La conscience artificielle : Une critique pensée et vécue

Michel Bitbol
Archives Husserl, CNRS/ENS, 45, rue d'Ulm, 75005 Paris, France

Chroniques Phénoménologiques, 10, 5-16, 2018

> *« Si la conscience n'est pas indispensable au mécanisme de la vie, il n'en est pas moins vrai que, sans elle, rien n'existerait ».* Rémy de Gourmont

Avant même de se demander si une conscience artificielle est possible, il faut s'immerger dans la part la plus intime et la plus difficile du problème de la conscience. Celle que la plupart des chercheurs en sciences cognitives et en intelligence artificielle s'efforcent de mettre de côté pour faire croire qu'ils sont tout près de résoudre le problème.

La part vraiment difficile du problème de la conscience, c'est ce que le philosophe américain William James et le philosophe japonais Nishida Kitarô ont appelé « l'expérience pure ». Selon Nishida Kitarô, l'expérience pure c'est l'expérience « telle quelle, sans ajouter la moindre discrimination réflexive. (…) (En elle) il n'y a pas encore de sujet ni d'objet ; (en elle) la connaissance et son objet sont en complète unité ». L'expérience pure au sens de Nishida « n'a aucune signification » [1], aucune polarité intentionnelle, contrairement à la conscience élaborée dont parle Husserl. Elle est simplement là, nue, ininterprétée et ininterprétante.

C'est parce qu'ils n'accordent pas assez d'importance à cet ultime niveau d'être, que beaucoup de chercheurs scientifiques et de philosophes de l'esprit restent plongés dans la confusion lorsqu'ils parlent de la conscience. Car, si l'on peut légitimement envisager une explication future par les neurosciences des *fonctions* cognitives attribuées à la conscience, l'expérience pure qui n'a nulle fonction mais une simple présence n'offre pas de prise à leurs méthodes. L'expérience pure actuelle des chercheurs demeure le présupposé silencieux et le point aveugle de toute connaissance

[1] Nishida Kitarô, « Une étude sur le bien », *Revue Philosophique de Louvain.* Quatrième série, Tome 97, N°1, 1999. pp. 19-29

objective. Elle imprègne les choses qui se montrent sans pour autant se montrer ; car "expérience" est le nom de la monstration elle-même.

Réaliser l'expérience pure exige donc d'inverser la direction de la recherche. Au lieu de faire un pas en avant vers les objets de connaissance et de saisie, il s'agit de faire un pas en arrière pour fusionner avec la vie même du connaître et du saisir. Réaliser l'expérience pure exige de ne pas voir les choses d'un endroit particulier, mais simplement de demeurer, de demeurer ici, dans cet élément qui n'a pas d'autre localisation que d'être le centre. Réaliser l'expérience pure signifie annuler la réalisation, dissoudre les intentions, et laisser l'œil pré-visuel grand ouvert. Cette position exceptionnelle de l'expérience pure a été admirablement indiquée par le philosophe allemand Ernst Cassirer, qui l'a attribuée à la conscience entière : « *Il n'y a pas à fixer le lieu de la conscience* […] *mais bien à changer le principe même de l'orientation. Au lieu de nous abandonner au mouvement qui porte la connaissance vers son objet, nous devons apercevoir un but auquel toute la connaissance tourne le dos* »[2]. Or, il existe une discipline qui adopte cette orientation paradoxale de la recherche ; elle a pour nom « phénoménologie », nous y reviendrons.

La quête contemporaine de la « conscience artificielle » apparaît dans ces conditions comme le produit tardif d'une erreur d'*orientation de l'attention* dont les conséquences sont manifestes, et manifestement désastreuses, mais que chaque effort visant à la compenser par un procédé objectif ne fait qu'approfondir. Du cœur du problème de la conscience, on peut dire : « non seulement nous ne voyons pas, mais nous ne voyons pas que nous ne voyons pas »[3]. Pour voir à nouveau, il y a deux méthodes. Il y a la méthode subite consistant à convertir la vision par l'*épochè* phénoménologique ou par l'« éveil » au sens Bouddhiste, qui tous deux consistent à se laisser glisser en arrière dans l'immanence indifférenciée de ce qui ce vit. Et il y a la méthode lente, par le débat d'idées, par l'usure des concepts incomplets, par leur élucidation puis par leur inactivation. C'est cette deuxième méthode, une méthode philosophique, que j'utiliserai ici, tout en sachant qu'elle ne

[2] E. Cassirer, *Philosophie des formes symboliques*, III, Minuit, 1972, p. 67.
[3] H. Maturana & F. Varela, *The Tree of Knowledge,* Shambala, 1987

peut pas suffire, et qu'elle est tout au plus une propédeutique à la conversion.

Comment se manifeste l'oubli fondateur, dans les recherches sur la conscience artificielle ? Qu'est-ce qui sous-tend la sincère conviction de ceux qui en poursuivent le projet ? Considérons par exemple cette déclaration de Stanislas Dehaene[4] : « Bien que la conscience soit souvent considérée comme le pinacle du cerveau, et quelque chose qu'il est impossible de conférer aux machines, j'aimerais me faire l'avocat de la position inverse ». Sa conviction comporte deux étages. Selon le premier étage, la conscience est quelque chose du cerveau, son pinacle, son *produit* le plus élevé ; et selon le second étage, ce qu'on connaît de « la manière dont les cerveaux engendrent la conscience » peut parfaitement être transféré à une machine. Je vais essayer de montrer que ce qui sous-tend cette double conviction, très répandue dans les milieux scientifiques, c'est une série de gestes d'arraisonnement de la part pensable et manipulable de la conscience au prix de la mise à l'écart ou de la négligence de sa part impensable et inaperçue. Impensable parce que *pensante* ; inaperçue parce qu'*apercevante*.

1. Le premier geste d'arraisonnement porte sur la définition, ou la re-définition, de la conscience. On redéfinit la conscience de manière à n'en retenir que les fonctions cognitives objectivables, quitte à admettre *parfois* que cela laisse un résidu dont on ne sait pas quoi dire et quoi faire.

2. Le second geste d'arraisonnement porte sur le vocabulaire des sciences cognitives. On décale subtilement le vocabulaire mentaliste qui nous sert à communiquer sur notre vie vécue en lui conférant une signification purement comportementale ou fonctionnelle. J'ai appelé cette stratégie la « zombification » du vocabulaire des sciences cognitives[5] ; par elle, on prête un nouveau sens *évidé* à des mots comme « esprit » et « conscience ». Puis on signale que des ordinateurs ou des réseaux neuronaux artificiels sont bien pourvus des caractéristiques dénotées par ce lexique appauvri, par cette novlangue cognitiviste. Enfin, dans un dernier et spectaculaire

[4] S. Dehaene, « What is consciousness, and could machines have it ? », <http://www.pas.va/content/accademia/en/publications/scriptavaria/artificial_intelligence/dehaene.html >

[5] M. Bitbol, *La conscience a-t-elle une origine ?* Flammarion, 2014, chapitre 8

tour de passe-passe, on compte sur l'ambiguité des mots de la vie mentale, sur la connotation vécue et éprouvée qu'ils ont malgré tout gardée dans les conversations quotidiennes en dépit de la pénétration du vocabulaire cognitiviste dans la culture, pour persuader tout un chacun que la conscience au sens *plein* du terme a bien été transférée à des artéfacts.

