Philosophiques Philosophiques

L'argument de la Simulation et le problème de la classe de référence : le point de vue du contextualisme dialectique

Paul Franceschi

Dossier. Usages de la réflexivité en philosophie allemande
Volume 43, numéro 2, Automne 2016

URI : id.erudit.org/iderudit/1038211ar
DOI : 10.7202/1038211ar

Aller au sommaire du numéro

Éditeur(s)

Société de philosophie du Québec

ISSN 0316-2923 (imprimé)
1492-1391 (numérique)

Découvrir la revue

Citer cet article

Paul Franceschi "L'argument de la Simulation et le problème de la classe de référence : le point de vue du contextualisme dialectique." *Philosophiques* 432 (2016): 371–389. DOI : 10.7202/1038211ar

Résumé de l'article

Je présente dans cet article une analyse de l'argument de la Simulation selon le point de vue du contextualisme dialectique, fondée sur le problème de la classe de référence. Je décris tout d'abord l'argument de la Simulation de manière détaillée. J'identifie ensuite la classe de référence et j'applique successivement l'argument à trois classes de référence distinctes : les simulations conscientes de leur propre nature de simulation, les simulations imparfaites et les simulations à immersion. Finalement, je montre qu'il existe trois niveaux de conclusion dans l'argument de la Simulation, selon la classe de référence choisie, qui engendrent des conclusions finales d'une nature très différente.

L'argument de la Simulation et le problème de la classe de référence : le point de vue du contextualisme dialectique

PAUL FRANCESCHI
Université de Corse
p.franceschi@univ-corse.fr

RÉSUMÉ. — Je présente dans cet article une analyse de l'argument de la Simulation selon le point de vue du contextualisme dialectique, fondée sur le problème de la classe de référence. Je décris tout d'abord l'argument de la Simulation de manière détaillée. J'identifie ensuite la classe de référence et j'applique successivement l'argument à trois classes de référence distinctes : les simulations conscientes de leur propre nature de simulation, les simulations imparfaites et les simulations à immersion. Finalement, je montre qu'il existe trois niveaux de conclusion dans l'argument de la Simulation, selon la classe de référence choisie, qui engendrent des conclusions finales d'une nature très différente.

ABSTRACT. — I present in this paper an analysis of the Simulation argument from a dialectical contextualist standpoint. This analysis is grounded on the reference class problem. I begin with describing in detail Bostrom's Simulation Argument. I identify then the reference class within the Simulation argument. I also point out a reference class problem, by applying the argument successively to three different reference classes : aware-simulations, imperfect simulations and immersion-simulations. Finally, I point out that there are three levels of conclusion within the Simulation Argument, depending on the chosen reference class, that yield each final conclusions of a fundamentally different nature.

1. L'argument de la Simulation

Je proposerai dans ce qui suit une analyse de l'*argument de la Simulation*, récemment décrit par Nick Bostrom (2003). Je m'attacherai tout d'abord à décrire en détail l'argument de la Simulation (SA), en mettant notamment l'accent sur la conséquence contraire à l'intuition qui en résulte. Je montrerai ensuite comment une telle conséquence peut être évitée, fondée sur l'analyse de la classe de référence qui sous-tend SA, sans qu'il soit nécessaire de renoncer à ses intuitions pré-théoriques.

L'idée générale qui sous-tend SA peut être ainsi énoncée. Il est très probable que des civilisations post-humaines posséderont une puissance de calcul informatique tout à fait hors de proportion avec celle qui est la nôtre actuellement. Une telle puissance de calcul extraordinaire devrait leur conférer la capacité de réaliser des simulations humaines tout à fait réalistes, telles notamment que les habitants de ces simulations aient une conscience de leur propre existence, en tous points similaire à la nôtre. Dans un tel contexte, on peut penser qu'il est probable que des civilisations post-humaines

consacreront une partie de leurs ressources informatiques à réaliser des simulations des civilisations humaines qui les ont précédées. Dans ce cas, le nombre des humains simulés devrait très largement excéder celui des humains authentiques. Dans de telles conditions, le fait de prendre en compte le simple fait que nous existons conduit à la conclusion qu'il est plus probable que nous fassions partie des humains simulés, plutôt que des humains authentiques.

Bostrom souligne ainsi que l'argument de la Simulation est basé sur les trois hypothèses suivantes :

1) il est très probable que l'humanité n'atteindra pas un stade post-humain.

2) il est très peu probable que les civilisations post-humaines réaliseront des simulations des races humaines qui leur sont antérieures.

3) il est très probable que nous vivions actuellement dans une simulation réalisée par une civilisation post-humaine.

et qu'il s'ensuit que l'une au moins de ces trois hypothèses est vraie.

Pour les besoins de la présente analyse, il s'avère également utile, à ce stade, de décrire la structure dichotomique sous-jacente de SA. La première étape du raisonnement consiste ainsi à considérer, par dichotomie, que : a) ou bien l'humanité n'atteindra pas un stade post-humain ; b) ou bien elle accédera effectivement à un tel stade post-humain. La première de ces deux hypothèses correspond à la disjonction (1) de l'argument. On considère ensuite l'hypothèse selon laquelle l'humanité accédera à un stade post-humain et poursuivra ainsi son existence durant de nombreux millénaires. Dans un tel cas, on peut également considérer qu'il est probable que les civilisations post-humaines posséderont à la fois la technologie et les aptitudes nécessaires pour réaliser des simulations d'humains. Une nouvelle dichotomie se présente alors : a) ou bien ces civilisations post-humaines ne réaliseront pas de telles simulations — il s'agit de la disjonction (2) de l'argument ; b) ou bien ces civilisations post-humaines réaliseront effectivement de telles simulations. Dans ce dernier cas, il s'ensuivra que le nombre d'humains simulés excédera largement celui des humains. La probabilité de vivre dans une simulation sera donc beaucoup plus grande que celle de vivre dans la peau d'un humain ordinaire. Il s'ensuit alors en conclusion que nous autres, habitants de la Terre, vivons probablement dans une simulation réalisée par une civilisation post-humaine. Cette dernière conclusion constitue la disjonction (3) de l'argument. Une étape supplémentaire conduit alors à considérer que l'une des hypothèses (1), (2) et (3) au moins est vraie. La structure dichotomique qui sous-tend SA peut ainsi être décrite étape par étape de la manière suivante :

(4)	ou bien l'humanité n'atteindra pas un stade post-humain, ou bien l'humanité atteindra un stade post-humain ;	dichotomie 1
(1)	l'humanité n'atteindra pas un stade post-humain ;	hypothèse 1.1
(5)	l'humanité atteindra un stade post-humain ;	hypothèse 1.2
(6)	les civilisations post-humaines seront capables de réaliser des simulations d'humains ;	de (5)
(7)	ou bien les civilisations post-humaines ne réaliseront pas de simulations d'humains, ou bien elles en réaliseront ;	dichotomie 2
(2)	les civilisations post-humaines ne réaliseront pas de simulations d'humains ;	hypothèse 2.1
(8)	les civilisations post-humaines réaliseront des simulations d'humains ;	hypothèse 2.2
(9)	la proportion des humains simulés excédera très largement celle des humains ;	de (8)
(3)	il est très probable que nous vivions actuellement dans une simulation réalisée par une civilisation post-humaine ;	de (9)
(10)	l'une des hypothèses (1), (2) et (3) au moins est vraie.	de (1), (2), (3)

Il convient également de mentionner un élément qui résulte de l'interprétation même de l'argument. Car, ainsi que le précise Bostrom (2005), l'argument de la Simulation ne doit pas être mal interprété. Il ne s'agit pas en effet d'un argument qui conduit à la conclusion que (3) est vraie, à savoir que nous vivons actuellement dans une simulation réalisée par une civilisation post-humaine. Le noyau de SA réside ainsi dans le fait que l'une des propositions (1), (2) ou (3) au moins est vraie.

