

HARD-NEUROSCIENCES: GREG EGAN

LA SCIENCE-FICTION EST-ELLE DE BON CONSEIL POUR UNE APPROCHE COGNITIVE DE LA MORALE ?

Nicolas BAUMARD

Institut Jean-Nicod (EHESS/ENS/CNRS)
École des hautes études en sciences sociales
1 bis avenue de Lowendal 75007 Paris
Mél : nicolas.baumard@free.fr

Résumé

Les valeurs, et en particulier la morale, ont longtemps été délaissées à la fois par les sciences cognitives et la science-fiction. Dans ses œuvres, l'auteur de science-fiction australien Greg Egan décrit des technologies capables de changer nos valeurs en intervenant sur notre cerveau. Nous défendons l'idée que ces spéculations rejoignent le paradigme de la modularité massive et peuvent aider à comprendre à quoi pourrait ressembler une explication naturaliste de la morale.

Abstract

HARD NEUROSCIENCES: GREG EGAN
IS SCIENCE FICTION A GOOD GUIDE FOR A COGNITIVE APPROACH TO MORALITY?

Values, and particularly morality, have long been neglected by both cognitive science and science fiction. In his work, the Australian science fiction writer Greg Egan depicts technology capable of changing our values by modifying our brains. We argue that these speculations are closely akin to the massive modularity paradigm and can help us figure out what a naturalistic explanation of morality could be.

Resumen

HARD NEUROSCIENCES: GREG EGAN
¿ES LA CIENCIA-FICCIÓN UN BUEN CONSEJERO PARA UN ENFOQUE COGNITIVO DE LA MORAL?

Los valores, y en particular la moral, los han desatendido durante mucho tiempo a la vez las ciencias cognitivas y la ciencia-ficción. En sus obras, el autor de ciencia-ficción australiano Grez Egan describe tecnologías capaces de cambiar nuestros valores interviniendo en nuestro cerebro. Defendemos la idea que estas especulaciones se aparentan al paradigma de la modularidad masiva y pueden ayudar a entender lo que podría ser una explicación naturalista de la moral.

Resumo

HARD NEUROSCIENCES: GREG EGAN

A FICÇÃO CIENTÍFICA PODE SER UM GUIA PARA UMA ABORDAGEM COGNITIVA DA MORAL?

Os valores, e em particular a moral, foram durante muito tempo colocados à parte tanto pelas ciências cognitivas quanto pela ficção científica. Em suas obras, o autor de ficção científica australiano Greg Egan descreve tecnologias capazes de mudar nossos valores ao intervir sobre o nosso cérebro. Defendemos a idéia de que essas especulações vão de encontro ao paradigma da modularidade massiva e podem ajudar a compreender como poderia ser uma explicação natural da moral.

Riassunto

HARD NEUROSCIENCES: GREG EGAN

LA FANTASCIENZA È UNA BUONA GUIDA PER UN APPROCCIO COGNITIVO ALLA MORALITÀ?

I valori, ed in particolare la moralità, sono stati a lungo trascurati sia dalle scienze cognitive che dalla fantascienza. Nelle sue opere, l'autore australiano di fantascienza Greg Egan descrive delle tecnologie che riescono a cambiare i nostri valori intervenendo sul nostro cervello. Difendiamo l'idea che queste speculazioni raggiungono il paradigma della modularità massiva e possono aiutare a capire in che termini potrebbe essere data una spiegazione naturalista della moralità.

1. Introduction

La morale a longtemps été absente du programme de recherche des sciences cognitives. D'une part, ces dernières se sont d'abord construites autour du traitement de l'information, en particulier de l'information perceptive et langagière ; d'autre part, la morale semblait trop éthérée, trop émotionnelle, trop culturelle pour faire l'objet d'un programme scientifique de recherche. Mais les temps changent : neuro-imagerie, neuropsychologie, psychologie évolutionniste, anthropologie cognitive entre autres proposent des expériences et des théories sur le fonctionnement de la morale humaine.