3. Le troisième geste d'arraisonnement s'accomplit à l'intérieur même du domaine objectif. On concentre l'attention des chercheurs sur les seuls corrélats *neuronaux* du traitement de l'information ou des fonctions cognitives ; et on laisse dans l'ombre des constituants biologiques et des processus physiques adjacents sous prétexte qu'ils ne semblent pas pertinents pour ces fonctions. C'est le cas du support glial des neurones, ou bien des champs électromagnétiques globaux à la fois engendrés par l'activité neuro-électrique et capable de l'influencer en retour. Ces processus apparemment marginaux ne pourraient-ils pas être les corrélats objectifs de certains aspects non-fonctionnels de la conscience ? Être incapable de répondre à ce genre de question intra-scientifique, rend le projet de fabriquer une conscience artificielle sur le modèle de celle qu'on qualifie de « naturelle » incertain et aléatoire parce que conduit à l'aveuglette.

4. Le quatrième geste d'arraisonnement est capital pour le projet d'une conscience artificielle, car il conditionne le sens qu'on peut donner à la *reconnaissance* par nous, êtres humains, de la conscience dans un artéfact. Il consiste à simplifier la vie intersubjective en supposant que nous *attribuons* explicitement une conscience à des êtres suffisamment semblables à nous, et que nous élaborons une théorie des autres esprits. Or, la tentative d'inférer la conscience des *alter-ego* sur la seule base des comportements laisse libre cours à un scepticisme radical ; et, de plus, son bien-fondé est contesté par plusieurs approches phénoménologiques de l'intersubjectivité. De même, l'idée que nous entretenons une *théorie* des autres esprits est très discutée à l'heure actuelle dans les sciences cognitives. On penche désormais plutôt vers des procédés infra-théoriques de la constitution d'intersubjectivité : la « simulation » du point de

vue de l'*alter-ego*[6], ou bien la co-émergence des points de vue dans la transaction[7].

5. Enfin, il faut se poser la question du *but* de tous ces arraisonnements. *Pourquoi* cherche-t-on à élaborer des artéfacts dotées de conscience ? Pour rendre ces artéfacts plus efficaces que des robots des premières générations, en leur conférant certaines fonctions d'autonomisation et d'utilisation conjointe de toutes les informations disponibles, que l'on attribue habituellement à la conscience des individus humains ? Pour rendre ces entités substituables à des êtres humains dans les fonctions *sociales* ? Ou pour les rendre substituables à *nous-mêmes* en tant qu'êtres vivant leur vie, en tant qu'êtres pour qui cela fait quelque chose de vivre ; avec pour horizon avoué celui de l'immortalité de la vie individuelle par son transfert dans des artéfacts ?

Je vais développer ces cinq modes d'arraisonnement du problème pré-rationnel de la conscience en les regroupant par blocs. Et j'essaierai de montrer quel est l'impact de chacun d'entre eux pour le problème de la conscience artificielle.

Commençons par le premier mode d'arraisonnement : chercher une définition objective de la conscience ; chercher à faire de la conscience une propriété objective, une propriété *physique*. Chercher cela parce que nous autres, membres tard-venus d'une civilisation occidentale qui déclare avoir vaincu ses origines mythiques, pensons vivre dans un monde fait uniquement d'objets physiques. À partir de là, tout ce qui paraît non-physique ne peut être qu'une propriété ou un sous-produit de quelque objet physique. Mais au nom de quoi pense-t-on cela ? Au nom d'une hypostase tacite des objets de la connaissance scientifique et de la manipulation technologique.

Rappelons-nous que la recherche scientifique consiste à isoler, à partir de l'expérience totale que nous avons du monde apparaissant, des foyers d'attention que nous prenons pour objets d'étude, puis à établir des relations systématiques, nommées des lois de la nature, entre les propriétés et les transformations de ces objets d'étude. Ces lois, parce qu'elles sont attestables par tous, sont exploitables à des fins collectives de transformation technologique de l'environnement.

---

[6] Goldman, A. I. (2006). *Simulating Minds*. Oxford, Oxford University Press.

[7] T. Fuchs & H. De Jaeger, « Enactive intersubjectivity: Participatory sense-making and mutual incorporation », *Phenomenology and the Cognitive Science,* 8, 465-486, 2009

Lorsqu'ils font cela, les chercheurs se contentent d'abord de *présupposer* l'expérience vécue, sans prétendre rendre raison de son origine. Entièrement tendues vers les choses qui sont circonscrites à partir de l'expérience des chercheurs, et vers la possibilité de maîtriser l'usage pratique de ces choses, les sciences se contentent dans un premier temps d'ignorer le milieu transparent à travers lequel elles acquièrent leurs connaissances. N'ayant pas affaire dans ce cas à un objet ou à une propriété d'objet, les sciences jouent à cache-cache avec l'insaisissable *condition préalable* pour que des objets se montrent.

Mais dans un deuxième temps, les chercheurs ont élaboré des stratégies à la fois astucieuses et fécondes pour contourner l'obstacle de principe. Leur stratégie principale, nous l'avons signalé en passant, consiste à *redéfinir* la conscience, de manière à transformer son concept en un objet légitime d'investigation. Les chercheurs redéfinissent si bien la conscience que cela les conduit à en soustraire l'aspect le plus caractéristique, mais aussi le plus inscrutable, et à n'en retenir que ce qu'ils peuvent saisir par leurs méthodes. Ils la redéfinissent par exemple comme une fonction de synthèse des informations, ou comme un processus de « méta-cognition » (un savoir que l'on sait), ou bien encore comme une fonction d'auto-représentation du sujet. Ils se concentrent en somme sur les niveaux les plus élaborés de la conscience que sont la conscience réflexive et la conscience de soi, en oubliant presque entièrement son niveau de base pourtant crucial qu'est *l'expérience pure*.