Cette nuance d'interprétation étant mentionnée, l'argument de la Simulation ne manque pas cependant de poser un *problème*. Car SA conduit à la conclusion que l'une des propositions (1), (2) ou (3) au moins est vraie, et que dans la situation d'ignorance où nous nous trouvons, on peut considérer ces dernières comme équiprobables. Ainsi que Bostrom le note : « *In the dark forest of our current ignorance, it seems sensible to apportion one's credence roughly evenly between (1), (2) and (3)* » (Bostrom, 2003). Cependant, selon notre intuition pré-théorique, la probabilité de (3) est nulle ou au mieux extrêmement proche de 0. Ainsi, la conclusion de l'argument a pour conséquence de faire passer la probabilité que (3) soit vraie, de zéro à une probabilité d'environ 1/3. Ainsi, le problème posé par l'argument de la Simulation est précisément qu'il fait passer — par sa conclusion disjonctive — une probabilité nulle ou quasi-nulle concernant (3) à une probabilité beaucoup plus considérable d'environ 1/3. Car une probabilité de 1/3 pour les propositions (1) et (2) ne possède rien de choquant a priori, mais se

révèle en revanche tout à fait contraire à l'intuition pour ce qui concerne la proposition (3). C'est en ce sens que l'on peut parler du problème posé par l'argument de la Simulation et de la nécessité de rechercher une *solution* à ce dernier.

De manière préliminaire, il convient de s'interroger sur ce qui constitue l'aspect paradoxal de SA. Qu'est-ce en effet qui confère une nature paradoxale à SA ? Car SA se distingue de la classe des paradoxes qui conduisent à une contradiction. Dans les paradoxes comme le Menteur ou le paradoxe sorite, le raisonnement correspondant conduit à une contradiction[1]. Rien de tel cependant ne se manifeste au niveau de SA qui appartient, de ce point de vue, à une classe différente de paradoxes dont fait également partie l'argument de l'Apocalypse et le problème de Hempel. Il s'agit en effet d'une classe de paradoxes dont la conclusion présente une nature contraire à l'intuition, et qui vient se placer en conflit avec l'ensemble de nos croyances. Dans l'argument de l'Apocalypse, ainsi, la conclusion selon laquelle la prise en considération de notre rang au sein de la classe des humains ayant jamais existé a pour effet qu'une apocalypse est beaucoup plus probable qu'on aurait pu l'envisager initialement vient heurter l'ensemble de nos croyances. De même, dans le problème de Hempel, le fait qu'un parapluie bleu confirme l'hypothèse que tous les corbeaux sont noirs se place en conflit avec l'ensemble constitué de nos connaissances. De manière similaire avec SA, ce qui apparaît finalement comme paradoxal, en première analyse, c'est que SA conduit à une probabilité de l'hypothèse selon laquelle nous vivons actuellement dans une simulation créée par des post-humains, qui est supérieure à celle qui résulte de notre intuition pré-théorique.

2. La classe de référence dans l'argument de la Simulation

La conclusion du raisonnement qui sous-tend SA, fondée sur le calcul du ratio futur entre les humains réels et les humains simulés, bien qu'elle se révèle contraire à l'intuition, résulte néanmoins d'un raisonnement qui apparaît a priori valide. Cependant, un tel raisonnement suscite une interrogation, qui se trouve liée à la *classe de référence* qui est inhérente à l'argument lui-même[2]. En effet, il s'avère que SA comporte, de manière indirecte, une classe de référence particulière, qui est celle des *simulations* d'humains.

1. Ainsi, le Menteur est à la fois vrai et faux. Dans le paradoxe sorite, un objet comportant un certain nombre de grains de sable est à la fois un tas et un non-tas. De même, dans le paradoxe de Goodman, une émeraude est à la fois verte et vleue, et donc à la fois verte et bleue après une certaine date. Enfin, dans le paradoxe de la Belle au bois dormant, la probabilité que la pièce soit tombée sur face avant le réveil de la Belle est de 1/2 en vertu d'un mode de raisonnement, et de 1/3 seulement selon un raisonnement alternatif.

2. William Eckhardt (2013, p. 15) considère que — de manière identique à l'argument de l'Apocalypse (Eckhardt 1993, 1997 ; Franceschi, 2009) — le problème inhérent à SA provient de l'usage de la rétro-causalité et du problème lié à la définition de la classe de référence : « *if simulated, are you random among human sims ? hominid sims ? conscious sims ?* ».

Mais qu'est-ce donc qui constitue une simulation ? L'argument original se réfère, de manière implicite, à une classe de référence qui est celle des simulations virtuelles d'humains, d'une très haute qualité et par nature indiscernables des humains authentiques. Toutefois, une certaine ambiguïté s'attache à la notion même de simulationm, et la question se pose de l'applicabilité de SA à d'autres types de simulations d'humains[3]. On peut en effet concevoir des types de simulations quelque peu différents qui, de manière intuitive, entrent également dans le champ de l'argument.

De manière préliminaire, il convient de s'attacher à préciser ici la nature des simulations réalisées par des moyens informatiques, auxquelles se réfère l'argument original. De manière implicite, SA se rapporte en effet à des simulations informatiques réalisées à l'aide d'ordinateurs de type classique composés de puces de silicium. Mais on peut envisager également que les simulations soient réalisées à l'aide d'ordinateurs construits à partir de composants qui utilisent les propriétés de l'ADN et la biologie moléculaire. Des recherches récentes ont en effet montré qu'il était possible de mettre en œuvre des algorithmes performants (Adleman, 1994, 1998) et de réaliser des composants d'ordinateurs (Benenson *et al.* 2001 ; MacDonald *et al.*, 2006) à partir de techniques de bio-calcul qui exploitent en particulier les combinaisons des quatre composants (adénine, cytosine, guanine, thymine) de la molécule d'ADN. Si un tel champ de recherche devait connaître une importante expansion et permettre de réaliser des ordinateurs au moins aussi performants que les ordinateurs de type classique, ce type de bio-ordinateurs pourrait légitimement entrer également dans le champ d'application de SA. Car le fait que les simulations soient réalisées à partir d'ordinateurs de type classique ou biologique[4] ne modifie en rien la portée de l'argument. Dans tous les cas, il en résulte que la proportion d'humains simulés sera beaucoup plus grande que celle d'humains réels, en raison des propriétés de la réalité simulée à l'aide de moyens numériques, car l'ordinateur ne connaît pas les limites physiques qui sont celles de la matière.

