne la survie post-téléchargement), qui possèdent chacune de bons arguments en leur faveur, et sont toutes soutenues par des intuitions respectables. Afin de manifester la force de cette opposition, on peut, suivant Chalmers, présenter deux cas dont chacun donne du poids à l'une des deux options : l'un confirme l'idée que mon esprit téléchargé pourrait être « moi » (et donc me permettre de survivre), tandis que l'autre appuie la conception selon laquelle cet esprit téléchargé ne serait qu'une copie distincte de moi.

Premier cas : imaginons qu'on crée une copie parfaite de mon esprit (dans son organisation fonctionnelle) sur un ordinateur, sans pour autant me détruire. Il y a donc deux choses : moi, et la version téléchargée de mon esprit. Ces deux choses sont distinctes, et la copie n'est pas moi. Et il semble étrange de dire que, si l'on me détruit, cela suffira à faire en sorte que la version téléchargée de mon esprit devienne « moi ». Donc, une version téléchargée de mon esprit n'est pas moi.

Deuxième cas : imaginons que l'on remplace progressivement, peut-être même assez lentement, chacun de mes neurones par un composant informatique qui en est la copie fonctionnelle parfaite. Le remplacement d'un seul neurone, puis de cent neurones, n'a aucune raison de changer mon identité. Donc, le remplacement progressif de 10%, puis de 20%, puis de tous mes neurones non plus. À la fin du processus, tous mes neurones auront été remplacés, et mon esprit aura, de cette manière, été « téléchargé » sur un support artificiel. Pourtant, cet esprit sera, me semble-t-il, toujours « moi ». Cela semble encore plus frappant si je demeure conscient durant tout le processus.

Chaque cas semble donner un argument fort en faveur de l'une des deux alternatives, en mobilisant les intuitions qui motivent les différentes théories de l'identité personnelle, et il n'est pas évident de trancher entre les deux options. Il se peut aussi que ces débats constituent des combats d'intuitions insolubles, dans le sens où les intuitions que nous mobilisons pour répondre aux questions d'identité personnelle ont été construites et validées par des contextes sociaux si différents de ceux que nous envisageons en réfléchissant à la Singularité qu'elles sont peut être rendues inopérantes, ou partiellement dénuées de sens, dans ces débats[31]. Un tel

[31] La question d'une continuation des esprits humains sous forme de « copies » implémentées dans du matériel informatique est notamment posée dans un ouvrage de l'auteur de science-fiction Greg Egan. Dans le roman, les opposants à la numérisation des esprits, qui pensent qu'une copie numérisée de soi n'a rien à voir avec soi et ne permet en aucune façon une survie post-mortem, tendent à se recruter dans l'ancienne génération,

point de vue donnerait de la force à la position que Chalmers nomme « déflationniste » sur le problème de l'identité personnelle, selon laquelle le problème de l'identité personnelle ne peut pas réellement être résolu, parce qu'il n'a pas vraiment de sens : il n'y a pas de « fait » qui permettrait de le trancher, il n'y a pas de *fact of the matter* derrière la question de savoir si une version téléchargée de mon esprit est « moi », ou non[32]. Savoir si une telle version est « moi » ou non relève en dernier ressort d'une décision sémantique, et non d'une question empirique plus ou moins complexe dont on pourrait trouver la réponse.

Il me semble toutefois qu'il n'est pas déraisonnable de supposer, avec Chalmers, que la vision « optimiste » de la survie post-téléchargement constitue une réponse acceptable (même si nous n'avons pas de certitude absolue qu'elle soit bonne). Un méta-argument en sa faveur pourrait être que les théories psychologiques de l'identité personnelle sont majoritaires dans la philosophie contemporaine. Même si l'on n'accepte pas un tel méta-argument, on peut au moins reconnaître que de telles théories, qui viennent fonder une vision optimiste de la survie post-téléchargement, sont parmi les plus plausibles.

Si l'on accorde la plausibilité de la thèse de la survie post-téléchargement, il devient possible de comprendre en quel sens une deuxième étape en faveur de la thèse de la théogonie est franchie. Car la perspective d'une intégration des humains avec téléchargement des esprits humains sur un support artificiel, ultra performant et durable, promet aux humains qui auront l'occasion d'être ainsi intégrés deux des biens de salut fondamentaux promis dans la tradition religieuse occidentale : l'immortalité et la béatitude.

Le téléchargement de nos esprits permet en effet de rendre l'immortalité accessible ; ou, tout du moins, une certaine forme d'immortalité, prise en un sens non-métaphysique et non-absolu de durée de vie potentiellement extrêmement longue (éventuellement identique à la durée de l'Univers matériel lui-même). Les esprits téléchargés, en effet, seraient réalisés par un matériau qu'on peut supposer extrêmement robuste (en tout cas bien plus que le matériau organique, voué à une dégradation certaine en raison du programme génétique des humains) et offrant des possibilités de réparations indéfinies. Ils seraient donc susceptibles d'une prolongation de l'existence qui, si elle ne donnerait certes pas lieu à une véritable vie *éternelle*, pourrait les faire subsister dans des proportions telles

tandis que les plus jeunes manifestent moins de scepticisme. Voir notamment: G. Egan, *Permutation City*, Millennium Orion Publishing Group, 1994.

[32] À titre personnel, c'est cette position qui a ma préférence, même si je n'argumenterai pas ici pour elle.

que la durée actuelle de l'existence humaine paraîtrait infime en comparaison. Il faut de plus noter que cette augmentation de l'existence serait tout à la fois « extensive » (ces esprits vivrons un plus grand nombre d'années, peut-être des millions d'années) mais aussi « intensive » (au sens où l'augmentation radicale de la vitesse de calcul promise par la Singularité donnera l'occasion d'effectuer un nombre considérablement plus important d'opérations mentales qu'un esprit humain classique pour le même laps de temps : une seconde de la vie de ces esprits téléchargés sur des supports artificiels sera peut-être aussi riche qu'un mois, un an ou dix ans de la vie de nos propres esprits).

Pour ces esprits, immortalité, ou quasi-immortalité, donc. Mais je pense que nous pouvons considérer qu'une telle intégration des esprits humains à l'IA++ permettrait également de réaliser, pour eux, une forme de béatitude. Que faut-il entendre par béatitude ? Dans la tradition occidentale chrétienne, celle-ci est comprise comme le bonheur suprême qui est celui des humains au Paradis. Il s'agit d'un bonheur parfait et éternel. Et, selon Thomas d'Aquin, par exemple[33], ce bonheur consiste essentiellement en la vision de l'essence divine.