Dans ce contexte, la nouvelle de Greg Egan « Axiomatique » (Egan, 1997) fournit une vision intéressante de ce à quoi pourrait conduire la conjonction de ces recherches. L'histoire est la suivante. Ne parvenant pas à oublier sa femme tuée lors d'un braquage de banque, le héros de cette nouvelle décide, à cours de solutions, de tuer son meurtrier récemment sorti de prison. Cette décision ne l'empêche pas cependant de considérer que cet acte serait un crime. Il achète alors une nanomachine qui modifiera pendant quelques jours certaines de ses configurations neuronales de manière à changer ses valeurs morales le temps de l'assassinat. En définitive, le changement de ses valeurs morales aura d'autres conséquences que la levée de son inhibition à tuer.

2. Science et science-fiction

Mon propos dans cet article n'est pas tant de discuter des mérites d'une théorie cognitive de la morale encore très programmatique que d'examiner dans quelle mesure la science-fiction peut entretenir des relations fructueuses avec la science en

train de se faire. Traditionnellement, la science-fiction est vue comme une façon d'anticiper les progrès techniques ainsi que leurs conséquences sur la société. Elle joue donc un rôle en dehors de la science : comme moyen de débattre de la science et de la vulgariser. C'est un autre type de relations, plus conjecturales et plus fortes, que j'aimerais examiner ici : la science-fiction peut-elle contribuer à la recherche scientifique en tant que telle ?

Cette relation forte science / science-fiction telle que nous l'examinons ici au travers de l'exemple théorie cognitive de la morale / nouvelles de Greg Egan demande donc d'abord pour être prise au sérieux de rapprocher les deux domaines. D'une part, la science-fiction, particulièrement à travers le courant dit *hard-science*, peut être un exercice très contraint par la connaissance scientifique et très en phase avec les développements scientifiques récents comme l'ont montré les courtes nouvelles parues dans chaque numéro de *Nature* durant l'année 2000. D'autre part, la science n'est pas une activité isolée qui produirait seule ses propres concepts. Ceux-ci circulent, allant des activités non scientifiques vers la science et vice-versa comme l'a montré l'histoire des sciences. Ils mobilisent souvent les mêmes ressources cognitives comme le suggèrent la didactique des sciences et l'anthropologie cognitive (Heintz, 2004) à travers l'étude de l'usages des analogies, des images, des paradoxes, des expériences de pensée, bref des intuitions produites au cours de l'histoire de notre espèce et utilisées aujourd'hui dans la vie quotidienne (e.g., la représentation spatiale pour l'infini algébrique, la physique naïve pour les phénomènes sociaux ou encore la théorie de l'esprit pour le comportement d'un atome vis-à-vis des électrons ou des « gènes égoïstes » dans la théorie des jeux évolutionnaires).

Popper (1975 [1935]) a proposé de définir les énoncés scientifiques comme étant des énoncés falsifiables, c'est-à-dire susceptibles d'être infirmés par la réalité (par observation et expérience). Il cherchait d'abord à montrer que le marxisme et la psychanalyse ne sont pas des activités scientifiques mais pour autant sa position n'était pas de réduire la science aux énoncés falsifiables. Il a mis en avant dans ses articles postérieurs le rôle de la *métaphysique*, c'est-à-dire l'ensemble des spéculations non falsifiables, sur lesquelles s'appuient les scientifiques pour proposer des énoncés scientifiques. Dans son acceptation la plus englobante, il s'agit de visions du monde, de *Weltanschauung*. Dans un sens plus restreint, il s'agit de programmes de recherche non falsifiables, de paradigmes au sens de Kuhn (1975) qui indiquent le type de questions à se poser et le type de réponses que l'on peut donner. Ainsi, en sciences cognitives, un long débat plus ou moins métaphysique a eu lieu autour de la conscience et d'une approche naturaliste de celle-ci (voir par exemple Dennett, 1991). De la même manière, il me semble que les nouvelles de Greg Egan sur la morale pourraient jouer un rôle dans le débat scientifique sur le type de vision du monde ou de programme de recherche susceptibles de se révéler les plus fructueux dans l'étude scientifique de la morale. Ma question devient donc : la science-fiction peut-elle être considérée comme une métaphysique ?