À titre préalable, on peut observer également que Bostrom se réfère explicitement à des simulations réalisées à l'aide de moyens informatiques. Cependant, la question se pose de savoir si les humains simulés ne pourraient pas consister en des copies physiques, parfaitement réussies, des

3. Nous laisserons de côté ici la question de savoir si l'on doit prendre en considération un nombre infini d'humains simulés. Tel pourrait être le cas si le niveau ultime de réalité était abstrait. Dans ce cas, la classe de référence pourrait inclure des humains simulés qui s'identifient, par exemple, à des matrices de très grands nombres entiers. Mais Bostrom répond à une telle objection dans sa FAQ (www.simulation-argument.com/faq.html) et indique que, dans ce cas, les calculs ne valent plus (le dénominateur est infini) et le ratio n'est pas défini. Nous laisserons donc de côté cette hypothèse, en concentrant notre argumentation sur ce qui constitue le cœur de SA, c'est-à-dire le cas où le nombre de simulations d'humains est fini.

4. Il en irait de même si les simulations étaient réalisées à partir d'ordinateurs de type quantique.

humains véritables. Dans un tel cas, les simulations[5] pourraient être extrêmement difficiles à discerner. A priori, une telle variation constitue également une version acceptable de SA. Cependant, on peut observer une différence avec l'argument original, qui met également en lumière le choix préférentiel fait par Bostrom des simulations de nature informatique. Il existe en effet dans l'argument original une disproportion très importante entre, d'une part, les humains simulés par des moyens informatiques et, d'autre part, les humains véritables. Cela constitue la prémisse (9) de l'argument : « la proportion des humains simulés excédera très largement celle des humains ». Comme le souligne Bostrom, les premiers seraient alors en nombre beaucoup plus grand que les seconds, en raison de la nature-même des simulations informatiques. C'est cette disproportion qui permet ensuite de conclure : (3) « nous vivons très probablement dans une simulation réalisée par une civilisation post-humaine ». Avec des simulations de nature physique, on n'aurait pas a priori une telle disproportion, et la portée de la conclusion serait quelque peu différente. Supposons ainsi que les post-humains parviennent à réaliser des simulations de nature physique, dont le nombre serait par exemple égal à celui des humains réels. Dans ce cas, la proportion des humains simulés serait de 1/2 (alors qu'elle est proche de 1 dans l'argument original). La prémisse (9) deviendrait alors : « la proportion des humains simulés et des humains réels sera de 1/2 ». Et cela permettrait seulement de conclure (3) « la probabilité que nous soyons des simulations réalisées par une civilisation post-humaine est égale à 1/2 ». On le voit, il en résulterait une version notablement atténuée de SA. La différence avec la version originale de SA est que l'argument de la simulation relatif à des simulations physiques s'applique avec moins de force que l'argument original. Cependant, si les conditions devaient changer et qu'il devait en résulter dans le futur pour les simulations physiques une disproportion de même nature qu'avec les simulations de nature informatique, SA s'appliquerait alors avec toute sa force. En tout état de cause, l'analyse qui suit s'appliquerait alors de manière identique à cette dernière catégorie de simulations.

Ces considérations préliminaires étant posées, nous nous intéresserons tour à tour à différents types de simulations d'humains qui sont susceptibles d'intégrer la classe de référence de SA, et aux conclusions qui en résultent en ce qui concerne l'argument. Car la question même de la définition de la classe de référence pour SA conduit à s'interroger sur l'inclusion ou non dans le champ de SA de plusieurs types de simulations. Cependant, la question de la définition de la classe de référence pour SA apparaît ainsi étroitement liée à la nature de la future taxinomie des êtres et des entités qui peupleront la Terre dans un futur proche ou éloigné. Il ne saurait être ici

5. Je remercie un expert anonyme d'avoir souligné ce point, ainsi que celui relatif aux ordinateurs construits à partir de composants utilisant les propriétés de l'ADN et la biologie moléculaire.

question de prétendre à l'exhaustivité, compte tenu de la nature spéculative d'un tel domaine. Mais on peut s'attacher toutefois à déterminer dans quelle mesure SA peut également s'appliquer à des simulations d'une nature différente de celles évoquées dans l'argument orignal, mais qui présentent une égale légitimité. Nous examinerons ainsi tour à tour : les simulations conscientes, les simulations imparfaites, et les simulations à immersion.

3. Le problème de la classe de référence : le cas des simulations conscientes

À ce stade, on ne peut encore véritablement parler de *problème* de la classe de référence au sein de SA. Pour cela en effet, il convient de montrer que le choix de l'une ou l'autre classe de référence a des conséquences tout à fait différentes en ce qui concerne l'argument, et en particulier que la nature de sa conclusion s'en trouve affectée, c'est-à-dire modifiée de manière fondamentale. Dans ce qui suit, nous nous attacherons désormais à montrer que, selon que l'on choisit l'une ou l'autre classe de référence, des conclusions radicalement différentes s'ensuivent quant à l'argument lui-même et que, par conséquent, il existe bien un *problème de classe de référence* au sein de SA. Nous considérerons pour cela successivement plusieurs classes de référence, en nous attachant à montrer comment des conclusions de nature fondamentalement différente en résultent pour ce qui est de l'argument lui-même.

La version originale de SA met en scène, de manière implicite, des simulations d'humains d'un certain type. Il s'agit de simulations de type virtuel, quasiment indiscernables des humains réels et qui présentent ainsi un degré de sophistication très élevé. Plus encore, il s'agit d'un type de simulations qui n'ont pas conscience qu'elles sont elles-mêmes simulées et qui sont donc persuadées d'être des humains authentiques. Cela résulte implicitement des termes de l'argument lui-même, et en particulier de l'inférence de (9) à (3) qui conduit à conclure que « nous » vivons actuellement dans une simulation indiscernable réalisée par des post-humains. De fait, il s'agit de simulations qui sont en quelque sorte abusées et trompées par les post-humains en ce qui concerne leur identité véritable. Pour les besoins de la présente discussion, nous dénommerons *quasi-humains*$^{-}$ les humains simulés qui n'ont pas conscience qu'ils le sont.