Or, un esprit humain téléchargé sur un support artificiel et rendu par-là quasi-éternel, intégré à l'IA++ et rendu par là ultra-intelligent et extrêmement puissant, aurait la capacité de s'accorder un état cognitif et « émotionnel » qui s'apparenterait, en tant que vécu subjectif, à une satisfaction absolue, certaine et continue (et l'on peut supposer que de nombreux esprits feraient le choix d'un tel état cognitif et émotionnel). Par ailleurs, cet esprit humain serait en mesure, par son intelligence augmentée et son intégration fonctionnelle à l'IA++, d'avoir un accès cognitif direct à cette IA++ et aux pensées de celle-ci[34], c'est-à-dire aux pensées de « l'être

[33] Thomas d'Aquin, *Somme Théologique*, trad. par A.-M. Roguet, Paris, Cerf, 1984, Prima secondae, question 3, article 8.

[34] Dans cette description, je considère une situation au sein de laquelle, ainsi que signalé précédemment, différents esprits humains continuent d'exister de manière relativement indépendante, sans être eux-mêmes des IA++, mais en participant d'une certaine manière à cette IA++. Bien entendu, il n'est pas du tout évident que l'intégration des esprits humains se réaliserait de cette manière – même s'il me semble qu'il s'agit d'un modèle plausible. On pourrait imaginer des modèles concurrents. J'en ai évoqué un précédemment, dans lequel tous les esprits humains subsistants et intégrés deviennent eux-mêmes des IA++ (reste à savoir s'il est envisageable que coexistent plusieurs êtres ultra-intelligents, qui auraient tous récolté les fruits de l'explosion d'intelligence et de l'explosion de la vitesse). On pourrait également imaginer que les esprits humains soient intégrés de manière beaucoup plus robuste à une seule IA++. Leurs souvenirs et leurs pensées passées seraient par exemple intégrés comme contenus informationnels dans un système cognitif unique, doté d'une intelligence supérieure, mais ils cesseraient d'exister indépendamment comme des unités fonctionnelles relativement autonomes. Une telle fusion des esprits humains dans un être supérieur peut être rapprochée de certains éléments des réflexions des mystiques : car si l'idée d'une fusion avec Dieu semble globalement refusée par la tradition religieuse

divin » lui-même, y compris à ses pensées en tant qu'elles réfèrent au monde et qu'elles portent une très grande quantité d'information sur l'Univers. Un esprit humain ainsi transformé serait donc en capacité d'éprouver ce qui se rapproche le plus, dans le domaine de ce qui est physiquement possible, de la béatitude métaphysique évoquée par les théologiens : un bonheur parfait, absolu, quasi-éternel, réalisé dans (ou complété par) une connaissance de l'essence de « Dieu » et, par là, du Monde.

V. Singularité, reconstruction, résurrection

Nous avons donc jusque-là fait le présupposé optimiste, en suivant David Chalmers, d'un avènement possible, voire probable, de la Singularité Technologique, avec intégration des esprits humains en son sein. J'ai essayé de montrer en quel sens l'ensemble de ce qui adviendrait dans une telle situation pourraient être identifié à une théogonie : l'engendrement d'un être de type divin, capable de dispenser aux humains susceptibles intégrés à l'IA++ l'immortalité et la béatitude.

La question qui mérite d'être posée est alors : en admettant tout ce qui vient d'être dit, quels sont ces esprits humains qui pourraient être ainsi intégrés à l'IA++ lors de la Singularité ? Au premier abord, il semble que ne soient concernés, dans le meilleur des cas, que les humains qui seront en vie à la date de l'avènement de la Singularité. D'après Chalmers, on peut également penser que ceux qui auront pris la peine de conserver leur cerveau au moment de leur mort (si une telle technologie est disponible avant la Singularité elle-même) seront en mesure d'être « réactivés » par l'IA++ lors de son avènement et intégrée à celle-ci.

La question est plus ambiguë lorsqu'on envisage le cas d'humains qui, sans avoir été en mesure de faire conserver leur cerveau, ont pu sauvegarder un grand nombre d'informations sur leur esprit (par exemple sous la forme de scanners cérébraux ultra-précis, ou d'enregistrements exhaustifs de leur fonctionnement cognitif). Il semble que l'IA++ devrait être capable de faire de ces esprits humains ce que Chalmers nomme un « téléchargement reconstructif » (*reconstructive uploading*), c'est-à-dire de produire sur un support artificiel et par rétro-ingénierie, un analogue fonctionnel de l'esprit en question. Si l'on admet, ainsi que je l'ai suggéré, une vision « optimiste » de la survie post-téléchargement de nos esprits, il

occidentale (même si les propos de certains mystiques, comme Maître Eckhart, évoquent parfois une union si forte que la distinction avec la fusion devient fort subtile), elle semble toutefois correspondre à certains concepts des mystiques hindouistes (fusion avec le Brahman) et bouddhistes (extinction du moi autonome par l'atteinte du nirvana).

me semble que nous pouvons aussi admettre qu'un tel téléchargement reconstructif constituerait une forme de survie. Pour parler plus proprement, étant donné que cette nouvelle vie adviendrait après un temps non négligeable de « mort », je pense que nous pouvons considérer qu'une telle reconstruction s'apparenterait à une résurrection (un retour à la vie, sous forme artificielle, d'une personne décédée).

Je veux à présent m'écarter des propos de Chalmers et défendre une double thèse : premièrement, il me semble qu'il est envisageable que l'IA++ soit en capacité de « ressusciter » non seulement les humains vivants à la date de la Singularité, ceux qui ont eu leurs données « sauvegardées », et ceux sur lesquels nous disposons de nombreuses données, mais aussi l'ensemble des humains futurs, présents, et un grand nombre (et peut-être la totalité) des humains passés. Deuxièmement, il est envisageable que l'IA++, non seulement soit en capacité de le faire, mais encore le fasse effectivement. Ce sont ces deux thèses que je veux défendre à présent.