3. Déterminer et changer les valeurs

On pourrait faire à propos de la science-fiction le même constat que celui qu'on a fait à propos des sciences cognitives : elle a longtemps laissé de côté le thème de la modification des valeurs pour se concentrer sur les modifications possibles des pouvoirs humains. Pour le dire à la manière des économistes, elle s'est plus intéressée aux questions d'efficacité et de rationalité qu'aux préférences,

considérées comme données. Les héros ou les créatures de science-fiction ont eu tendance à pouvoir réaliser tout ce que les humains normaux voulaient beaucoup plus qu'à pouvoir vouloir toute autre chose que ce que les humains voulaient.

C'est l'une des originalités de Greg Egan que de s'attaquer de front à ce problème, tant à l'aspect scientifique de l'implémentation des valeurs qu'à ses conséquences en philosophie politique. Si la technologie nous permet de choisir nos buts, à partir de quoi allons-nous pouvoir choisir ceux-ci ? Il y revient à plusieurs reprises au cours de son œuvre. Ainsi, dans « Fidélité » (Egan, 1994 [1991]), les héros, après plusieurs mariages invariablement achevés dans une lassitude réciproque, ont la possibilité de stabiliser de manière définitive les poids synaptiques des réseaux neuronaux responsables de la cristallisation amoureuse. Ils peuvent ainsi décider de fixer le désir qu'ils ont pour la personne au moment où ils leur semblent être le plus attachés l'un à l'autre. Leurs sentiments ne changent pas alors même qu'ils ont conscience que sans cet artifice, ils se seraient probablement quittés quelques années plus tard. Dans « Raisons d'être heureux » (Egan, 1999a [1997]), un enfant, atteint d'un cancer du cerveau perd les neurones impliqués dans les centres du plaisir et de l'évaluation émotionnelle. Après avoir végété pendant toute son adolescence, un nouveau traitement lui permet de projeter de nouveaux neurones et de retrouver la possibilité d'évaluer ses actions et ses perceptions. Mais le traitement projette des neurones dans toutes les directions et le héros se retrouve ainsi à aimer au plus haut point tous les styles de musique, l'homosexualité tout autant que l'hétérosexualité. Il lui est impossible de décider d'une conduite à prendre puisque toutes les possibilités lui paraissent également plaisantes. Il devra choisir de manière arbitraire quelles préférences il décide de conserver et à quel degré d'intensité. Enfin, dans « La Cité des permutants » (Egan, 1999b [1994]), les héros sont des humains numérisés, avec l'éternité devant eux, qui ont la possibilité non seulement de modifier à leur guise leur environnement virtuel, ses habitants, leur propre corps mais aussi de décider quelle sera leur personnalité, comment elle évoluera, ce qui les intéressera. Si la question scientifique de la modélisation du cerveau est un peu évacuée, le problème philosophique est poussé à son paroxysme et l'un des héros sera conduit à choisir des préférences de manière complètement arbitraire qui se modifieront de manière tout à fait stochastique, le faisant se passionner pendant des années pour la sculpture des pieds de chaises avant de passer subitement à l'entomologie.

Terminons cette revue en notant que la question n'est déjà plus complètement du ressort de la science-fiction. En sciences politiques, Fukuyama (2002) a montré que notre biotechnologie balbutiante (neuropharmacologie, génétique, etc.) pose déjà des questions de choix des préférences et ouvre la voie à des solutions biologiques à des problèmes politiques. Par ailleurs, en philosophie politique (Rawls, 1971 ; Nozick, 1974), les valeurs morales et plus généralement les préférences de chacun sont elles aussi mises à distance, les chercheurs proposant des métapréférences, des règles pour déterminer à quels types de préférences (altruisme, estime de soi, travail, etc.) on doit donner la priorité pour bien vivre en société.

4. La morale comme un module cognitif

En quoi peut-on dire qu'Egan dans sa nouvelle « Axiomatique » rejoint les recherches actuelles sur les bases cognitives de la morale ? On peut évidemment regretter le flou qui entoure dans la nouvelle l'action de la nanomachine sur les neurones. Pour autant, changer les configurations neuronales n'est pas la même

chose que prendre un filtre magique ; cela implique que l'on puisse un jour rendre compte du comportement moral humain en termes de neurotransmetteurs et de décharges électriques, en termes donc d'entités naturelles. Comme pour la conscience, il n'est pas sûr que l'adhésion au programme naturaliste soit complètement partagée, y compris au sein des sciences cognitives (pour ne pas parler des sciences sociales qui ont traditionnellement la charge de rendre compte du comportement moral).