Un excellent exemple de cette prise de possession définitionnelle est la caractérisation par Dehaene de ce qu'il faut implémenter sur un artéfact pour qu'on puisse le considérer comme conscient. Dans l'un de ses articles[8], il dénombre *quatre* fonctions nécessaires à cet effet : 1) un espace de travail global permettant le partage et la comparaison des informations disponibles (c'est-à-dire un lieu où s'effectue la fonction de synthèse) ; 2) « un répertoire d'auto-connaissance », c'est-à-dire la connaissance de soi, le « connais-toi toi-même », la méta-cognition ; 3) plus

[8] S. Dehaene, « What is consciousness, and could machines have it ? », <http://www.pas.va/content/accademia/en/publications/scriptavaria/artificial_intelligence/dehaene.html >

spécifiquement, la connaissance réflexive sur sa propre connaissance, la capacité d'en évaluer la validité ; 4) une théorie de l'esprit des autres. Dans un article plus récent[9], il ne retient que deux de ces fonctions, en y absorbant les deux autres : 1) Une fonction de synthèse, encore appelée « disponibilité globale de l'information pertinente » ou capacité « d'intégration et de coordination », et 2) une fonction d'auto-surveillance, d'évaluation réflexive et de confiance dans les critères de ses propres choix. Ces fonctions sont censées être réalisées neurobiologiquement, selon Dehaene et son équipe, au sein de l'« espace de travail neuronal global » des régions fronto-pariétales du cortex cérébral. Suivant cette perspective, deux pistes s'ouvrent en vue de produire une conscience artificielle : implémenter, sur un artéfact, des *fonctions cognitives* analogues à celles qui sont attribuées à la conscience humaine ; ou bien s'astreindre, en plus, à utiliser une *architecture* de la transmission et de la distribution d'information analogue à celle du cerveau. Autrement dit, synthétiser, réfléchir, et utiliser l'information de toutes les manières possibles et pragmatiquement efficaces ; ou bien le faire à la manière particulière du réseau neuronal biologique.

La question de savoir laquelle des deux pistes emprunter (imitation purement fonctionnelle, ou imitation de l'architecture cérébrale) se pose d'autant plus que des théories autres que celle de l'espace de travail neuronal global ont été proposées pour rendre raison des mêmes fonctions cognitives. Ainsi, dans la théorie que soutient Giulio Tononi, la conscience est caractérisée comme « information intégrée », c'est-à-dire comme fonction de discrimination des éléments de l'environnement, et comme capacité à *intégrer* le produit des divers actes de discrimination en une seule représentation et un seul schéma de réponse coordonnée. Dans cette théorie, l'intégration est le mot qui vaut pour la fonction de synthèse. Quant à la fonction de méta-cognition, elle est considérée comme le trait saillant, pour ne pas dire unique, de la conscience, dans la théorie des « pensées d'ordre supérieur » réalisées par des boucles récursives neuronales, qui a été proposée par David Rosenthal[10].

[9] S. Dehaene, H. Lau, & S. Kouider, « What is consciousness, and could machines have it ? », *Science,* 358, 486-492, 2017

[10] D. Rosenthal, *Consciousness and Mind*. Oxford: Oxford University Press, 2005.

Cela étant admis, on peut obtenir des machines ayant l'une, l'autre, ou toutes, des trois caractéristiques suivantes qui, *vues de l'extérieur,* rendent plausible de leur attribuer une conscience[11] :

1. Ces machines peuvent avoir un *comportement manifeste* analogue à celui d'êtres humains conscients (c'est le test de Turing, dans une version exigeante qui inclut l'imputation d'une conscience à la machine) ;

2. Ces machines peuvent avoir des *propriétés cognitives* analogues à celle d'êtres humains conscients ; des propriétés qui sous-tendent les comportements mais qui sont plus riches qu'eux, parce qu'elles tiennent en réserve des réponses inédites à des situations inattendues ;

3. Ces machines peuvent avoir en plus, pour réaliser ces propriétés cognitives, une architecture interne analogue à celle du cerveau d'êtres humains conscients, y compris et surtout les éléments qui sont censés être des corrélats voire des substrats de conscience chez les êtres humains. Elles peuvent ainsi être dotées d'un espace de travail global vers lequel convergent les modules spécialisés dans chaque fonction cognitive particulière.

Le problème c'est que réaliser ces trois caractéristiques, tout particulièrement la première d'entre elles, ne garantit pas que les machines aient une quatrième et dernière caractéristique ; une dernière caractéristique qui n'en est justement pas une, parce qu'elle est la condition pour que l'expression « attribuer une caractéristique » ait le sens qu'elle a pour nous. La quatrième pseudo-caractéristique qui semble omise, c'est encore et toujours l'expérience pure, *ce que cela fait d'être* un être (humain ou artéfactuel) ; ou encore le « sentiment » *éprouvé* (« sentience », en anglais) qui accompagne la « sensibilité » *observée* (« sensitivity » en anglais)[12]. Comme l'a timidement concédé Dehaene dans un article récent, « Nous affirmons qu'une machine dotée des fonctions (de synthèse et de métacognition) se comporterait *comme si* elle était

[11] D. Gamez, Progress in machine consciousness, <http://davidgamez.eu/papers/Gamez07_ProgressMachineConsciousness.pdf>

[12] Schilhab T. S. S. « Why did subjective experiences develop ? », *Evolution and Cognition* 4(1): 63–69, 1998. <http://cepa.info/4259>

consciente (…). (Mais) ne laissons-nous pas de côté la composante expérientielle ? »[13].

Une telle lacune des théories de la conscience artificielle ne peut pas leur être reprochée comme une imperfection, car elle est simplement un effet de leur *méthode* qui les conduit à opérer sous l'hypothèse d'une « clôture causale » du domaine physique. Puisque toutes les causes des phénomènes observables (comme les comportements) sont énonçables en termes physiques objectifs, le fait de la subjectivité, le fait de l'expérience vécue, devient simplement superflu, au mieux épiphénoménal. Si ces théories ne sont pas *entièrement* muettes à propos du fait élémentaire que l'information synthétisée et réfléchie *apparaît,* c'est donc à condition d'enfermer ce fait élémentaire de l'apparaître dans un ultime retranchement verbal, ou dans un ultime noeud de quelque diagramme du mode d'opération de l'esprit, sans aucun rôle avéré. Cet ultime retranchement verbal est ce que Ned Block appelle la « conscience phénoménale », opposée à la « conscience d'accès ». La conscience d'accès est définie comme une fonction par laquelle des informations, par exemple sensorielles, deviennent capables d'infléchir des opérations mentales, de se rendre disponibles pour le rapport verbal, et d'être utilisées pour guider des actions. La conscience phénoménale est alors le résidu de cette définition de la conscience d'accès ; elle est ce que la définition n'a pas pu capturer, mais qui est pourtant manifestement *là*.

Comment les théoriciens fonctionnalistes affrontent-ils cette question têtue mais marginalisée de l'expérience vécue, lorsqu'il s'agit d'implémenter une conscience sur des artéfacts ? Ils semblent avoir trois stratégies explicites et une stratégie implicite pour cela.

La première stratégie explicite consiste à prendre la question à bras le corps, en annonçant (non sans imprudence) des techniques permettant de doter les artéfacts de « conscience phénoménale », en plus de la « conscience d'accès » fonctionnelle[14]. Nous en discuterons plus tard.