À ce stade, il s'avère que l'on peut également concevoir des simulations indiscernables qui présentent un degré tout à fait identique de sophistication mais qui, à l'inverse, auraient conscience qu'elles sont simulées. Nous appellerons ainsi *quasi-humains*$^{+}$ des humains simulés ayant conscience qu'ils sont eux-mêmes des simulations. De telles simulations sont en tous points identiques aux *quasi-humains*$^{-}$ auxquels SA se réfère de manière implicite, à la seule différence qu'elles sont cette fois clairement conscientes de leur nature intrinsèque de simulation. De manière intuitive, SA s'applique également à ce type de simulation. À priori, on ne possède pas de justification pour écarter un tel type de simulation. Plus encore, plusieurs raisons conduisent à penser que les *quasi-humains*$^{+}$ pourraient être plus nombreux

que les *quasi-humains*⁻. Pour des raisons éthiques (a) tout d'abord, on peut penser que les post-humains pourraient être enclins à préférer les *quasi-humains*⁺ aux *quasi-humains*. Car le fait de conférer une existence aux *quasi-humains*⁻ constitue une tromperie sur leur identité véritable, alors qu'un tel inconvénient est absent lorsqu'il s'agit des *quasi-humains*⁺. Une telle tromperie pourrait raisonnablement être considérée comme non éthique et conduire à une forme ou une autre d'interdiction des *quasi-humains*⁻. Une autre raison (b) milite pour le fait de ne pas écarter, a priori, les simulations d'humains ayant conscience de leur propre nature de simulation. En effet, on peut penser que le niveau d'intelligence acquis par certains *quasi-humains* dans un futur proche pourrait être extrêmement élevé et faire que, dans ce cas, les simulations deviendraient très rapidement conscientes qu'elles sont elles-mêmes des simulations. On peut penser qu'à partir d'un certain degré d'intelligence, et en particulier celui susceptible d'être obtenu par l'humanité dans un futur assez proche (Kurtzweil, 2000, 2005 ; Bostrom, 2006), les *quasi-humains* devraient être à même — au moins beaucoup plus facilement qu'actuellement — de recueillir les preuves qu'ils sont l'objet d'une simulation. Plus encore, le concept même de « simulation non consciente qu'elle est une simulation » pourrait être entaché de contradiction, car il faudrait alors limiter son intelligence et, dès lors, il ne s'agirait plus d'une simulation indiscernable et suffisamment réaliste[6]. Ces deux raisons inclinent à penser que les *quasi-humains*⁺ pourraient bien exister en plus grand nombre que les *quasi-humains*⁻ ou même qu'ils pourraient constituer le seul type de simulation mis en œuvre par les post-humains.

À ce stade, il s'avère nécessaire d'envisager les conséquences de la prise en considération des *quasi-humains*⁺ au sein de la classe de référence des simulations inhérente à SA. Pour cela, considérons tout d'abord la variation de SA (dénommons-la SA*) qui s'applique, de manière exclusive, à la classe des *quasi-humains*⁺. Un tel choix, tout d'abord, n'a pas de conséquence sur la disjonction (1) de SA, qui se réfère à une possible disparition de notre humanité avant qu'elle n'ait atteint le stade post-humain. Cela n'a pas d'effet non plus sur la disjonction (2), selon laquelle les post-humains ne réaliseront pas de *quasi-humains*⁺, c'est-à-dire de simulations conscientes d'êtres humains. En revanche, le choix d'une telle classe de référence a une conséquence directe sur la disjonction (3) de SA. Certes, il s'ensuit, de la même manière que pour l'argument original, la conclusion de premier niveau selon laquelle le nombre des *quasi-humains*⁺ excédera largement le nombre des humains authentiques (la *disproportion*). Cependant, il ne s'ensuit plus désormais la conclusion de second niveau selon laquelle « nous »

6. Il paraît difficile d'écarter ici le cas où les *quasi-humains*- découvrent, au moins de manière fortuite, qu'ils sont des humains simulés, devenant de ce fait à partir de cet instant des *quasi-humains*⁺. Cependant, pour avantager le paradoxe, nous considérerons ici que la notion même de simulation indiscernable n'est pas entachée de contradiction.

sommes actuellement des *quasi-humains*$^{+}$. En effet, une telle conclusion (appelons-la l'*auto-applicabilité*) ne s'applique plus à nous désormais, puisque que nous n'avons pas conscience d'être simulés et sommes tout à fait convaincus d'être des humains authentiques. Ainsi, dans ce contexte particulier, *l'inférence de (9) à (3) ne prévaut plus*. En effet, ce qui constitue la conclusion *inquiétante* de SA ne résulte plus désormais de l'étape (9), puisque nous ne pouvons nous identifier aux quasi-humains^{+}, ces derniers ayant clairement conscience qu'ils évoluent dans une simulation. Ainsi, à la différence de la version originale de SA basée sur la classe de référence qui associe les humains aux *quasi-humains*$^{-}$, cette nouvelle version associant les humains et les quasi-humains^{+}, n'est pas associée à une telle conclusion inquiétante. La conclusion qui s'ensuit désormais, on le voit, s'avère tout à fait *rassurante*, et en tout état de cause très différente de celle, profondément *inquiétante*[7], qui résulte de l'argument original.

À ce stade, il apparaît qu'une question se pose : doit-on identifier, dans le contexte de SA, la classe de référence aux *quasi-humains*$^{-}$ ou bien aux *quasi-humains*$^{+}$?[8] Il s'avère qu'aucun élément objectif, dans l'énoncé de SA, ne vient conforter le choix a priori des quasi-humains^{-} ou des quasi-humains^{+}. Ainsi, toute version de l'argument qui comporte le choix préférentiel des *quasi-humains*$^{-}$ ou bien des *quasi-humains*$^{+}$ apparaît comme comportant un *biais*. Tel est ainsi le cas pour la version originale de SA, qui comporte ainsi un biais en faveur des *quasi-humains*$^{-}$, qui résulte du choix par Bostrom d'une classe des simulations qui s'assimile exclusivement à des *quasi-humains*$^{-}$, c'est-à-dire à des simulations non conscientes de leur nature de simulation et qui sont par conséquent abusées et trompées par les post-humains sur la nature même de leur identité. Et tel est également le cas pour SA* la version alternative de SA qui vient d'être décrite, qui comporte un

7. Bostrom (2003) considère que le fait d'attendre que nous vivions dans une simulation n'affecterait notre vie quotidienne que de manière modérée : « *Supposing we live in a simulation, what are the implications for us humans? The foregoing remarks notwithstanding, the implications are not all that radical* ». On peut penser toutefois que l'effet devrait en être beaucoup plus profond, compte tenu du fait que le niveau fondamental de la réalité ne se situe pas où le croient les sujets de la simulation et que par conséquent, un grand nombre de leurs croyances sont tout à fait erronées. Ainsi que le souligne David Chalmers (2005) : « *The brain is massively deluded, it seems. It has all sorts of false beliefs about the world. It believes that it has a body, but it has no body. It believes that it is walking outside in the sunlight, but in fact it is inside a dark lab. It believes it is one place, when in fact it may be somewhere quite different* ».