En illustrant ce à quoi pourrait ressembler une machine construite à partir d'une théorie naturaliste de la morale, Greg Egan contribue à rendre plus tangible ce programme de recherche encore relativement spéculatif. Il faut garder en tête que de larges pans des sciences sociales comme des sciences cognitives ne considèrent pas qu'il existe des intuitions et des émotions morales, et que le comportement moral n'est seulement qu'un calcul égoïste en situation sociale. Ainsi par exemple, alors qu'Adam Smith dans sa *Théorie des sentiments moraux* (2003 [1759]) avait défendu un programme de recherche sur les émotions morales universelles (et en particulier l'empathie), les économistes ainsi que de nombreux sociologues soucieux de théories parcimonieuses refusent d'analyser les phénomènes sociaux en termes de sentiments moraux. Rappelons enfin qu'en sciences cognitives, le rôle des émotions est encore mal intégré aux autres processus cognitifs (Damasio, 1994).

Notons pour commencer que cette nanomachine qui reconfigure transitoirement quelques réseaux neuronaux ne transforme le héros ni en machine à tuer, comme le serait un robot d'Asimov obéissant à des lois « criminelles » de la robotique, ni en assassin conditionné, comme le serait un humain dressé dans *Le meilleur des mondes*.

« Il ne s'agissait pas d'avoir une voix dans la tête qui réciterait un laïus mal écrit que je pouvais choisir de tourner en ridicule ou d'ignorer. Il ne s'agissait pas non plus de passer une sorte de décret mental que je pourrais ensuite court-circuiter, ou auquel je pourrais échapper en ergotant. Les implants axiomatiques avaient été élaborés à partir d'analyses de structures neurales existantes, étudiées sur de vrais cerveaux. Ils n'étaient pas basés sur des axiomes exprimés par le langage. » (Axiomatique, 1997 ; p. 17, je souligne)

Egan ne suggère donc pas dans le passage que l'on vient de lire que les humains sont de simples circuits composés d'« interrupteurs » que l'on pourrait brancher, débrancher ou court-circuiter. Si la morale était régulée de cette manière, ce ne serait pas de la morale, mais un comportement plus basique que l'on rencontre chez des espèces aux capacités cognitives bien plus réduites (par exemple un instinct d'évitement de la violence comme le suggérait l'éthologie classique, voir Lorenz, 1969). De nombreuses théories sociobiologiques (Wilson, 1975), inspirées de l'étude des insectes, ont des difficultés à rendre compte de la dimension évaluative de la morale : les modèles évolutionnaires montrent que des comportements altruistes peuvent être sélectionnés dans certaines circonstances. Ils n'expliquent pas pourquoi les humains ont recours à des normes et à des intuitions de devoir.

Il ne s'agit pas non plus de conditionnement de type béhavioriste (ou culturaliste en sciences sociales) qui agirait par répétitions et que l'on pourrait combattre par de pareilles répétitions. La morale n'est pas (qu')une habitude mais une capacité spécifique, fonctionnant dans un format spécifique qui n'est pas celui du langage et auquel le langage n'a pas accès. Les changements moraux doivent se faire dans ce format et non pas en « ergotant », de même que les valeurs morales ne sont pas exprimées par le langage mais comme intuition spécifique.

La façon dont Egan nous décrit le changement des valeurs morales nous semble aller dans le sens du cadre de la modularité massive (Sperber, 2002). Ce paradigme s'appuie sur l'analyse de Fodor (1983) qui proposait une distinction entre les processus périphériques (perception, langage) qui seraient modulaires et les processus centraux (raisonnement) qui ne le seraient pas. L'exemple paradigmatique du processus modulaire est le langage. Ainsi, par modulaire, Fodor entendait notamment des processus *spécialisés* pour un type de stimulus (les voix humaines), *automatiques* (ils se déclenchent dès qu'un stimulus est présent, sans qu'un individu puisse s'y opposer — on ne peut pas ne pas vouloir comprendre une phrase en français), *encapsulés* (le traitement n'est pas influencé par les sorties des autres modules, ainsi les illusions d'optique persistent), et enfin *impénétrables* (nous n'avons pas accès à la façon dont sont traitées les informations mais seulement au résultat de ce traitement, dans un format spécifique comme les intuitions syntactiques).