La deuxième stratégie, aux antipodes de la première, consiste à déclarer qu'il n'y a « en réalité » *rien de tel* qu'une

[13] S. Dehaene, H. Lau, & S. Kouider, « What is consciousness, and could machines have it ? », *loc. cit.*
[14] D. Gamez, Progress in machine consciousness, op. cit.

expérience vécue. Ou du moins que la conscience phénoménale est une simple *illusion* engendrée par la dynamique fonctionnelle de la conscience d'accès, et que, par conséquent, implémenter une conscience d'accès s'accompagnerait automatiquement d'une conscience phénoménale en ce sens fantasmagorique. Plus radicale encore que les célèbres déclarations de Daniel Dennett et Susan Blackmore sur le caractère illusoire de la « phénoménologie apparente », on peut penser ici à une interview donnée par le neurobiologiste Michael Graziano au *New York Times* : À la question « Comment le cerveau va-t-il au-delà du traitement de l'information pour devenir subjectivement conscient de l'information ? », ce chercheur répond : « Il ne le devient pas. … Car il n'y a rien de tel qu'une impression subjective ; *il y a* seulement de l'information dans un dispositif de traitement des données »[15]. Voilà ce qu'on peut appeler un éliminativisme de l'espèce la plus extrême. Mais ce parachèvement de l'éliminativisme laisse songeur : quel genre d'état de conscience a-t-il fallu atteindre pour considérer l'apparaître conscient comme illusoire ? Que doit-on être devenu pour ne pas voir que l'illusion est-elle-même une expérience, ce qui rend vaine l'affirmation que l'expérience est une illusion ? Et qu'a donc fait notre civilisation pour avoir ainsi transformé certains d'entre nous, parmi les plus brillants, en des êtres devenus aveugles au fait de voir ? Ne nous sommes-nous pas conditionnés à oublier notre humanité vécue au profit de nos aspects « machiniques », afin de nous préparer à considérer des machines comme substituts acceptables des êtres humains ?

La troisième stratégie consiste non plus à nier, mais à marginaliser la question de l'expérience consciente, à ne pas nécessairement l'éliminer, mais à la traiter *d'extra-scientifique* (ce qui revient à la traiter comme quantité négligeable dans notre système de valeurs épistémologiques). C'est le cas par exemple dans un article assez récent de Cohen et Dennett : « Loin d'être un obstacle pour la science, le 'problème difficile' (de l'origine de l'expérience vécue) doit sa difficulté apparente au fait qu'il est extérieur à la science (…) car seuls les produits des *fonctions cognitives* (les rapports verbaux, les pressions de boutons etc.) permettent d'étudier empiriquement

[15] Michael S. A. Graziano, "Are We Really Conscious?" *New York Times*, Oct. 10, 2014, https://www.nytimes.com/2014/10/12/opinion/sunday/are-we-really-conscious.html.

la conscience »[16]. L'approche de la conscience artificielle qui va avec cette marginalisation sans négation consiste à s'efforcer de reproduire les seuls comportements et fonctions cognitives attribués à la conscience, et à *supposer* qu'une conscience phénoménale émergera *peut-être* spontanément de là, à la manière (dirait Bergson) d'une « phosphorescence » laissée par les fonctions cognitives dans leur sillage. La locution « peut-être » reste posée à titre de précaution, mais la conviction la plus répandue est qu'il ne peut pas en être autrement, que la pleine implémentation des fonctions de la conscience d'accès ne saurait manquer de laisser affleurer une conscience phénoménale.

Cela nous amène à la dernière stratégie, implicite, pour affronter le problème de l'expérience vécue. Cette stratégie-là peut être vue comme un stratagème rhétorique permettant de cacher le « problème difficile » aux yeux d'un grand public ébahi par les progrès de l'intelligence artificielle. Mais le stratagème finit par intoxiquer ceux qui l'ont inventé, en leur faisant croire qu'ils sont tout près d'arriver à transmuter le plomb de la matière en l'or de la conscience. Quel est donc le stratagème ? Tout simplement, je l'ai signalé antérieurement, utiliser le vocabulaire de l'expérience vécue pour décrire le comportement simili-conscient d'un artéfact. Donner à ce vocabulaire un usage, et donc un sens, nouveau, purement fonctionnel voire comportemental : un usage qui pourrait tout aussi bien convenir à un zombie ou à une machine privée de « sentiment » (de « sentience », en anglais). Je ne trouve pas plus bel exemple de cela, une fois de plus, que l'article déjà cité de Dehaene et son équipe, paru en octobre 2017. Là même où les auteurs ont concédé qu'ils ont peut-être laissé de côté la composante expérientielle de la conscience, ils n'hésitent pas à écrire qu'une machine capable de réaliser les fonctions de synthèse et de métacognition « (…) *saurait* qu'elle *voit* quelque chose, (…) *souffrirait d'hallucinations* lorsque ses mécanismes d'auto-surveillance tomberaient en panne, et pourrait avoir l'*expérience* des *mêmes illusions perceptives* que les êtres humains »[17]. En droit, et selon la réserve qu'ils ont

[16] M. Cohen, D. Dennett, « Consciousness cannot be separated from functions », *Trends in the Cognitive Sciences,* 15, 358-364, 2011

[17] S. Dehaene, H. Lau, & S. Kouider, « What is consciousness, and could machines have it ? », *loc. cit.*

eux-mêmes exprimée, les auteurs auraient dû écrire qu'une telle machine peut se comporter *comme si* elle savait qu'elle voit quelque chose, *comme si* elle souffrait d'hallucinations, *comme si* elle était victime d'illusions perceptives. Mais ils ont retiré le « *comme si* » deux lignes après l'avoir posé, et ils ont même attribué à leur machine une « expérience » deux lignes avant de reconnaître qu'ils ont peut-être négligé l'expérience vécue. Seul l'analyste voit que leur vocabulaire expérientiel est en vérité factice et « zombifié », tandis que le grand public (qui n'entend que les raccourcis verbaux), a toutes les chances de se laisser abuser.

Un autre exemple de lexique évidé, ou « énucléé » de son œil expérientiel, est celui des *émotions,* dont beaucoup de spécialistes de la conscience artificielle déclarent qu'ils peuvent (ou pourront un jour) doter leurs artéfacts. Mais ils le disent, une fois de plus, au nom de leur capacité à implémenter soit un comportement analogue à celui d'un être humain ému, soit une structure cognitive analogue à celle du cerveau d'un être humain éprouvant une émotion. À la faveur d'un tel défléchissement objectif du vocabulaire des affects émotionnels, il semble aller de soi qu'un artéfact (disons un réseau neuronal artificiel) doté d'un mécanisme analogue de propagation de l'information, à la suite de l'activation d'un système fonctionnellement analogue à l'amygdale du cerveau, pourra être dit « capable d'émotions ». Mais quelles émotions ? Des émotions observables et objectivées, ou des émotions éprouvées ? Des émotions apparentes prenant la forme de commotions visibles, ou bien ces soudaines transformations du sens vécu du monde en quoi consistent les émotions dans l'acception phénoménologique qu'en retient Sartre ? En utilisant le vocabulaire de l'expérience dans une acception objective, on ne fait qu'escamoter ce problème sans le résoudre.