8. Pour les besoins de la présente discussion, nous présentons les choses sous la forme d'une alternative entre les *quasi-humains*– et les *quasi-humains*+. Cependant, on pourrait concevoir que les post-humains — peut-être des civilisations post-humaines différentes — créent à la fois des *quasi-humains*– et les *quasi-humains*+. On aurait alors une situation tripartie comportant des humains, des *quasi-humains*– et les *quasi-humains*+. Dans un souci de simplification, nous pouvons assimiler ici une telle situation à celle qui prévaut lorsque les post-humains créent seulement des *quasi-humains*–, puisqu'il suffit que ces derniers soient présents en très grand nombre pour créer l'effet inquiétant propre à SA.

biais particulier en faveur des *quasi-humains*$^{+}$, des simulations conscientes de leur propre nature de simulation. Cependant, le choix de la classe de référence se révèle ici fondamental, car il comporte une conséquence essentielle : si l'on choisit une classe de référence qui associe les simulations aux quasi-humains^{-}, il en résulte la conclusion *inquiétante* que nous vivons actuellement très probablement dans une simulation. En revanche, si l'on choisit une classe de référence qui identifie les simulations aux quasi-humains^{+}, il s'ensuit un scénario qui, de manière *rassurante*, ne comporte pas une telle conclusion. À ce stade, il apparaît bien que le choix des *quasi-humains*$^{-}$, c'est-à-dire des simulations non conscientes, dans la version originale de SA, au détriment des simulations conscientes, constitue un choix arbitraire. En effet, qu'est-ce qui permet de préférer le choix des *quasi-humains*$^{-}$, par rapport aux *quasi-humains*$^{+}$? Une telle justification fait défaut dans le contexte de l'argument. À ce stade, il s'avère que l'argument original de SA comporte un biais qui conduit au choix préférentiel des *quasi-humains*$^{-}$, et à la conclusion alarmante qui lui est associée[9].

4. Le problème de la classe de référence : le cas des simulations imparfaites

Le problème de la classe de référence dans SA se rapporte, ainsi que cela a été mentionné plus haut, à la nature même et au type des simulations mises en œuvre dans l'argument. Ce problème se limite-t-il au choix préférentiel, en ce qui concerne l'argument original, des simulations non conscientes, au détriment du choix alternatif des simulations conscientes, qui correspondent à des simulations très sophistiquées d'humains, capables de créer l'illusion, mais dotées de la conscience qu'elles sont elles-mêmes des simulations ? Il apparaît que non. En effet, comme cela a été évoqué plus haut, on peut également concevoir d'autres types de simulations pour lesquelles l'argument fonctionne également, mais qui se révèlent d'une nature quelque peu différente. En particulier, on peut concevoir que les post-humains conçoivent et implémentent des simulations identiques à celles de l'argument original, mais qui ne présentent toutefois pas un caractère aussi parfait. Une telle situation présente un caractère tout à fait vraisemblable et ne présente pas les inconvénients d'ordre éthique qui pourraient accompagner les simulations indiscernables mises en scène dans l'argument original. Le choix de réaliser ce type de simulations pourrait résulter du niveau technologique nécessaire, ou bien de choix délibérés et pragmatiques, destinés à faire économiser du temps et des ressources. Il pourrait s'agir par exemple de simulations d'excellente qualité telles que les habitants scientifiques des simulations ne pourraient en découvrir la nature artificielle qu'après, par exemple, dix années de recherche. De telles simulations pourraient être réa-

9. Ce type de biais s'analyse en une instance du *biais d'uni-polarisation* (Walton, 1999, p. 76-81 ; Franceschi, 2014, p. 587-592) où la classe de référence est celle des simulations et la dualité associée est conscience/inconscience.

lisées en très grand nombre et compte tenu de leur nature moins coûteuse en ressources, pourraient se présenter en plus grand nombre encore que les *quasi-humains*⁻. Pour les besoins de la présente discussion, nous appellerons *simulations imparfaites* cette catégorie de simulations.

À ce stade, on peut se demander quelles sont les conséquences sur SA de la prise en considération d'une classe de référence qui s'assimile à des *simulations imparfaites* ? Dans ce cas, il s'ensuit bien, de manière identique à l'argument original, la conclusion de premier niveau selon laquelle le nombre de *simulations imparfaites* excédera largement le nombre des humains authentiques (la *disproportion*). Mais là aussi, il ne s'ensuit plus désormais la conclusion de second niveau selon laquelle « nous » sommes actuellement des *simulations imparfaites* (l'*auto-applicabilité*). Cette dernière ne s'applique plus à nous désormais, et une conclusion de nature rassurante s'y substitue puisque nous avons clairement conscience de ne pas être de telles simulations imparfaites. Finalement, il s'avère que la conclusion qui résulte de la prise en considération de la classe des *simulations imparfaites* est de même nature que celle qui s'ensuit lorsqu'on considère la classe des *quasi-humains*⁺.

5. Le problème de la classe de référence : le cas des simulations à immersion

Ainsi que nous l'avons vu, le fait d'étendre la classe de référence de SA aux simulations conscientes conduit à une conclusion d'une nature différente de celle qui résulte de l'argument original. Il en va de même pour une autre catégorie de simulations — les simulations imparfaites — qui entraînent une conclusion de même nature que les simulations conscientes, et qui s'avère, en tout état de cause, différente de celle qui résulte de la prise en considération des simulations mentionnées dans l'argument original. À ce stade, la question se pose de savoir si on ne peut pas assimiler la classe de référence à d'autres types de simulations pertinentes du point de vue de SA et dont la prise en considération engendrerait une conclusion par essence différente de celle qui s'ensuit lorsqu'on considère les simulations de l'argument original, ou encore des simulations conscientes ou imparfaites.

En particulier, la question se pose de savoir s'il ne pourrait pas exister dans un futur plus ou moins proche des simulations humaines qui soient telles qu'elles s'appliqueraient à nous-mêmes — dans un sens éventuellement différent de l'argument original — et qui seraient notamment telles que la conclusion d'*auto-applicabilité* inhérente à SA s'ensuive alors. Des éléments de réponse peuvent nous être fournis en considérant une évolution des concepts de réalité virtuelle qui sont d'ores et déjà mis en œuvre dans différents domaines tels que la psychiatrie, la chirurgie, l'industrie, l'entraînement militaire, le divertissement, etc. En psychiatrie notamment, des univers virtuels sont utilisés afin de mettre en œuvre des techniques relevant des thérapies comportementales, et présentent des avantages par rapport à la mise en situation classique in vivo (Powers & Emmelkamp, 2008). Dans ce

type de traitement, le sujet lui-même est simulé à l'aide d'un avatar, et l'univers dans lequel il évolue fait également l'objet d'une simulation de la manière la plus réaliste possible. Des résultats probants ont ainsi été obtenus dans le traitement de certaines phobies (Choy *et al.*, 2007, Parsons & Rizzo, 2008), ainsi que du trouble de stress post-traumatique (Cukor *et al.*, 2009 ; Baños *et al.*, 2011).