Le paradigme de la modularité massive propose d'étendre cette analyse à tous les processus cognitifs. Ainsi même les processus cognitifs de haut niveau comme la logique ou la cognition sociale sont des processus modulaires. L'activité du cerveau doit être vue comme une somme de modules différents en interaction et en compétition permanente, dont l'activation dépend du contexte et des interactions modulaires (Sperber, 2005). Plus généralement, l'idée de modularité massive efface la distinction toujours prégnante entre deux types de systèmes cognitifs, d'un côté les systèmes simples (perception, action, sexe, émotion), de l'autre les systèmes plus complexes (logique, social, conscience, esthétique). On retrouve ici un souci naturaliste constamment illustré chez Egan. Dans l'étude psychologique de la morale, ce paradigme s'oppose aux approches constructivistes développées par Piaget (1932) et Kohlberg (1981) pour lesquelles la morale se développe graduellement avec les autres capacités cognitives grâce à la maturation et à l'expérience. L'hypothèse modulariste propose plutôt que les humains soient équipés d'un système spécialisé (qui apparaît à un moment spécifique du développement) qui leur permet de se comporter moralement (ressentir de la culpabilité, apprendre les normes particulières à leur culture, etc.) de manière naturelle, de la même façon qu'ils apprennent sans effort à parler ou à marcher.

Ce paradigme suggère donc une certaine autonomie et une certaine spécificité des processus moraux qui permettent de rendre cohérentes nombre d'expériences et de théories sur la morale. En neuropsychologie, Damasio (1994) a mis en évidence des troubles du cortex préfrontal affectant spécifiquement les émotions impliquées lors des décisions sociales et Blair (1995) s'est penché sur des criminels capables de comprendre des normes morales sans pour autant en ressentir la nécessité morale. En imagerie cérébrale, à l'aide de dilemmes moraux, plusieurs études ont montré que des zones cérébrales différentes étaient impliquées dans les raisonnements utilitaristes et déontologiques (pour une revue, voir Greene et Haidt, 2002). En psychologie sociale, un certain nombre de domaines moraux indépendants ont été mis en évidence comme le dégoût (Haidt *et al.*, 1997) ou la hiérarchie (Shweder *et al.*, 1997). Le paradigme de la modularité massive s'accorde parfaitement avec les théories sur l'évolution de l'esprit (Pinker, 1997 ; Tooby et Cosmides, 1992) : il propose que l'esprit soit composé, comme le reste de l'organisme, de modules sélectionnés pour des tâches très spécifiques (e.g., détecter les tricheurs, éviter les substances contagieuses, etc.). Enfin, Sperber et Hirschfeld (2004), Boyer (2001) et Atran (2002) ont proposé que ces modules innés stabilisent préférentiellement certaines représentations culturelles (e.g., honneur, vengeance, etc.).

5. Comment imaginer intervenir sur la morale sans voir l'homme comme un robot ?

On pourrait objecter à ces théories et ces expériences qu'elles ne rendent pas compte de la complexité et de la richesse de la cognition morale humaine. Tournons-nous vers le personnage d'Egan qui vient d'inspirer par ses canaux nasaux la nanomachine qui va modifier ses intuitions morales. Le personnage ne ressent rien, aucun changement dans son esprit. Il pratique divers tests mais ceux-ci, faute de contexte pertinent, ne lui donnent aucune indication. La nanomachine n'a donc créé chez lui aucune impulsion au meurtre et ce n'est qu'en se rendant chez le meurtrier de sa femme qu'il commence à se rendre compte, non qu'il a envie de tuer le meurtrier mais plutôt que la question de la moralité de cet acte n'a plus de sens chez lui.