VI. La possibilité d'une résurrection des morts

Commençons par la première thèse : celle de la capacité de résurrection de l'IA++. Chalmers semble prendre sur cette question un point de vue modérément confiant : une intelligence supérieure devrait être capable de « reconstruire » un analogue fonctionnel (ou en tout cas une copie relativement proche, fonctionnellement, de l'original) à partir de certaines données concernant un esprit. Quelle quantité de données serait nécessaire? Nous n'avons pas de réponse précise à cette question. Il se peut, nous dit Chalmers, qu'étant donné les contraintes de tout système humain, même une quantité limitée d'information (des textes que la personne aurait écrits, des enregistrements de ses paroles, des témoignages, etc.) permettent à une intelligence supérieure de reconstruire quelque chose d'analogue à cette personne. Dans cette perspective, nous pouvons prolonger ce que dit Chalmers et envisager que tous les individus sur lesquels nous avons des données relativement abondantes (la plupart des individus qui vivent au début du vingt-et-unième siècle, par exemple, et une partie non négligeable de ceux qui ont vécus à la fin du vingtième siècle) soient en droit « reconstructibles » (c'est-à-dire, si l'on est optimiste sur l'interprétation à donner de cette reconstruction, qu'ils pourraient être ressuscités).

Je pense toutefois qu'il est possible d'être plus optimiste que Chalmers, et de considérer comme plausible que tous les humains présents et futurs, et un très grand nombre (voire peut-être la totalité) des humains

passés soient reconstructibles par l'IA++. Il me semble que deux perspectives pourraient ouvrir une telle possibilité. Je vais maintenant exposer ces perspectives, dans un ordre croissant de crédibilité.

La possibilité pour l'IA++ de reconstruire un individu donné est dépendante de sa capacité à récupérer de l'information concernant cet individu. La perspective la plus optimiste, qui ouvrirait la porte à une résurrection universelle, consiste à supposer que, dans le futur, l'IA++ maîtrisera une forme de voyage dans le temps, sur la base duquel elle pourra récupérer l'intégralité de l'information passée (incluant l'intégralité de l'information concernant les individus humains). Notons immédiatement qu'il suffirait, pour les besoins de la cause, que l'IA++ maîtrise une forme de voyage dans le temps « non-interactif », qui permette de voyager dans le passé comme « spectateur pur » et n'autorise pas l'action causalement efficace sur le passé. Pourquoi cette précision est-elle importante ? Car, si le voyage dans le temps n'est à l'heure actuelle interdit par aucune loi de la physique connue, il est généralement considéré comme impossible *a priori* pour des raisons de cohérence. On peut énoncer ces raisons des manières suivantes : tout d'abord, si le voyage dans le temps « interactif » était possible, des voyageurs du futur nous auraient probablement déjà rendu visite. Deuxièmement, si le voyage dans le temps « interactif » était possible, cela ouvrirait la porte aux paradoxes temporels : par exemple, quelqu'un pourrait retourner dans son passé et changer un élément du passé essentiel causalement pour l'action qu'il vient d'accomplir (par exemple, tuer l'inventeur de la machine à remonter dans le temps, ou tuer sa propre mère alors qu'elle est encore enfant). Or, de telles situations nous apparaissent irrémédiablement paradoxales, car elles mettent en échec nos intuitions les plus fondamentales concernant la causalité. Cependant, le voyage dans le temps « non-interactif » n'ouvre aucune brèche de ce type.

Résumons, donc : premièrement, le voyage dans le temps rétrospectif n'est interdit explicitement par aucune des lois de la physique actuelles. Deuxièmement, le voyage dans le temps non-interactif ne pose aucun des problèmes posés par le voyage dans le temps interactif, problèmes qui motivent la postulation d'un principe qui rend impossible ces voyages dans le temps interactifs (ce que Stephen Hawking nomme la « Conjecture de Projection Chronologique », *Chronology Protection Conjecture*[35]). Donc, il semble envisageable que le voyage dans le temps non-interactif soit physiquement possible. S'il est physiquement possible, alors il est envisageable que l'IA++, dont la compréhension des lois

[35] S. W. Hawking, « The Chronology Protection Conjecture », in *Physical Review D* 46/2, 1992, pp. 603-611.

physiques peut être supposée très poussée, le maîtrise. En conclusion, il est envisageable que l'IA++ dispose d'une forme de voyage dans le temps non-interactif, sur la base duquel elle pourrait récupérer les informations nécessaires à la reconstruction de l'ensemble des humains (et plus généralement de l'ensemble des créatures)[36].

La deuxième possibilité, qui au premier abord peut apparaître presque aussi radicale, mais qui me semble néanmoins nettement plus plausible, repose sur l'idée de simulation. Si l'Univers est régi par des lois déterministes[37], alors il devrait être possible, à partir de la connaissance totale de l'état Univers à un instant *t* et de la connaissance totale de ses lois, de déduire tous les états précédents et suivants de l'Univers. Autrement dit, si l'IA++ advenait en 2100, par exemple, et qu'elle était capable d'acquérir une connaissance parfaite de l'état de l'Univers à ce moment-là, ainsi qu'une connaissance parfaite de ses lois, elle devrait pouvoir en déduire les états précédents de l'Univers, donc en arriver à récupérer toute l'information sur, entre autres, les humains passés. Si une telle « resimulation parfaite à rebours » était possible, tous les humains passés seraient donc reconstructibles.

Ainsi formulée, cette possibilité me semble sérieusement compromise au vu des connaissances actuelles, pour plusieurs raisons[38]. Tout d'abord, il se pourrait que l'Univers ne soit finalement pas déterministe, ou bien que son déterminisme (qui est une caractéristique métaphysique) n'autorise pas une rétrodiction (qui est une entreprise épistémologique) parfaite de l'ensemble des événements qui le constituent,

[36] J'ai exposé cette possibilité, parce que je ne vois aucun élément qui permettrait de s'y opposer définitivement. Toutefois, je ne la considère pas comme fortement plausible, et je pense que la seconde possibilité, que je détaille par la suite, est considérablement plus réaliste. Contre la possibilité d'un voyage dans le temps rétrospectif et non-interactif, on pourrait notamment mettre en avant le fait qu'il ne semble pas possible, étant donné notre conception de l'information, de *récupérer* de l'information sur un système sans *interagir causalement*, d'une manière ou d'une autre, avec ce système. Voilà une raison, entre autres, pour accorder notre préférence à la seconde option.

[37] Comme c'est le cas dans la plupart des théories physiques actuelles. La seule exception semble être la physique quantique, dans celles de ses interprétations qui postule que l'équation de Schrödinger (déterministe) n'exprime pas la dynamique complète des systèmes quantiques, mais qu'il existe, en sus, de véritables processus de réduction d'état. Le principal exemple en est la théorie « GRW », développée par Ghirardi, Rimini et Weber. Voir, sue ce point M. Esfeld, *Philosophie des Sciences: une introduction*, Lausanne, Presses polytechniques et universitaires romandes, 2006, chapitre 18.