Dans ce contexte, on peut penser que des évolutions de ce concept de réalité virtuelle pourraient donner lieu à la réalisation d'humains simulés, qui exigeraient un haut degré de réalisme. Cela nécessiterait notamment que soient menées à leur terme les recherches actuelles portant en particulier sur la simulation du cerveau humain. Il se pourrait ainsi que des avancées significatives soient obtenues dans un futur plus ou moins proche (Moravec, 1998 ; Kurzweil, 2005 ; Sandberg et Bostrom, 2008 ; De Garis *et al.*, 2010). On peut concevoir également que nous disposions alors de la faculté de nous immerger dans des univers simulés en empruntant les personnalités d'humains ainsi simulés, tout en ayant véritablement — le temps de l'immersion — l'impression qu'il s'agit de notre existence réelle[10]. En outre, une même simulation humaine pourrait se présenter sous la forme de multiples variations qui correspondraient à la finalité — thérapeutique, scientifique, ludique, utilitaire, historique, etc. — recherchée lors de l'immersion. On peut imaginer par exemple que certaines variations pourraient ne comporter que les éléments importants de la vie de la personnalité simulée, en négligeant les détails inintéressants. Pour les besoins de la présente discussion, nous pouvons dénommer ce type de simulations : *simulations à immersion*. Dans ce contexte, les humains pourraient ainsi recourir fréquemment à l'immersion dans une personnalité humaine antérieure simulée. On peut en outre envisager que les individus utilisent des simulations de leur propre personne : il pourrait s'agir ainsi de simulations d'eux-mêmes à des époques antérieures de leur vie, avec toutefois de légères variations en fonction de la finalité recherchée lors de l'immersion considérée. Dans de telles circonstances, on peut concevoir que des quantités très grandes de ce type de simulations soient réalisées par des moyens informatiques. En tout état de cause, il apparaît que les simulations à notre disposition seraient en nombre beaucoup plus important que les habitants de notre planète. Dans ce contexte, il s'avère que de manière identique à l'argument original, SA fonctionne si l'on raisonne par rapport à une classe de référence qui s'identifie à ce type de *simulations à immersion*.

10. Une simulation complète d'un cerveau humain est également dénommée *upload*. Une définition (Sandberg & Bostrom, 2008, p. 7) en est la suivante : « *The basic idea is to take a particular brain, scan its structure in detail, and construct a software model of it that is so faithful to the original that, when run on appropriate hardware, it will behave in essentially the same way as the original brain* ».

À ce stade, il convient de se demander quel serait l'effet sur SA de l'assimilation de la classe de référence aux *simulations à immersion*. Dans un tel contexte, il s'avère que la conséquence de premier niveau fondée sur la *disproportion* humains/simulations s'appliquerait ici, de manière identique à l'argument original. En second lieu, et c'est là une conséquence importante, la conclusion de second niveau fondée sur l'*auto-applicabilité* s'appliquerait désormais, puisque nous pouvons en conclure que « nous » sommes également, dans ce sens étendu, des simulations. En revanche, il ne s'ensuivrait plus la conclusion alarmante, qui est celle de l'argument original et qui se manifeste à un troisième niveau, que nous sommes des simulations non conscientes, puisque le fait que nous soyons en ce sens des simulations n'implique pas ici que nous soyons trompés sur notre identité première. Ainsi s'ensuit-il finalement, à la différence de l'argument original, une conclusion *rassurante* : les humains sont occasionnellement des *simulations à immersion*, tout en étant conscients qu'ils les utilisent.

Ne pourrait-on objecter ici que nous n'avons pas encore atteint l'état où nous pouvons nous identifier, ne serait-ce que de manière temporaire, à de telles *simulations à immersion* et que cela ne rend pas pertinents les développements qui précèdent vis-à-vis de SA ? Au sens strict, la réalité virtuelle mise en œuvre à notre époque peut en effet être considérée comme étant d'une nature trop grossière pour être assimilée aux simulations très réalistes évoquées par Bostrom. Cependant, on peut penser qu'il suffirait que des *simulations à immersion* d'excellente qualité, qui seraient de nature à donner l'illusion au moins le temps de leur utilisation qu'il s'agit d'une existence réelle, puissent être réalisées, pour que de telles simulations deviennent pertinentes pour la classe de référence de SA. L'hypothèse qu'un tel palier technologique, fondé sur une explosion de l'intelligence artificielle, puisse être atteint d'ici quelques dizaines d'années, a ainsi été avancée (Kurzweil, 2005 ; Eden *et al.*, 2013). Si une telle évolution technologique devait se produire d'ici, par exemple, quelques dizaines d'années, ne pourrions-nous pas alors considérer légitimement que telles simulations entrent également dans la classe de référence de SA ? Compte tenu de cette possible proximité temporelle, il apparaît ainsi opportun de s'intéresser au cas des *simulations à immersion* et d'en évaluer les conséquences vis-à-vis de SA[11].

11. Ce qui précède montre également qu'en examinant SA avec attention, on constate que l'argument recèle une seconde classe de référence. Cette seconde classe de référence est celle des *post-humains*. Qu'est-ce donc qu'un post-humain ? Doit-on assimiler cette classe aux civilisations très largement supérieures à la nôtre, à celles qui évolueront au XXV^e^ siècle ou bien au XLIII^e^ siècle ? Les descendants de notre actuelle race humaine qui vivront au XXII^e^ siècle devraient-ils être comptés parmi les post-humains s'ils devaient accomplir des progrès technologiques considérables dans le domaine des simulations ? En tout état de cause, la définition de la classe des post-humains apparaît étroitement liée à celle des simulations. Car si l'on s'intéresse, dans un sens étendu, à des *simulations à immersion*, alors les post-humains peuvent être assimilés à une génération d'humains pas très éloignée de nous. Si l'on considère des *simula-*

6. Les différents niveaux de conclusion selon la classe de référence choisie

Finalement, la discussion qui précède met l'accent sur le fait que si on considère SA à la lumière du problème de la classe de référence qui lui est inhérente, il existe en réalité plusieurs niveaux dans la conclusion de SA : (C1) la disproportion ; (C2) l'auto-applicabilité ; (C3) la non-conscience (le fait inquiétant que nous soyons trompés, dupés sur notre identité première). En fait, la discussion précédente montre que (C1) est vraie quelle que soit la classe de référence choisie (par restriction ou par extension) : les quasi-humains^{-}, les quasi-humains^{+}, les simulations imparfaites et les simulations à immersion. En outre, (C2) est également vraie pour la classe de référence originale des quasi-humains^{-} et pour celle des simulations à immersion, mais se révèle toutefois fausse pour la classe des quasi-humains^{+} et celle des simulations imparfaites. Enfin, (C3) est vraie pour la classe de référence originale des quasi-humains^{-}, mais se révèle fausse pour les quasi-humains^{+}, les simulations imparfaites et les simulations à immersion. Ces trois niveaux de conclusion sont représentés sur le tableau ci-dessous :

Niveau	*Conclusion*	*Cas*	*Quasi-humains*$^{-}$	*Quasi-humains*$^{+}$	*Simulations imparfaites*	*Simulations à immersion*
C1	la proportion des humains simulés excédera largement celle des humains (*disproportion*)	C1A	vrai	vrai	vrai	vrai
	la proportion des humains simulés n'excédera pas largement celle des humains	C1Ā	faux	faux	faux	faux

tions imparfaites, il convient alors de les associer à une époque plus lointaine. En revanche, si l'on considère, dans un sens plus restrictif, des simulations d'humains complètement indiscernables pour notre humanité actuelle, il convient alors de s'intéresser à des post-humains d'une époque beaucoup plus lointaine. Ainsi, la classe des post-humains apparaît-elle étroitement corrélée à celle des simulations, car le degré d'évolution des simulations s'avère lié au niveau atteint par les civilisations post-humaines qui les mettent en œuvre. Pour cette raison, nous avons limité ici la discussion à la classe de référence des simulations.