« Est-il mal de tuer ? Mais je ne parvenais pas à me concentrer sur la question. Je trouvais même difficile de croire que je me l'étais posée dans le passé. L'idée elle-même me semblait obscure et complexe, pareille à un théorème mathématique ésothérique. L'idée d'aller jusqu'au bout de mon plan initial me retournait l'estomac — mais il s'agissait simplement de peur, non d'un sursaut moral. » (Axiomatique, 1997 ; p.22)

On retrouve ici l'idée que nous avons des systèmes spécialisés produisant des intuitions particulières. Ainsi, nous avons des intuitions sur la numérosité (Dehaene, 1996) qui nous permettent de comparer des quantités, mais nous n'avons aucune intuition sur un espace à quatre dimensions. Nous devons pour manipuler pareils concepts nous en remettre non à notre sens commun mais à des outils mathématiques inventés spécifiquement pour cette tâche et qui pallient les limites de notre cerveau. Certaines personnes, dites acalculiques, sont même dépourvues, de manière congénitale, d'intuitions de numérosité. Elles doivent pour compenser cette déficience handicapante dans la vie courante mettre en place toutes sortes de stratagèmes, sans être capables de retrouver les intuitions normales. Dans le domaine moral, Anderson *et al.*, (1999) ont montré que des patients qui souffraient de troubles du cortex préfrontal à la suite d'un accident cérébro-vasculaire très précoce (avant 16 mois) étaient incapables d'apprendre des normes morales, à la différence de personnes ayant eu un accident à l'âge adulte et qui étaient toujours capables de manipuler les normes qu'elles avaient apprises autrefois.

Les intuitions morales du personnage ont donc changé sans pour autant qu'il puisse en identifier la raison : il n'a pas accès à ses processus cognitifs et aucune information n'est venue modifier la situation. Il doit donc *a posteriori* rationaliser son jugement. C'est ce que nous faisons couramment dans la vie quotidienne comme le montrent nombre d'expériences de psychologie sociale.

« Alors pourquoi ? En fin de compte, je crois que j'ai considéré qu'il s'agissait d'une question d'honnêteté. Aussi déplaisant que cela puisse être, je voulais accepter que je voulais vraiment tuer Anderson. Même si j'avais été dégoûté par la simple idée de tuer, je devais passer à l'acte, ne serait-ce que par honnêteté envers moi-même. Si j'avais reculé, j'aurais été hypocrite. Je me serais trompé moi-même. » (Axiomatique, 1997 ; p.24)

J. Haidt (2001) a réalisé une expérience qui éclaire bien cette situation. Il propose à des sujets de donner leur avis sur l'histoire suivante :

« Julie et Mark sont frère et sœur. Ils visitent ensemble la France pendant leurs congés universitaires. Une nuit, ils passent la nuit seuls dans une cabine de plage. Ils se disent qu'il serait intéressant et amusant de faire l'amour. Ce serait une expérience nouvelle pour tous deux. Julie prend la pilule, et Mark utilise un préservatif, pour se protéger. Ils apprécient tous les deux, mais décident de ne pas le refaire. Ils gardent cette nuit entre eux, comme un secret qui les rend encore plus proches l'un de l'autre. Que pensez-vous ? Pouvaient-ils faire ce qu'ils ont fait ? » (Haidt, 2001 ; p. 814, je traduis)

La plupart des sujets désapprouvent la situation et après avoir tenté de donner quelques raisons objectives (risque d'avoir des enfants malades, désapprobation de l'entourage, etc.) qui ne peuvent tenir dans ce cas précis, continuent de soutenir leur jugement sans pour autant être capable de le justifier.

"I don't have, like, a point that says OK, that's why it's wrong. But it's like, a gut thing where I think it's wrong. I mean, you could try to possibly change my mind, but I probably wouldn't." (Haidt, communication personnelle)

Leurs réactions mettent en évidence le fait que, au moins pour certaines classes de jugements moraux, nous avons d'abord des intuitions, puis, seulement dans un second temps, nous cherchons des raisons capables de les étayer, à la manière d'un avocat qui prendrait comme donnée l'idée que son client est innocent puis chercherait ensuite comment démontrer ce jugement.