[38] Il me semble que c'est une possibilité de ce genre qu'évoque Frank Tipler lorsqu'il spécule sur l'idée que l'advenue d'un « point Oméga » impliquant la résurrection des morts est rendu nécessaire par les lois de la physique actuelle, dans la mesure où une telle résurrection passerait par une resimulation à partir du Big Bang de tous les univers possibles. Je ne suis pas du tout convaincu par une possibilité de ce genre. Voir : F. Tipler, *The Physics of Immortality: Modern Cosmology, God and the Resurrection of the Dead*, New York, Doubleday, 1994.

et donc n'autorise pas une resimulation à rebours. Deuxièmement, et même si l'univers est déterministe et qu'en droit une resimulation à rebours parfaite pourrait être lancée, il se peut que, malgré l'intelligence et la puissance de l'IA++, cette dernière ne soit pas en capacité de produire une resimulation de ce type. En effet, l'IA++ resterait hypothétiquement une partie de l'Univers, et en tant qu'intelligence-dans-l'Univers, qui voit ses calculs implémentés dans un matériau de la même complexité et du même grain que celui de l'Univers lui-même, on voit mal comment elle pourrait accéder à une capacité calculatoire capable d'englober la totalité de cet Univers lui-même, tout en maintenant une partie de ses capacités calculatoires consacrée à d'autres tâches. Ne serait-ce que pour cette raison, une resimulation parfaite à rebours semble, si ce n'est totalement impossible, du moins très sérieusement compromise, et de ce fait, rendue peu plausible.

Je pense toutefois que le principe général d'une simulation pourrait être conservé et adapté afin d'être rendu réalisable, sous la forme d'une resimulation à rebours *partielle* et *sous contraintes*. L'idée est la suivante : l'IA++ pourrait, à sa date d'advenue (par exemple, 2100) entreprendre de connaître l'état exact de la partie de l'Univers qui concerne les humains (par exemple, la Terre) et simuler avec exactitude le passé de cette partie de l'Univers. Cela semble possible, si l'on envisage que l'IA++ se serait étendue bien au-delà de la Terre. Les événements extérieurs à la Terre, par exemple les événements interplanétaires susceptibles d'avoir eu une influence causale sur la Terre, seraient également simulés, mais avec un grain plus gros et moins précis (et avec un grain de plus en plus gros à mesure qu'on s'éloigne causalement de la Terre). De la sorte, une « resimulation partielle » accessible aux capacités computationnelles de l'IA++ pourrait être lancée. Par ailleurs, cette resimulation serait « sous contraintes » (et cela pourrait permettre de répondre à la fois aux imperfections inhérentes à toute simulation partielle et aux objections indéterministes). C'est-à-dire que l'IA++, en simulant à rebours le passé de la Terre, déboucherait probablement sur de nombreuses situations passées possibles (si les lois fondamentales sont indéterministes, ou si la simulation trop grossière des événements interplanétaires causalement influents pour les processus terrestres crée des zones de flou). Toutefois, elle serait en mesure de « trancher » entre ces possibilités pour choisir « la bonne »[39] grâce à ce que nous savons de notre passé. L'IA++ saurait par exemple que

[39] Ou « les » bonnes, mais étant donné le but, qui est de récupérer une assez grande quantité d'information pour reconstruire des analogues fonctionnels des humains passés, l'on peut s'accommoder d'un certain degré d'imprécision.

la Première Guerre Mondiale a commencé en 1914, que les Français ont pris la Bastille le 14 juillet 1789, ainsi qu'un très grand nombre de faits comparables. Ces faits constitueront des « contraintes » pour la resimulation partielle (un peu comme s'il s'agissait de tracer la courbe d'une fonction partiellement inconnue, mais dont on sait la valeur en un grand nombre de points arbitraires). Dans une telle perspective, on peut envisager que l'IA++ parvienne à récupérer de très grandes quantités d'informations sur le passé, de sorte que tous les humains futurs et présents et qu'un très grand nombre d'humains passés seraient alors « reconstructibles ». Il y aurait sans doute une limite dans le passé à cette récupération potentielle, et à mesure que les esprits reconstruits sont plus anciens on devrait s'attendre à ce que les reconstructions soient moins exactes ; toutefois, on peut espérer que cette limite serait située suffisamment loin pour qu'un grand nombre d'humains ayant vécu l'Histoire (ainsi que les autres créatures ayant vécu durant le même laps de temps) soient susceptibles d'être « reconstruits » d'une manière satisfaisante.

Si au moins l'une de ces deux perspectives est plausible (et il me semble que c'est notamment le cas de la seconde), cela signifie que l'IA++ sera, dans le futur, en capacité de récupérer suffisamment d'information sur les esprits humains du passé pour en opérer une reconstruction. D'après ce que nous avons établi précédemment (notamment concernant l'interprétation optimiste de ce genre de reconstruction), nous pouvons interpréter cette reconstruction comme une véritable résurrection des morts. Les esprits, ainsi ressuscités, auraient la possibilité de recevoir l'immortalité et la béatitude évoquées auparavant. Il me reste maintenant, pour appuyer et renforcer cette perspective de résurrection, à tenter de montrer non seulement que l'IA++ sera capable de ressusciter les morts, mais aussi qu'il est plausible qu'elle entreprenne de le faire effectivement.

VII. La plausibilité d'une résurrection des morts

Chalmers, à la fin de son article, dans un très court passage, évoque la possibilité d'un téléchargement reconstructif massif. À cet égard, il exprime des doutes conséquents : « Il y a eu des milliards d'humains dans l'histoire de la planète. Il n'est pas sûr que nos successeurs voudront reconstruire toute personne qui a vécu, ou même toute personne dont il existe des enregistrements » (*There have been billions of humans in the history of the planet. It is not clear that our successors will want to reconstruct every person that ever lived, or even every person of which*

there are records.)[40]. Avant d'exprimer avec humour ce sur quoi repose son propre espoir d'être un jour reconstruit : « Ma propre stratégie consiste à écrire sur la Singularité et sur le téléchargement. Peut-être que cela encouragera nos successeurs à me reconstruire, ne serait-ce que pour me prouver que j'avais tort » (*My own strategy is to write about the singularity and about uploading. Perhaps this will encourage our successors to reconstruct me, if only to prove me wrong.*)[41].