Il est vrai que la stratégie d'escamotage a été remarquablement opérante dans le passé des sciences, et c'est ce qui semble la légitimer aux yeux de ses défenseurs. Appeler « chaleur » une quantité mesurable plutôt qu'une qualité perçue a permis l'établissement de la science thermodynamique ; appeler « vie » un processus d'homéostasie physico-chimique plutôt que le fait du « vécu » a permis l'essor de la biologie moléculaire. Mais tenter d'objectiver la qualité même de

l'expérience, le vécu même de l'être vivant, c'est atteindre la limite fondamentale du processus de connaissance objective et se priver de ressources pour reconnaître cette limite.

Cette critique nous conduit à revenir à la première stratégie, la plus audacieuse ; celle qui déclare que donner aux machines le « sentiment » (« sentience ») n'est pas hors de portée. Après tout peut-être pourrons-nous, volontairement ou involontairement, méthodiquement ou à titre d'effet secondaire, fabriquer des artéfacts dotés, *en plus* des fonctions cognitives de la conscience, de sa composante dite « phénoménale » ? Peut-être sera-t-il alors justifié d'employer les mots de l'expérience au sens propre et non pas « zombifié » ? Que cela puisse arriver involontairement ou secondairement est une thèse souvent soutenue par les fonctionnalistes, nous l'avons vu. Selon eux, la conscience phénoménale étant factuellement associée aux fonctions de synthèse des représentations et de méta-cognition remplies par des esprits humains, ou à des processus de traitement de l'information suffisamment complexes, discriminatifs, et intégrés, ces processus doivent non seulement être nécessaires pour qu'une entité manifeste des comportements évoquant la conscience, mais aussi suffisants pour engendrer *tous* les aspects de la conscience, y compris l'aspect « phénoménal ».

Il y a cependant une autre option, qui, appartient également à la famille des conceptions physicalistes. Selon elle, comme l'illustre la figure du Zombie de Chalmers, le simple accomplissement des fonctions informationnelles attribuées à la conscience ne suffit pas nécessairement à engendrer sa composante « phénoménale ». L'engendrement de l'expérience vécue est alors attribué aux propriétés physiques du substrat matériel des fonctions informationnelles : soit à des propriétés physiques déjà connues, soit à des propriétés inédites relevant d'une physique encore inconnue. Cette ligne de recherche a donné lieu à un grand nombre de propositions concernant l'hypothétique « mécanisme physique » de la conscience, y compris phénoménale. La réduction objective du vecteur d'état global de systèmes macromoléculaires comme les microtubules des neurones, a ainsi été présentée par Penrose et Hameroff comme candidate au titre de « mécanisme de la conscience » au sens plein et entier du terme. L'effet tunnel de la mécanique quantique, les condensats de Bose dans la matière biologique,

les champs électromagnétiques quantifiés émis par l'activité neuro-électrique de l'encéphale [18], ou encore le bain moléculaire offert par le substrat glial au réseau neuronal, sont d'autres possibles substrats physiques de l'expérience vécue, par-delà les fonctions cognitives qui sont greffées sur ces substrats. S'il en allait ainsi, si l'accomplissement des fonctions cognitives ne suffisait pas à engendrer la conscience phénoménale, mais que la base physique particulière sur laquelle sont implémentées ces fonctions chez les êtres vivants était seule capable de la faire émerger, alors il est probable que la plupart des machines comportementalement et fonctionnellement équivalentes à nous seraient dénuées de conscience phénoménale. Et comme nous n'avons aucune certitude concernant *lequel* de ces éléments de substrats physiques sous-tend ou ne sous-tend pas la conscience phénoménale (la multiplicité des théories à ce sujet en témoigne), la recherche d'un artéfact authentiquement conscient, c'est-à-dire doté d'expérience, n'est qu'un tâtonnement dans le noir.

Plus embarrassant encore : les théories physiques de la conscience phénoménale que je viens d'évoquer n'ont d'autre plausibilité que celle qu'elles tirent du remplissement, par le mécanisme qu'elles postulent, de la *fonction* centrale de la *conscience d'accès* qu'est le pouvoir de synthèse des informations. Les condensats de Bose, la non-séparabilité quantique, les champs électromagnétiques quantifiés, ne sont convoqués dans la réflexion sur la conscience qu'en raison de leur puissante capacité *intégratrice*. *Et de rien d'autre.* Qu'est-ce qui nous garantit alors qu'un mécanisme *quantique* d'intégration de l'information est plus apte qu'un mécanisme *classique* d'intégration de l'information, à engendrer l'expérience vécue ? Rien de ce que nous connaissons sur la mécanique quantique ne nous le certifie et ne le laisse même soupçonner. On ne trouve rien dans aucune science objective, pas plus la théorie quantique que les autres, qui évoque de près ou de loin l'expérience vécue, et qui nous rapprocherait donc d'un compte-rendu scientifique de la genèse de cette expérience vécue.

[18] Sur toutes ces propositions, voir : M. Bitbol, *Physique et philosophie de l'esprit,* Flammarion, 2000 ; P. Uzan, *Conscience et physique quantique,* Vrin, 2013, chapitres VIII et IX

Au total, nous n'avons rigoureusement aucun critère nous permettant de savoir, ni même de *deviner,* qu'un artéfact fabriqué par nous est ou n'est pas doté de conscience phénoménale. Il est vrai que nous pourrions tomber dessus par hasard, et mettre en place les conditions d'une conscience phénoménale sans le faire exprès ; mais dans ce cas, nul signe, pas le plus petit indice, ne nous permettrait de savoir que nous avons réussi (ou de savoir le contraire). C'est ce que signale à juste titre le neurobiologiste Jesse Prinz : « À quel degré de proximité avec le cerveau humain un ordinateur doit-il parvenir, avant que nous puissions dire qu'il est probablement conscient ? Il n'y a aucune manière de répondre à cette question »[19]. Et comme il n'y a par principe nul moyen objectif de savoir si cet être artéfactuel a un « quelque chose que cela fait d'être lui subjectivement », notre ignorance à ce propos est insurmontable. En particulier, il n'est pas question de la surmonter en prenant les prescriptions cognitives utilisées pour construire un artéfact *déclaré* conscient au nom de ses comportements ou de son architecture, comme seul *critère* de la réussite de l'opération. Car alors, le cercle logique de la démarche serait hermétiquement clos, et il s'agirait d'un cercle vicieux. La vraie question demeurerait, et elle se tiendrait *sous* le niveau de la logique.