C2	nous sommes très probablement des simulations (*auto-applicabilité*)	C2A	vrai	faux	faux	vrai
	nous ne sommes très probablement pas des simulations	C2Ā	faux	vrai	vrai	faux
C3	nous sommes des simulations inconscientes de leur nature de simulation (*non-conscience*)	C3A	vrai	faux	faux	faux
	nous ne sommes pas des simulations inconscientes de leur nature de simulation	C3Ā	faux	vrai	vrai	vrai

Figure 1. *Les différents niveaux de conclusion dans SA*

ainsi que sur l'arborescence suivante :

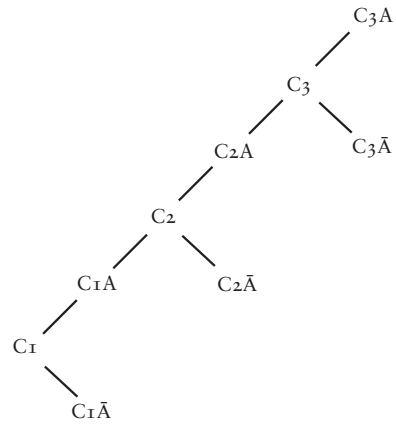

Figure 2. *Arbre des différents niveaux de conclusion de SA*

Alors même que la conclusion originale de SA laisse penser qu'il n'existe qu'un seul niveau de conclusion, il s'avère toutefois, ainsi que cela vient d'être souligné, qu'il existe en réalité plusieurs niveaux de conclusion dans SA, dès lors qu'on examine l'argument selon une perspective plus large, à la lumière du problème de la classe de référence. La conclusion de l'argument original (C3A) est elle-même inquiétante et alarmante, en ce sens qu'elle conclut à une probabilité beaucoup plus forte que nous ne l'avions imaginé a priori, que nous soyons des humains simulés à leur insu. Cependant, l'analyse qui précède montre que, selon la classe de référence choisie, des conclusions de nature très différente peuvent être inférées par l'argument de la simulation. Ainsi, une conclusion de nature tout à fait différente est associée au choix de la classe de référence des *quasi-humains*$^{+}$ ou des *simulations imparfaites*. La conclusion qui en résulte est que nous ne sommes pas de telles simulations (C2$\bar{A}$). Enfin, une autre conclusion possible, elle-même associée au choix de la classe des *simulations à immersion*, est que nous faisons éventuellement partie d'une telle classe de simulation, mais que nous en avons conscience et que cela ne présente donc rien d'inquiétant (C3$\bar{A}$).

L'analyse qui précède met finalement en lumière ce qui pêche dans la version originale de SA, et qui se situe à un double niveau. En premier lieu, l'argument original focalise sur la classe des simulations non conscientes de leur propre nature de simulation. Il s'ensuit la succession de conclusions selon lesquelles il existera une plus grande proportion d'humains simulés que d'humains authentiques (C1A), que nous faisons partie des humains simulés (C2A) et finalement que nous sommes, plus probablement que nous ne l'aurions imaginé a priori, des humains simulés non conscients de l'être (C3A). Cependant, ainsi que cela a été évoqué plus haut, la notion même de simulation d'humains se révèle ambiguë, et une telle classe peut en réalité être définie de différentes manières, compte tenu qu'il n'existe pas, dans SA, de critère objectif permettant de choisir une telle classe d'une manière qui ne soit pas arbitraire. On peut en effet choisir la classe de référence en identifiant les simulations à des simulations *non conscientes*, c'est-à-dire des *quasi-humains*$^{-}$. Mais le choix alternatif d'une classe de référence qui s'identifie à des simulations *conscientes* d'être elles-mêmes des simulations c'est-à-dire des *quasi-humains*$^{+}$, possède une égale légitimité. Dans l'argument original, un critère objectif permettant de choisir la classe de référence, d'une manière qui ne soit pas arbitraire, fait défaut. Ainsi, le fait de privilégier, dans l'argument original, le choix des *quasi-humains*$^{-}$ — avec la conclusion alarmante qui leur est associée — par rapport aux *quasi-humains*$^{+}$, constitue un *biais*, alors même que le choix d'une classe de référence qui s'identifie aux *quasi-humains*$^{+}$, conduit cette fois à une conclusion rassurante.

En second lieu, il apparaît que l'on peut définir la classe de référence de SA à un certain niveau de *restriction* ou d'*extension*. Le choix dans l'ar-

gument original des *quasi-humains*$^{-}$ se situe à un certain niveau de restriction. Mais si l'on se place maintenant à un certain niveau d'extension, la classe de référence inclut désormais les *simulations imparfaites*. Et si l'on se place à un niveau plus grand encore d'extension, les simulations incluent non seulement les simulations imparfaites, mais également les *simulations à immersion*. Mais selon que la classe sera choisie à tel ou tel niveau de restriction ou d'extension, une conclusion tout à fait différente s'ensuivra. Ainsi, le choix, à un plus grand niveau d'extension, des *simulations imparfaites* entraîne une conclusion rassurante. De même, à un niveau d'extension plus grand encore, qui inclut cette fois les *simulations à immersion*, il s'ensuit également une nouvelle conclusion rassurante. Ainsi, l'analyse qui précède montre que dans la version originale de SA le choix se porte de manière préférentielle, par restriction, sur la classe de référence des *quasi-humains*$^{-}$, à laquelle est associée une conclusion inquiétante, alors même qu'un choix par extension prenant également en considération les *simulations imparfaites* ou les *simulations à immersion* conduit à une conclusion rassurante.

Ne peut-on objecter, à ce stade, que l'analyse qui précède conduit à modifier le scénario original de SA et qu'il ne s'agit plus désormais du même problème[12] ? À cela, on peut répondre que l'analyse précédente se fonde sur des variations de SA qui préservent la structure même de l'argument original. Ce que montre la présente analyse, c'est que cette même structure est susceptible de produire des conclusions de nature très différente, dès lors que l'on fait varier la classe de référence dans des limites raisonnables qui correspondent au contexte de SA, et alors même que l'énoncé original de SA laisse supposer un type unique de conclusion. Bostrom lui-même insiste sur le fait que c'est la structure de l'argument qui en constitue le noyau véritable :

> The *structure* of the simulation argument does not depend on the nature of the hypothetical beings that would be created by the technologically mature civilizations. If instead of computer simulations they created enormous numbers of brains in vats connected to a suitable virtual reality simulation, the same effect could in principle be achieved (Bostrom, 2005).