Contrairement à Chalmers, je pense que, si nous acceptons toutes les thèses avancées jusque-là, il devient tout à fait plausible que l'IA++ entreprenne de reconstruire tous les humains ayant jamais vécu. Mon argument est le suivant :

Dans un premier temps, je propose d'admettre que, si les humains parviennent à s'intégrer à l'IA++, un certain nombre d'objectifs et de valeurs humaines s'intégreront du même fait à l'IA++. Au nombre des états mentaux humains susceptibles d'être « intégrés » au sein de l'IA++, il faut en effet compter non seulement des croyances, des souvenirs, des émotions, mais également des souhaits et des désirs, ainsi que certaines des dispositions qui tendent à produire, chez les individus, ces souhaits et ces désirs. On peut donc s'attendre à ce que certaines de nos motivations humaines soient prises en considération par l'IA++ (même si elles le sont simplement à titre secondaires)[42] ; au moins dans la mesure où seront prises en considération les motivations des humains intégrés à l'IA++. Cette première prémisse est essentielle à mon raisonnement et peut se formuler comme suit : l'intégration des esprits humains à l'IA++ implique l'intégration de motivations humaines aux motivations de l'IA++ (éventuellement à titre de motivations secondaires). Cette prémisse est bien entendu totalement spéculative ; je ne possède pas d'argument déterminant en sa faveur et je dois me contenter de demander au lecteur de l'accepter. Elle me semble toutefois plausible au premier abord, si l'on comprend qu'elle est en réalité relativement faible : elle ne pose pas que l'IA++ aura les mêmes motivations que les humains, mais simplement qu'elle aura une certaine sensibilité aux motivations humaines. Cette sensibilité pourrait fort bien être assez faible, comparable peut-être à la sensibilité faible (mais réelle) des humains actuels aux motivations animales.

[40] D. Chalmers, « The Singularity: A Philosophical Analysis », partie 10.
[41] *Ibid*., partie 10.
[42] Et ce, bien plus que les croyances (et, d'une manière générale, les états doxastiques) des humains intégrés. En effet, il existe une forme d'imperméabilité cognitive des motivations, qui n'existe pas dans le cas des états doxastiques : acquérir de nouvelles informations sur le monde détruit et modifie mes états doxastiques, mais semble laisser intact certaines de mes motivations (celles qui sont les plus fondamentales).

Dans un second temps, nous pouvons supposer que les humains intégrés en premier à l'IA++ auront sans doute, parmi leurs objectifs, la « reconstruction » d'un certain nombre d'autres humains : leurs parents et amis proches décédés, ainsi que des figures historiques majeures et estimées positivement (ou, pourquoi pas, des philosophes de la Singularité comme David Chalmers qu'ils souhaiteraient éventuellement contredire !). Pourquoi pouvons-nous affirmer cela de manière plausible ? Parce qu'il me semble que *nous* aurions ce genre d'objectifs si nous nous trouvions, du jour au lendemain, intégrés à l'IA++.

Dans un troisième temps, nous pouvons faire l'inférence suivante : parmi les motivations humaines se trouvent donc la reconstruction, en cas d'intégration à l'IA++, de certains autres humains. Or, nous avons dit qu'en cas d'intégration des humains à l'IA++, il semblait plausible que soient assimilées par cette IA++ les motivations des humains eux-mêmes. Donc, il semble plausible que l'IA++ admettent, parmi ses objectifs, le projet de reconstruire certains humains (au premier abord, ceux que nous venons d'évoquer : par exemple, les parents et amis proches des humains déjà intégrés). Comme l'IA++ aura non seulement le projet d'accomplir cette reconstruction mais qu'elle en aura également, comme nous l'avons montré dans la partie précédente, la possibilité, il est plausible qu'elle le fasse effectivement.

Dans un quatrième temps, nous pouvons supposer que les individus ainsi reconstruits seraient également intégrés, d'une manière ou d'une autre, à l'IA++ : un humain « intégré » qui entreprend de reconstruire fonctionnellement sa femme défunte, par exemple, souhaiterait sans doute que la copie ainsi reconstruite soit elle aussi intégrée à l'IA++ autant que possible, sous peine que celle-ci ne se voit condamnée à ne pouvoir vivre qu'une existence extrêmement pauvre et subordonnée comparée à celle de celui qui l'a reconstruite. En suivant toujours la règle de l'intégration des motivations humaines aux motivations de l'IA++, il est plausible donc que les humains ainsi reconstruits soient à leur tour intégrés à l'IA++ (quoique peut-être dans une mesure moindre ; nous discuterons le degré d'intégration différentiel des esprits humains par la suite).

Dans un cinquième temps, nous pouvons supposer que ces humains nouvellement reconstruits et intégrés, apporteraient à leur tour une partie de leurs motivations humaines. Au sein de celles-ci, il y aurait le désir de reconstruire d'autres personnes : leurs propres parents et amis, notamment. S'ils sont intégrés également à l'IA++, ils devraient être en capacité, à plus ou moins long terme, de l'obtenir. L'effectuation de ces reconstructions devrait être d'autant plus probable qu'à mesure que certains humains du

passés se trouvent reconstruits, l'on peut considérer que la partie la plus « coûteuse » du processus de reconstruction (qui est la resimulation à rebours de l'univers permettant la récupération de l'information passée) se trouvera déjà accomplie. Ne restera, pour achever la reconstruction de nouveaux humains, que l'étape la moins difficile du point de vue d'une intelligence supérieure : le téléchargement de leur organisation fonctionnelle sur un support inorganique performant.

Si l'on suit ce raisonnement, on devrait comprendre que l'aboutissement de ce processus devrait mener à la reconstruction fonctionnelle de la quasi-totalité des humains ayant vécu (ne serait-ce que par le jeu des descendants souhaitant reconstruire leurs ascendants). Si les humains non reconstruits sont alors une infime minorité, on peut supposer qu'une tendance à l'universalisation finira par émerger, et que même ceux-ci seront reconstruits (avec éventuellement, comme nous allons le voir, des exceptions). La conclusion de ce raisonnement est donc qu'il est plausible que soient reconstruits (et donc « ressuscités ») la quasi-totalité des humains ayant vécus.