Que vaut cette tentative d'obtenir un résultat qu'aucun d'entre ceux qui l'ont voulu ne pourra *par principe* attester ? Elle vaut exactement ce que vaut une proposition *par principe* non testable ; une proposition dont rien ne permettra jamais de décider si elle est vraie ou fausse. *Une proposition dont la valeur de vérité est* par principe *indécidable est une proposition dénuée de sens.* Par extension, la tentative de « fabriquer » un robot « conscient » dans la pleine acception du terme, qui inclut la conscience phénoménale, est une activité privée de sens. C'est ce que ne comprennent pas la plupart des chercheurs en sciences cognitives, parce qu'en tant que scientifiques, ils ont été éduqués à se mettre en quête d'une vérité plutôt qu'à s'interroger sur la question du sens et du non-sens de ce qu'ils font pour y parvenir. Cette interrogation sur le

[19] Prinz, J. J. « Level-headed mysterianism and artificial experience ». *Journal of Consciousness Studies,* 10, 111-132, 2003

sens ou le non-sens, développée en amont de l'attribution d'une valeur de vérité, comme condition de possibilité d'une telle attribution, relève du domaine propre de la philosophie. Or, nous venons de le voir, l'analyse philosophique prononce sur ce point un verdict sans ambiguïté : *la conscience (phénoménale) artificielle est un concept dénué de sens.*

Mais si ce concept est dénué de sens, c'est qu'au fond il tente d'enfermer un fragment de ce qui se vit en première personne dans une définition et une procédure en troisième personne. Seule la transposition du débat dans le domaine de la deuxième personne pourrait éviter ce face-à-face exclusif et stérile entre la méthode en troisième personne et l'aperçu en première personne de ceux qui l'appliquent. Car seule la dynamique des échanges en deuxième personne articule entre eux le pôle des vécus en présence avec leur pôle d'entente formelle traité comme une sphère d'objectivité, les premières personnes et leur foyer commun de visée dit « en troisième personne ». Notre enquête doit donc à présent se réorienter vers un questionnement sur la constitution d'intersubjectivité et sur son possible élargissement à des entités robotiques.

Comment reconnaissons-nous un autre être comme *alter-ego* ? Lui *attribue*-t-on alors une conscience, une expérience vécue ? Cette dernière question semble triviale, mais elle est plus épineuse qu'il n'y paraît. Après tout, reconnaître un être comme *alter-ego* n'impose pas forcément de lui attribuer une conscience en propre, mais simplement de ne pas la lui dénier, et de supposer implicitement qu'on se meut avec lui dans un espace d'expérience vécue. Supposons quand même qu'on attribue une conscience à un *alter-ego,* sur quels critères se fonde-t-on pour cela ? Les critères d'attribution d'une conscience à quelque autre être sont-ils constants ou varient-il en fonction des présupposés culturels et de l'état de la technologie ?

Seule la phénoménologie peut nous guider dans le labyrinthe des procédés de constitution d'intersubjectivité à partir d'une subjectivité, puisqu'elle seule prend la subjectivité au sérieux ; non pas (évidemment) comme un objet d'étude extérieur, mais comme un milieu à explorer de l'intérieur. Cependant, la phénoménologie n'est pas unanime sur la question de l'intersubjectivité. La variété des théories phénoménologiques de l'intersubjectivité risque dès lors d'avoir pour conséquence

une variété de positions quant à l'attitude qu'un être humain pourrait ou devrait avoir à l'égard d'un artéfact se comportant *comme s'il* était conscient.

Par souci de simplicité, je me concentrerai sur Husserl, dont les textes contiennent suffisamment de ressources et de variantes pour notre thème de recherche. Le premier mode de constitution de l'intersubjectivité au sens de Husserl est fondé sur une approche dite « cartésienne ». « Je doute », écrit Husserl, suppose déjà « je suis » [20]. Le doute cartésien étant une forme de réduction phénoménologique, il en résulte qu'à l'issue de la réduction, le champ de mon expérience propre me saute soudain aux yeux ; et que le problème devient alors de savoir si je peux étendre cette certitude personnelle à d'autres êtres qui me sont semblables. Mais Husserl a aussi exploré une voie très différente, qu'il appelle la « réduction intersubjective ». Dans cette approche, « nous pouvons exercer une réduction sur les actes des autres appréhendés par l'intropathie. En nous projetant en eux, nous pouvons y opérer une *épochè* phénoménologique comme si nous étions eux-mêmes » [21]. Ici, l'intropathie (ou empathie) est considérée comme un mouvement préliminaire qui nous installe d'emblée dans une modalité intersubjective de l'*épochè*, au lieu d'être un simple ajout à une modalité subjective standard de l'*épochè*. Cela ouvre la voie à une approche alternative, résolument non-solipsiste, de l'intersubjectivité. De même qu'il existait une post-constitution d'intersubjectivité du point de vue de la réduction cartésienne (selon l'ordre « *ego* d'abord, *alter-ego* après »), Husserl considère qu'il peut y avoir une *pré*-constitution de l'intersubjectivité du point de vue de la réduction intersubjective (où *l'ego* et l'*alter-ego* sont virtuellement concomitants, l'*alter-ego* devenant, lui aussi, une *« certitude (tacite) de l'expérience »).*

Commençons par la post-constitution de l'intersubjectivité, à partir du sujet « ego ». Ici, on part du moi considéré comme une « monade » au sens de Leibniz, puis on tente de constituer des *alter-egos* par extrapolation du champ de conscience transcendantal de ce moi : « Chaque *ego* est une 'monade' » [22]

---

[20] E. Husserl, *Méditations cartésiennes* §9, Vrin, 2014, p. 48

[21] E. Husserl, *Philosophie première 2,* Presses Universitaires de France, 1972, 47e Leçon, p. 189

[22] *Husserliana* XIII, *Zur Phänomenologie der Intersubjektivität 1920-1928,* Beilage XXX, Martinus Nijhoff, 1973 p. 233 ; cité par N. Depraz, *Transcendance et*

écrit Husserl. Mais comment les monades sortent-elles de leur sphère, de ce point de vue solipsiste de la réduction cartésienne ? Comment naît la conviction qu'il existe d'autres *egos* et d'autres expériences ? Le processus décrit par Husserl suppose une interconvertibilité de mon corps propre habité de proprioceptions, et de mon corps objectivé accessible par la perception. Puis, étant donnée l'analogie entre le corps objectivé de l'autre et mon corps objectivé, et étant donné que j'imagine que l'autre voit en tant que corps objectivé ce corps-ci que je vis d'abord en tant que corps propre, j'associe réciproquement un corps propre capable de « sentiment » au corps objectivé de l'autre. Selon Husserl, « À travers un jeu d'*imagination* sur mes potentialités, je peux mettre le centre spatial de ma sphère primordiale, le 'ici' où se trouve mon corps propre, à la place du 'là-bas' où je perçois le corps objectivé des autres »[23]. Mais bien sûr, l'imagination n'est pas tout ; à cela s'ajoute l'empathie, avec le potentiel de partage émotif qu'elle comporte : « (…) les monades ont des fenêtres, poursuit Husserl. ... Les fenêtres sont des intropathies »[24].