En outre, les différents niveaux d'extension qui sont utilisés ici pour mettre en évidence les variations de la classe de référence de SA sont destinés à illustrer comment différents niveaux de conclusion peuvent en résulter. Mais si l'on souhaite conserver jusqu'à la forme même de l'argument original, on peut alors limiter la variation de la classe de référence à ce qui constitue véritablement le noyau de la présente analyse, en ne considérant qu'une classe de référence qui s'identifie aux quasi-humains. La classe de référence est alors constituée à la fois des quasi-humains^{-} et des quasi-humains^{+}. Car cela suffit pour engendrer une conclusion rassurante — qui n'est pas prise en

12. Je remercie un expert anonyme d'avoir soulevé cette objection.

considération dans l'argument original — et modifier ainsi la conclusion générale qui résulte de l'argument. Dans ce cas, il s'agit de la même classe de référence que celle qui sous-tend l'argument original, à la seule différence que des simulations ayant conscience qu'elles sont simulées en font désormais partie. Car ces dernières, dont l'existence possible n'est pas évoquée dans l'argument original, possèdent cependant un droit égal à la légitimité dans le contexte qui est celui de SA.

Finalement, le choix préférentiel dans l'argument original de la classe des *quasi-humains*⁻ apparaît comme un choix arbitraire que rien ne vient justifier, alors même que d'autres choix possèdent une égale légitimité. Car l'énoncé de SA ne comporte aucun élément objectif permettant d'effectuer le choix de la classe de référence d'une manière non arbitraire. Dans ce contexte, la conclusion inquiétante associée à l'argument original apparaît également comme une conclusion arbitraire, alors même qu'il existe plusieurs autres classes de référence qui possèdent un degré égal de pertinence vis-à-vis de l'argument lui-même, et desquelles découlent une conclusion tout à fait rassurante[13,14].

Références

Adleman, Leonard. « Molecular Computation of Solutions to Combinatorial Problems », Science, vol. 266, 1994, p. 1021-1024.

——. « Computing with DNA », Scientific American, vol. 279(2), 1998, p. 54-61.

Baños, R. M. and V. Guillen, S. Quero, A. García-Palacios, M. Alcaniz, C. Botella. « A Virtual Reality System for the Treatment of Stress-Related Disorders », International Journal of Human-Computer Studies, vol. 69, no. 9, 2011, p. 602-613.

Benenson, Y. T. and Paz-Elizur, R. Adar, E. Keinan, Z. Livneh, E. Shapiro. « Programmable and Autonomous Computing Machine Made of Biomolecules », Nature, vol. 414, 2001, p. 430-434.

Bostrom, Nick. « Are You a Living in a Computer Simulation ? », Philosophical Quarterly, vol. 53, 2003, p. 243-55.

Bostrom, Nick. « Reply to Weatherson », *Philosophical Quarterly*, vol. 55, 2005, p. 90-97.

——. « How Long Before Superintelligence ? », *Linguistic and Philosophical Investigations*, vol. 5, no. 1, 2006, p. 11-30.

Chalmers, David. « The Matrix as Metaphysics », dans C. Grau (dir.), *Philosophers Explore the Matrix*, New York, Oxford University Press, 2005.

13. Le double affaiblissement de SA qui en résulte permet finalement de réconcilier SA avec nos intuitions pré-théoriques, car le scénario inquiétant de l'argument original coexiste désormais avec plusieurs scénarios d'une nature tout à fait rassurante.

14. La présente analyse constitue une application directe à l'argument de la Simulation de la forme de *contextualisme dialectique* décrit dans Franceschi (2014).

Je remercie deux experts anonymes pour *Philosophiques*, pour leurs commentaires très utiles concernant une version précédente de cet article.

Choy, Yujuan and A. Fyer, J. Lipsitz. « Treatment of Specific Phobia in Adults », *Clinical Psychology Review*, vol. 27, no. 3, 2007, p. 266-286.

Cukor, Judith and J. Spitalnick, J. Difede, A. Rizzo, B. O. Rothbaum. « Emerging Treatments for PTSD », *Clinical Psychology Review*, vol. 29, no. 8, 2009, p. 715-726.

De Garis, Hugo and C. Shuo, B. Goertzel, L. Ruiting. « A World Survey of Artificial Brain Projects, Part i: Large-Scale Brain Simulations », Neurocomputing, vol. 74, no. 1-3, 2010, p. 3-29.

Eckhardt, William. « Probability Theory and the Doomsday Argument », Mind, vol. 102, 1993, p. 483-488.

——. « A Shooting-Room View of Doomsday », *Journal of Philosophy*, vol. 94, 1997, p. 244-259.

——. *Paradoxes in Probability Theory*. Dordrecht, New York, Springer, 2013.

Eden, A. and J. Moor, J. Søraker, E. Steinhart (eds.). *Singularity Hypotheses: A Scientific and Philosophical Assessment*, Londres, Springer, 2013.

Franceschi, Paul « A Third Route to the Doomsday Argument », *Journal of Philosophical Research*, vol. 34, 2009, p. 263-278.

Franceschi, Paul « Eléments d'un contextualisme dialectique », dans J. Dutant, D. Fassio & A. Meylan (dir.), *Liber Amicorum Pascal Engel*, Genève, Université de Genève, 2014, p. 581-608.

Kurzweil, Ray. *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*, New York & Londres, Penguin Books, 2000.

——. *The Singularity is Near*, New York, Viking Press, 2005.

MacDonald, J. and Y. Li, M. Sutovic, H. Lederman, K. Pendri, W. Lu, B. L. Andrews, D. Stefanovic, M. N. Stojanovic. « Medium Scale Integration of Molecular Logic Gates in an Automaton », *Nano Letters*, 6, 2006, p. 2598-2603.

Moravec, Hans. « When Will Computer Hardware Match the Human Brain? », *Journal of Evolution and Technology*, 1998, vol. 1.

Parsons, T. D. & A. Rizzo. « Affective Outcomes of Virtual Reality Exposure Therapy for Anxiety and Specific Phobias: A Meta-Analysis », *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 39, no. 3, 2008, p. 250-261.

Powers, M. B. and P. Emmelkamp. « Virtual Reality Exposure Therapy for Anxiety Disorders: A Meta-Analysis », *Journal of Anxiety Disorders*, vol. 22, no. 3, 2008, p. 561-569.

Sandberg, Anders & Nick Bostrom. *Whole Brain Emulation: a Roadmap, Technical Report #2008-3*, Future of Humanity Institute, Oxford University, 2008.

Walton, Douglas. *One-Sided Arguments: A Dialectical Analysis of Bias*, Albany State University of New York Press, 1999.