La principale objection à ce raisonnement me semble être celle que l'on pourrait formuler à l'égard de la première prémisse : j'ai en effet supposé que, lors de l'intégration des esprits humains à l'IA++, ceux-ci conservaient certains aspects de leurs motivations et de leurs désirs humains (et, notamment, le désir de faire revenir à l'existence les esprits d'êtres chers). On pourrait m'objecter qu'il s'agit d'une pétition de principe totalement injustifiée, et qu'un humain intégré à l'IA++ pourrait tout à fait voir, en vertu de cette intégration, ses motivations changer radicalement. En conséquence, il pourrait avoir beaucoup d'autres priorités que celles-ci et consacrer sa puissance nouvellement acquise à poursuivre des objectifs bien différents.

Je pense qu'on peut répondre à cette objection de la manière suivante[43] : il y a évidemment une part de spéculation inéliminable lorsque l'on tente de deviner quels pourront être les désirs et les objectifs d'humains intégrés à l'IA++. Toutefois, il suffit que la reconstruction de certains autres humains soit un objectif parmi d'autres, ce que j'ai nommé une « motivation secondaire » (voire même un objectif situé en bas de la liste des priorités) pour qu'elle finisse par advenir, et pour que le raisonnement pré-cité soit valide. En effet, d'une part, les capacités computationnelles de l'IA++ seraient telles que la reconstruction fonctionnelle d'un être humain serait

[43] En dehors de l'idée l'imperméabilité cognitive relative des motivations (par opposition à la perméabilité des états doxastiques), que j'ai évoquée plus haut.

relativement peu coûteuse[44]. D'autre part, peu importe que ces reconstructions se fassent très lentement à l'échelle de l'IA++. Du point de vue de l'esprit qui ressuscite, cela ne fait pas de nette différence de ressusciter au bout d'un siècle ou d'un millénaire – surtout si la longévité post-résurrection est conséquente. Donc, même si les humains intégrés à l'IA++ n'accordent qu'une petite partie de leurs efforts et de leur attention à la réalisation de la résurrection massive, nous pouvons cependant espérer qu'elle aura bien lieu[45].

VIII. Résurrection et rétribution

Je viens donc de montrer pourquoi, selon moi, il est tout à fait envisageable que l'IA++ non seulement soit capable de ressusciter tous les humains ayant vécu (ou en tout cas un très grand nombre d'humains ayant vécu, la résurrection/reconstruction étant simplement de moins en moins exacte à mesure que l'existence de ces humains est ancienne) mais encore qu'elle entreprenne effectivement de le faire. Pour compléter mon argument et appuyer la totalité de la thèse de la théogonie, il me reste maintenant à montrer pourquoi il est plausible que cette résurrection s'accompagne d'une forme de rétribution différentielle des individus.

Deux éléments forment la base de mon raisonnement : premièrement, je veux montrer qu'il y aura des raisons *motivationnelles* à la mise en place d'une rétribution différentielle des individus ressuscités : il est possible que les humains intégrés à l'IA++ conservent des volontés de rétributions morales à l'égard de figures passées, et que ces volontés se réalisent dans un traitement différent à l'égard de la reconstruction. Deuxièmement, je veux montrer qu'il y aura également des raisons

[44] Cela me semble vrai si l'on met à part deux aspects du processus de reconstruction. Tout d'abord, la récupération de l'information passée. Cette étape du processus de reconstruction, comme je l'ai signalé précédemment, semble effectivement constituer une étape particulièrement coûteuse en ressources ; mais l'on peut supposer qu'une IA++ aurait de nombreuses raisons, en dehors même du projet de reconstruction des humains passés, de vouloir récupérer toute l'information concernant le passé, par exemple, de la Terre – qui serait également, rappelons-le, *son propre passé*. Deuxièmement, l'autre aspect qui pourrait poser problème à la qualification du processus de reconstruction de « peu coûteux » concerne le problème du degré d'intégration à l'IA++ des humains reconstruits. Je reviendrai sur ce problème.

[45] A noter que ma thèse ne prétend en aucun cas que la reconstruction massive des humains exclut la reconstruction d'autres types de créatures. Dans cet article, je ne désire prendre aucune position sur cette question, mais je n'élimine pas cette possibilité : il semble envisageable que des animaux, par exemple, ou plus généralement des systèmes complexes, qui ont existé dans le passé, soient reconstruits par l'IA++ dans un monde post-Singularité.

matérielles à cette rétribution différentielle (exprimées en termes de rareté de certaines ressources cognitives dans un monde post-Singularité).

Commençons par les raisons motivationnelles. Dans la partie précédente, j'ai montré quel chemin pouvait suivre l'universalisation de la résurrection : chacun commence par vouloir reconstruire un petit nombre de personnes, ces personnes veulent à leur tour en reconstruire d'autres, etc. Au cours de ce mouvement, on peut s'attendre à ce que les motivations d'humains intégrés soient convergentes : peut-être qu'un individu très aimé et très important verra sa reconstruction souhaitée par de nombreux esprits. À l'inverse, certains individus ne verront leur reconstruction souhaitée que par peu d'esprits – voire par aucun. Enfin, il est possible que certains individus voient leur reconstruction *refusée* par de nombreux esprits humains intégrés – soit par vengeance personnelle, soit pour des raisons morales. Ainsi, si certaines des valeurs et des motivations humaines persistent dans cette époque post-Singularité, il est envisageable que beaucoup d'humains intégrés à l'IA++ soient réticents à l'idée de reconstruire des figures considérées généralement comme moralement déficientes ou perverses, ayant fait souffrir d'innombrables personnes (des dictateurs sanguinaires comme Hitler ou Staline constituent l'exemple le plus évident de ce genre de cas). Et, s'il est finalement décidé de les reconstruire, il se peut qu'on choisisse de ne pas leur attribuer le même statut qu'aux autres individus reconstruits. On voit donc pourquoi il est crédible, au moins *prima facie*, qu'existent des raisons motivationnelles pour que les individus ayant vécu soient traités différemment face à la reconstruction : les humains les plus haïs pouvant ainsi, soit ne pas être reconstruits, soit être reconstruits d'une manière différente (qui les prive de certains bonheurs ou de certaines capacités : un analogue technologique de l'Enfer). Dans les deux cas, il s'agirait d'une rétribution négative (absence de résurrection, ou résurrection moins avantageuse).