Nous voyons qu'à ce niveau encore rudimentaire de la constitution d'intersubjectivité, tout ce qui est requis pour faire d'un autre être un alter-ego, pour lui reconnaître un corps propre ressenti comme tel, c'est la possibilité imaginative d'inter-substitution entre l'action propre et le comportement de l'autre, et c'est aussi la reconnaissance chez l'autre des signes de l'émotion dans des circonstances où l'on est soi-même susceptible d'être ému. La réalisation d'artéfacts capables de simuler des états mentaux et des affects, c'est-à-dire de se comporter *comme s'ils* avaient des états mentaux et des affects, suffit à mettre en branle un tel processus d'identification. Quant à la preuve absolue que l'autre est bien le siège de « sentiment » ou d'expérience vécue, elle n'existe pas plus pour les autres êtres humains que pour les robots. La seule chose qui la rend plus crédible dans le premier cas que dans le second est une clause « *ceteris paribus* » (toutes choses étant

---

incarnation : Le statut de l'intersubjectivité comme altérité à soi chez Husserl, Vrin, 1995, p. 323-324.

[23] J. Tryssesoone (2006), "Les chemins de l'intersubjectivité dans la philosophie de Husserl", *Bulletin d'analyse phénoménologique,* 2, 3-76

[24] *Husserliana* XIII, *Zur Phänomenologie der Intersubjektivität 1920-1928,* Beilage XXX, Martinus Nijhoff, 1973 p. 233 ; cité par N. Depraz, *Transcendance et incarnation : Le statut de l'intersubjectivité comme altérité à soi chez Husserl,* op. cit., p. 323-324.

égales par ailleurs). Tous les constituants biologiques, physiques, chimiques étant égaux entre nous par ailleurs, toutes nos positions dans un réseau de relations sociales, de dépendances naturelles et d'héritages historiques étant semblables par ailleurs, j'ai de fortes chances de ne pas me tromper si je juge que le comportement organisé d'un autre être humain exprime les contenus de sa conscience phénoménale. En revanche, dans le cas du robot, faute de ces similitudes étendues, je ne suis pas sûr du tout que les conditions nécessaires et suffisantes de la conscience phénoménale sont réalisées lorsque les conditions d'un comportement et d'une organisation cognitive qui l'évoque le sont.

À ce stade, il semble qu'il y ait quelque chose de difficile à faire pour constituer *l'alter-ego*; il semble que l'*alter-ego* soit plus problématique que l'*ego*. C'est ce que j'ai appelé la post-constitution laborieuse de l'*alter-ego*.

Mais cette post-constitution pourrait bien dériver d'une *épochè* phénoménologique incomplète, où le *moi* reste seul incontesté. À rebours de cela, Husserl a remarqué (anticipant Patočka[25]) qu'une *épochè* vraiment radicale ne serait pas une *épochè* subjective ; ce serait plutôt une *épochè pré*-subjective. « En tant que phénoménologue, je me suis mis *moi-même* hors circuit, à l'égal de tous les autres, de telle sorte qu'il ne reste même pas un *solus ipse* »[26]. Or, si notre point de départ n'est pas la subjectivité mais la pré-subjectivité, si notre point de départ n'est pas un *ego* mais un champ d'expérience pré-égotique, la constitution d'intersubjectivité ne diffère en rien de celle de la subjectivité. Aucune dissymétrie fondamentale ne demeure entre *ego* et *alter-ego*.

Mais en quoi consiste exactement ce champ d'expérience pré-égotique? Tout simplement en une *pure expérience présente*. Ainsi que le déclare Husserl: « Je ne suis donné à moi-même d'une manière absolument immédiate que dans le pur *présent* de ma vie »[27]. Comment alors mon moi est-il constitué dans la pure expérience présente ? Il est constitué par

[25] J. Patočka, *Papiers phénoménologiques,* Grenoble : Jérôme Millon, 1995, p. 195

[26] J. Tryssesoone (2006), "Les chemins de l'intersubjectivité dans la philosophie de Husserl", loc. cit. La phrase citée est un commentaire par J. Tryssesoone d'un cours de 1911 donné par Husserl : E. Husserl, *Problèmes fondamentaux de la phénoménologie,* Paris : Presses Universitaires de France, 1992 ; *Husserliana* XIII : *Zur Phänomenologie der Intersubjektivität, Erster Teil : 1905-1920,* text n°6, Den Haag : Martinus Nijhoff, 1973, 152/154

[27] E. Husserl, *Philosophie première 2,* op. cit. 241/175

l'interconnexion d'une chaîne d'événements passés et d'une poussée d'attentes futures, articulés par la mémoire présente. Car la mémoire ne se contente pas de donner accès au passé ; elle découpe les attentes sur fond d'un passé et leur donne alors le sens d'un projet (le projet d'un moi qui s'identifie à ce passé).

« Un souvenir, écrit Husserl, me donne accès au transcendantal de deux manières. (...) D'une part, le 'je me souviens' m'est à présent donné dans le cadre de ma vie transcendantale actuelle ; et d'autre part, ce 'je me souviens' évoque ma vie transcendantale passée »[28]. Husserl élabore ici le concept d'une réduction phénoménologique à deux niveaux. Le premier niveau est la réduction qui me ramène de l'attitude naturelle consistant à se souvenir du passé vers la conscience d'un état actuel d'évocation d'un souvenir ; et le second niveau est la réduction qui me ramène à mon état antérieur par une forme d'auto-empathie : elle me reconduit vers un être qui vivait alors dans sa vie présente cette expérience même que je perçois maintenant comme passée. Le premier niveau de la réduction est un niveau de présentation directe de l'acte présent de se souvenir. Le deuxième niveau de la réduction concerne quant à elle la présentation indirecte d'une situation passée qui a été vécue à l'époque comme présente. La présentation directe d'un acte présent est simplement appelée une « présentation », alors que la présentation indirecte d'un acte passé alors vécu comme présent s'appelle une « présentification »[29].

Le dispositif d'une réduction à deux niveaux, et la différence entre « présentation » et « présentification », servent chez Husserl de modèle pour le problème de l'intersubjectivité. Nous avons vu qu'une subjectivité personnelle peut être constituée à partir d'une expérience présente neutre en accédant à l'expérience vécue passée par une combinaison de souvenirs et d'empathie de soi. D'autres subjectivités peuvent être constituées à partir de la même expérience présente neutre en actualisant leur expérience située par les effets d'empathie qu'éveillent leurs discours et leurs comportement. Comme la mémoire, l'empathie *(einfühlung)* est une présentification dans

[28] E. Husserl, *Philosophie première 2,* op. cit. 121/85

[29] Le mot « présentification » a été choisi par Jean-Paul Sartre comme traduction française du mot allemand « *Vergegenwärtigung* » qu'utilise Husserl dans sa *Phénoménologie de la conscience intime du temps.*

l'expérience neutre, une présentification dont l'objet intentionnel présuppose une autre expérience. Aucun écart, aucune dissymétrie ne subsiste alors entre l'*ego* et l'*alter-ego.* Les deux sont constitués à partir du flux neutre de l'expérience vécue actuelle, et par des procédures similaires.