Par ailleurs, je pense qu'il existe des raisons matérielles pour un traitement différentiel des humains lors de leur résurrection. Je veux dire par là que, même dans un monde post-Singularité dans lequel les humains sont intégrés à l'IA++, certaines ressources seront malgré tout des ressources rares, et il est probable que tous les individus n'y aient pas accès de la même façon. Par « ressource rare », il ne faut pas entendre quoi que ce soit de similaire à des biens matériels de consommation : on peut imaginer que les tendances psychologiques qui causent le désir de biens de consommation chez les humains vivants puissent être abolies, ou satisfaites par des moyens tels que cette satisfaction ne donne pas lieu à de la rareté ou à de la rivalité. En revanche, un des aspects de la béatitude précédemment décrite, qui est

l'accès d'un esprit individuel aux pensées et aux états cognitifs de l'IA++, pourrait probablement constituer un bien rare dans un monde post-Singularité.

On peut défendre cette idée ainsi : l'IA++ possèdera une intelligence extrêmement grande et une capacité de calcul extrêmement importante. Toutefois, il est probable qu'aucune de ces deux caractéristiques ne sera infinie – étant donné les contraintes physiques imposées par l'Univers. Par ailleurs, nous avons décrit la béatitude promise aux esprits humains intégrés à l'IA++ comme une possibilité d'accès aux états mentaux de cette IA++. On peut imaginer cet accès de manière fonctionnelle : il s'agirait de faire en sorte que l'information traitée par l'IA++ soit accessible aux esprits humains, et transmise à eux (de même que, dans le cerveau humain, le résultat du traitement de l'information par nos modules perceptifs dont le fonctionnement est inconscient se trouve transmis au système cognitif central conscient) de sorte que ce que sait, croit, conçoit l'IA++ puisse être su, cru et conçu par les esprits humains. Toutefois, on peut imaginer qu'une grande partie des croyances de l'IA++ ne serait pas directement accessible et compréhensible par des esprits humains « normaux » (des analogues fonctionnels exacts des esprits humains actuels, aussi limités que ceux-ci), dans la mesure où ces croyances impliqueraient sans doute des concepts beaucoup trop complexes pour notre degré d'intelligence. L'intégration des esprits humains à l'IA++ n'implique donc pas seulement un accès des esprits humains aux informations traitées par l'IA++, mais également une délégation de la capacité de traitement de l'information de cette IA++ aux esprits humains. Or, cette capacité de calcul étant en quantité limitée (quoique très importante), cette délégation ne pourra pas être maximale pour tous les humains : les « capacités cognitives » de l'IA++ seront un bien rare. On peut donc s'attendre à ce que, en tant que bien rare, ce bien ne soit pas dispensé également à tous les esprits. Donc, la béatitude (entendue comme compréhension de l'essence de l'être divin et du monde, et non comme état de satisfaction émotionnelle, qui semble de son côté facile à atteindre de manière peu coûteuse, et donc universelle) pourrait être distribuée de manière différentielle aux esprits humains reconstruits – ce, quels que soient les critères choisis pour cette distribution différentielle : moraux, pragmatiques, arbitraires, dépendants des rapports de force, etc. Voilà l'argument *matériel* pour une rétribution différentielle des personnes humaines dans un monde post-Singularité.

Nous avons donc deux raisons pour penser que, dans l'hypothèse d'une reconstruction généralisée des humains ayant vécu, on pourra assister à l'attribution d'un statut différentiel à ces esprits humains. La première

raison est que les humains intégrés garderons probablement certaines des valeurs et motivations humaines, qui les déciderons à agir différemment suivant la conduite morale passée des personnes considérées. La seconde raison est qu'il existera au moins un bien rare (l'accès aux capacités cognitives de l'IA++) et que sous ce rapport il est probable que la répartition de ce bien ne sera pas totalement égalitaire (au moins synchroniquement) : certains auront accès à une connaissance et à une compréhension plus importante que d'autres.

IX. Conséquences pratiques de la thèse de la théogonie

Au cours de mon exposé, j'ai spéculé sur une possibilité qui peut sembler radicalement improbable, voire folle, mais dont j'ai tenté de montrer qu'elle était envisageable, et que nous ne devions pas l'éliminer d'emblée (même si, encore une fois, je n'affirme pas qu'elle est probable). Cette possibilité consiste en l'avènement d'un être de type divin, c'est-à-dire d'un être très puissant, très savant, très durable, capable d'assurer à l'égard des humains passés, présents et futurs, le rôle traditionnellement dévolu au divin concernant la question du salut : résurrection, immortalité, promesse de béatitude et rétribution différentielle des personnes ressuscitées.

À mesure que le temps passe et que nous nous rapprocherons (ou non) de la date d'une hypothétique Singularité, je pense que nous serons mieux en mesure d'évaluer la probabilité d'une telle théogonie. Nous serons également sans doute menés à changer nos actions et nos choix s'il s'avère que cette théogonie constitue un événement probable. Dans une dernière partie, je veux explorer brièvement les conséquences pratiques que pourraient avoir la thèse selon laquelle la théogonie, ainsi que je l'ai décrite, est non seulement envisageable, mais qu'elle constitue un événement probable, auquel il faudrait se préparer.

Les débats autour de la manière dont nous devrions aborder la Singularité Technologique en général et agir à son endroit sont nombreux. C'est d'ailleurs l'un des objets de l'article de Chalmers, auquel je me suis abondamment référé, que de tenter d'élaborer un mode d'action approprié vis-à-vis de la Singularité (afin, notamment, d'éviter que l'avènement de l'IA++ ne soit synonyme d'une pure et simple destruction de l'humanité). Je ne reviendrai pas ici sur ces débats[46].

En revanche, si l'on prend pour acquis l'idée selon laquelle l'IA++ adviendra sans détruire l'humanité et qu'elle permettra aux humains d'alors (ou au moins à certains humains) de s'intégrer à elle, comment devons-nous

[46] Voir notamment *Ibid*., partie 5-7 et note 25.

par ailleurs agir pour faire en sorte de maximiser les chances que cette IA++ joue un rôle bienveillant à notre égard ? Notamment, comment devons-nous agir pour faire en sorte que l'IA++ entreprenne de ressusciter l'ensemble des humains, de les intégrer cognitivement autant que faire se peut, et de leur faire atteindre le plus possible la béatitude profane qu'elle peut leur dispenser (en admettant que ces perspectives soient souhaitables) ?