Mais alors, la perspective entière du problème des « autres consciences » (y compris celles des robots), est profondément transformée. Il ne s'agit plus de penser « j'*ai* une expérience, enfermée en ce lieu et en ce corps, et je me demande si ces êtres humains et ces robots *ont* aussi une expérience enfermée dans leur lieu et leur corps ». Car, selon l'approche de la constitution phénoménologique, l'expérience vécue, la conscience phénoménale, *n'est pas* quelque chose que j'ai ou que vous avez, ce n'est pas quelque chose d'enfermable dans les limites d'un lieu ou d'un corps.

Essayons de comprendre cela, car c'est le point crucial. Les simples expressions, « j'ai », « vous avez », supposent que le sujet du verbe (je ou vous) préexiste, et que ce qu'il possède lui est en quelque sorte surajouté, selon le vieux schéma grammatical et aristotélicien de la substance et des prédicats. Or, la phénoménologie (telle que la présente Husserl lorsqu'il dépasse ses racines cartésiennes) conduit à inverser cette hiérarchie. Le sujet de l'expérience est constitué à partir de l'expérience, au lieu que l'expérience soit celle d'un sujet. Dans certaines conditions de mémorisation et d'activité de synthèse, l'expérience présente s'identifie comme celle d'un sujet durable, au lieu qu'un sujet existant puisse être dit *avoir* cette expérience ; et l'expérience se déploie dans une communauté de sujets liés par l'empathie, au lieu que chaque sujet puisse être dit *produire* une expérience pour son compte.

Aborder l'expérience du point de vue phénoménologique permet aussi de comprendre la relativité anthropologique de sa distribution dans l'environnement visible et tangible. Il y a plusieurs manières, culturellement déterminées, de constituer les sujets d'expérience, et donc plusieurs manières de distribuer l'expérience dans l'environnement. Philippe Descola a répertorié quatre distributions de l'expérience, dans l'espace des cultures. La nôtre s'appelle naturalisme, et culmine dans le physicalisme. Elle tend à dénier la conscience aux choses naturelles, et à la réserver dans un premier temps aux êtres humains. Mais comme l'option du retrait s'est avérée

prodigieusement efficace pour l'élaboration des sciences, elle a été poussée dans un deuxième temps jusqu'au bout, jusqu'à l'absurde, en traitant tout, y compris nous-mêmes, comme des objets inertes dont la conscience n'est qu'illusion. D'autres options existent pourtant dans des cultures différentes ; comme par exemple l'animisme, qui tient que « tout est plein d'âmes », qui attribue une personnalité, un esprit conscient, à toutes sortes d'êtres, animaux, végétaux, voire massifs montagneux. La question de savoir si des robots sont ou seront conscients se transcrit alors en interrogation sur le régime de distribution *des* expériences dans l'espace rendu disponible par constitution des sujets *dans* l'expérience neutre actuelle.

Pour l'instant, nous hésitons. Nous nous arc-boutons sur le naturalisme dans sa version forte, physicaliste, et nous voulons savoir si les objets physiques « robots » peuvent être conscients, après avoir donné un sens objectif à ce prédicat relevant du vécu « subjectif ». Mais l'arrivée concrète des robots dans nos vies est en passe de faire craquer ce vêtement naturaliste trop étroit, et d'ouvrir une nouvelle époque dans notre culture ; une époque post-naturaliste bien plus que post-humaniste. En effet, dans le domaine des attitudes et des êtres-au-monde, on fait de moins en moins la différence entre traiter un être comme conscient et lui attribuer une conscience. Et dans le domaine des doctrines, le panpsychisme se répand, comme un compromis baroque entre le physicalisme et une forme moderne d'animisme, à la faveur de l'expansion du domaine de l'empathie. Mais si le panpsychisme témoigne ainsi de la mutation profonde du mode de distribution de l'expérience, il n'a pas encore intégralement brisé le cadre métaphysique naturaliste qui a sous-tendu la modernité occidentale. Seule une synergie entre phénoménologie et disciplines contemplatives pourrait asséner le coup fatal à ce cadre métaphysique manifestement inapproprié.

L'ironie historique est frappante : l'ultime produit artéfactuel d'une science construite sous un présupposé naturaliste est en train de rendre le naturalisme caduc. Mais avec la fin du naturalisme, toute la problématique des consciences artificielles est renversée. La vraie question post-naturaliste est celle de savoir comment cultiver *LA* conscience actuelle (vécue au singulier du présent absolu) dans de nouveaux espaces de

socialisation et avec de nouveaux besoins de sens, plutôt que de savoir comment la bouturer sur des artéfacts.

Et d'ailleurs, pourquoi voudrait-on plus que cela ? Pourquoi voudrait-on que les robots soient « réellement » le siège ou le centre de perspective d'une conscience phénoménale, en supplément de leur comportement analogue à celui d'un être conscient ? Qu'est-ce que cela nous apporterait de plus par rapport au « comme si » ? Une seule chose : la potentialité de s'incarner *soi-même,* de se ré-incarner faudrait-il dire, dans un robot et d'affranchir ainsi le flux d'identification vécu que nous appelons « ego » de ce corps malade et mortel. Le test pour savoir si quelqu'un a réussi à se *convaincre* de cette possibilité serait de lui poser cette simple mais dérangeante question : accepteriez-vous de mourir à l'instant si vous saviez que votre structure cognitive et vos *habitus* comportementaux ont été intégralement téléchargés dans un robot ? Ou auriez-vous un doute, celui que la structure cognitive en question, n'ayant aucun vécu associé, vous seriez alors mort pour de bon ?

Mais il y a aussi des raisons de *craindre* d'en faire trop en conférant une conscience phénoménale à un robot. Cette crainte c'est d'élever nos artéfacts au rang de véritables *sujets* vis-à-vis desquels un comportement éthique, et pas seulement utilitaire, s'impose ; mais de n'avoir pourtant aucun moyen de savoir si c'est le cas et encore moins de le prouver. Car, redisons-le, la seule preuve de la conscience n'est pas une preuve mais une évidence : l'évidence d'être, que vous vivez en ce moment même, en première personne du singulier, avant même qu'une personne (« vous ») se soit identifiée comme son sujet, et avant même qu'une fantaisie de toute-puissance ne vous ait fait ressentir comme désirable de la prêter à des robots.