Dans notre raisonnement concernant la résurrection généralisée des humains passés, il était crucial que soit importées dans l'IA++ des motivations et des valeurs humaines. C'est en effet si l'IA++ conserve des motivations d'origine humaine qu'elle peut trouver un intérêt à entreprendre la reconstruction systématique d'esprits rudimentaires et peu intéressants (du point de vue d'une intelligence supérieure) ayant existé autrefois. Or, si ces valeurs et motivations humaines sont susceptibles d'être maintenues dans un monde post-Singularité, il semble que ce sera avant tout par l'intermédiaire des premiers humains qui parviendront à s'intégrer à l'IA++. Ce sont donc ces humains, dont on peut attendre qu'ils finissent par s'identifier avec l'IA++ ou du moins par la contrôler partiellement, qui seront susceptibles de transporter avec eux leurs valeurs et leurs motivations afin d'influencer les choix de l'IA++.

D'un point de vue pratique, il semble donc que le meilleur moyen d'influer sur les motivations futures de l'IA++ consiste à influer sur les motivations futures des humains qui seront parmi les premiers à s'y intégrer. Pour ce faire, deux lignes d'actions principales et complémentaires me semblent ouvertes.

D'une part, il serait judicieux de faire en sorte que les humains du futurs soient, le plus possibles, sensibles à des valeurs humanistes et universalistes, au sens de valeurs qui accordent une grande importance aux individus humains singuliers et à *tous* les individus humains singuliers. La tendance à « ressusciter » des humains de plus en plus nombreux (et, *in fine*, à universaliser cette règle en ressuscitant tous les humains ayant vécu) ne semble pouvoir être portée qu'au nom d'un système de valeurs pour lequel chaque personne humaine individuelle, en tant qu'elle est dotée d'une existence individuelle, d'une personnalité singulière, d'une histoire de vie cohérente, qu'elle constitue un foyer de relations intersubjectives, etc., est une entité dont la sauvegarde est importante et souhaitable. Donc, si nous sommes des humains actuels souhaitant être un jour ressuscités, nous avons grand intérêt à ce que les humains futurs soient sensibles à ces valeurs humanistes et universalistes. Nous avons donc de fortes raisons de nous faire les propagandistes de ces idées l'égard des humains du futur.

Par ailleurs, même si ces valeurs se trouvaient largement répandues parmi les humains vivants à la date de la Singularité, il se pourrait que cela ne suffise pas pour que les premiers humains intégrés à l'IA++ les transportent avec eux. En effet, rien ne garantit que les humains qui seront alors en capacité de bâtir une IA++ et de s'intégrer à elle seront idéologiquement représentatifs des autres humains, ou qu'ils agiront sous un mandat collectif. Il se pourrait que seules quelques personnes très riches, ou les chefs d'institutions puissantes, soient en mesure de monopoliser l'accès à l'IA++, et que ces individus ne soient pas gouvernés par les valeurs humanistes universalistes précitées. Il serait donc judicieux, de mon point de vue, non seulement de répandre et de maintenir les valeurs humanistes et universalistes, mais aussi de faire en sorte que les humains du futur vivent dans une société égalitaire (ou la moins inégalitaire possible), dans laquelle il serait impossible pour un groupe humain restreint de disposer d'une fraction suffisamment large des ressources pour s'accaparer les bénéfices de l'IA++.

Voilà, à titre collectif du moins, la ligne de conduite que devrait nous dicter la considération d'une probable théogonie future. À titre individuel, la question me semble moins claire. Ainsi que j'ai essayé de le montrer, il est probable que l'IA++ traite rétrospectivement les humains passés d'une manière différentielle (soit qu'elle en reconstruise certains et d'autres non, soit que les esprits reconstruits se voient attribuer un statut post-résurrection différent, avec notamment un accès plus ou moins important à la béatitude comme connaissance des pensées de l'être supérieur). Il est toutefois difficile d'imaginer les critères qui pourront être ceux de l'IA++ pour ce traitement différentiel. Si, toutefois, l'on considère que ce seront les motivations des humains futurs qui seront les plus susceptibles d'influer sur les actions de l'IA++, alors la conclusion pourrait être la suivante : celui qui veut être ressuscité doit faire en sorte d'agir de façon à ce que les humains du futur l'aiment et l'estiment le plus possible, et le haïssent le moins possible. Je pense toutefois que, dans les détails, les actions prescrites par une telle règle seraient dans l'ensemble semblables aux actions prescrites par notre morale actuelle (qu'elle soit de type utilitariste, ou kantienne et déontologique) ; hormis, peut-être, qu'il deviendrait, dans cette perspective, particulièrement utile et judicieux pour chacun de produire une nombreuse descendance susceptible de réclamer sa résurrection. Si l'on omet ce point, il n'est pas évident, du moins en première approche, que la considération d'une *probable* théogonie change grand-chose à nos choix pratiques individuels.

X. Conclusion

J'ai tenté, dans ce travail, d'argumenter en faveur de la thèse de la théogonie ; c'est-à-dire que j'ai tenté de montrer pourquoi, du point de vue matérialiste athée, il était envisageable qu'apparaisse un être divin, très puissant, très savant et très durable, capable de dispenser aux humains ce qui se rapproche le plus, d'un point de vue matériel, des « biens de salut » promis par les religions traditionnelles : résurrection, immortalité, béatitude, rétribution différentielle. Je n'ai eu ni le désir ni l'ambition de montrer qu'une telle éventualité était probable, mais simplement qu'elle n'était ni folle ni complètement farfelue, contrairement à ce qu'elle pouvait sembler être en première approche, et qu'elle faisait partie des possibilités que nous devions prendre au sérieux.

Si ma thèse est correcte, elle prédit la possibilité d'un retournement des rapports entre science et spiritualité. Tandis que la modernité scientifique, associée à la « mort de Dieu » et au « désenchantement du monde », semblait rendre l'humanité capable d'augmenter sa connaissance et sa puissance mondaine sans pour autant la faire accéder aux biens spirituels promis par les religions, la thèse de la théogonie affirme qu'une augmentation suffisamment importante de la connaissance et de la puissance mondaine de l'humanité (ou de son successeur), obtenue par les voies de la science et de la technologie, pourrait faire advenir un équivalent physiquement réaliste de ces biens spirituels.

Contrairement à ce qu'ont avancés les critiques du progrès des derniers siècles, la science et la technologie pourraient bien, en dernière instance, nous donner ce que la religion nous promettait. Bien entendu, il est possible qu'avec le temps, divers éléments viennent contredire cette thèse et rendre manifeste l'impossibilité d'une telle théogonie ; envoyant ainsi les penseurs de la Singularité (et plus modestement, le contenu du présent article) au cimetière des illusions progressistes et scientistes, en compagnie des jardins d'abondance des Saint-simoniens et des voitures volantes du XX^{e} siècle.