Il ne s'agit pas de dire que la réflexion morale explicite ne joue aucun rôle dans nos jugements moraux mais plutôt qu'elle n'agit qu'en interaction avec des modules cognitifs qui traitent l'information d'une certaine manière. La discussion morale, le travail de persuasion, la culture proposent ainsi des jugements qui cadrent ou non avec la configuration d'entrée d'un certain domaine. Kahneman et Tversky (2000) ont montré comment ce processus fonctionnait pour nos jugements probabilistes : un traitement qui sauve en moyenne 200 personnes parmi 600 n'est pas la même chose qu'un autre qui tuera 400 personnes sur 600. De la même façon, en déclarant que le mariage homosexuel est dégoûtant ou va dans le sens de la justice, deux types de modules et deux types d'intuitions différentes sont produits.

Il existe de nombreux exemples de manipulation consciente de nos intuitions comme lorsque nous choisissons de changer de côté dans une rue pour éviter un clochard dont la proximité nous aurait fait pitié et nous aurait forcés à lui donner de l'argent. C'est également le cas en philosophie politique lorsqu'on soumet nos idées à ce que Rawls (1971) appelle « l'équilibre réfléchi », c'est-à-dire un va-et-vient entre nos intuitions de justice et les conséquences pratiques qui en découlent. Si ces dernières posent problème, nous reformulons autrement nos principes moraux qui interagissent alors différemment avec nos modules moraux : injustes de prime abord, les inégalités sociales nous apparaîtront justes si nous nous disons qu'est juste ce qui améliore la situation des plus pauvres.

6. Où est l'axiomatique : dans nos têtes ou dans le ciel ?

La chute de la nouvelle nous donne un indice sur le type de module et les configurations neuronales affectés par la nanomachine. Au moment de tuer le meurtrier de sa femme, le narrateur s'aperçoit que cette dernière n'a plus de valeur pour lui et que son deuil n'a aucun sens.

« Tout était si clair maintenant. Je comprenais. Je comprenais l'absurdité de tout ce que j'avais ressenti pour Amy — mon « amour », mon « chagrin »... c'était des conneries. Elle n'était que de la viande. Elle n'était rien. » (Axiomatique, 1997 ; p.28)

Egan suggère ici que lorsque l'on juge que tuer est immoral, on s'appuie d'abord sur des émotions d'empathie et des indices d'attachement. Et les indices qui activent l'inhibition de tuer sont les mêmes que ceux qui activent l'empathie pour les êtres chers ou le souci des partenaires sociaux. S'il y avait peut-être d'autres moyens de modifier les intuitions morales concernant le meurtre, le fabricant de la machine a choisi d'agir sur un des réseaux les plus basiques. Affaiblir l'empathie pour la souffrance du meurtrier implique d'affaiblir également celle pour tous les autres membres du genre humain qui correspondent à la configuration d'entrée de ce module. Voilà donc l'axiomatique dont parle le titre de la nouvelle. Egan suggère ici qu'il existe des axiomes moraux qui organisent le comportement humain, des axiomes tout aussi fondamentaux que ceux des robots positroniques d'Asimov. Le héros de la nouvelle ne déclare-t-il pas au cours du délire alcoolique qui suit sa décision d'utiliser la nanomachine :

« Je criais : « HAL enfreint la Première Loi de la Robotique ! HAL enfreint la Première Loi ». » (Axiomatique, 1997 ; p.21)

Changer une intuition, une émotion pour résoudre un problème peut avoir, comme en mathématique, des conséquences sur de nombreux autres problèmes qui utilisaient cet axiome. Ces conséquences sont d'autant plus nombreuses que cet axiome est d'un niveau « élevé ». Chez certains psychopathes, cet « axiome » est d'un niveau plus élevé qu'on pourrait le penser. La sensibilité aux émotions est très affaiblie (Blair, 1995) entraînant, entre autres symptômes, un affaiblissement de l'empathie qui favorise le crime mais qui n'est qu'une des multiples conséquences d'un dysfonctionnement émotionnel.

On peut se faire une idée des caractéristiques d'entrée de ce module en pensant aux intuitions que nous avons concernant les traitements infligés aux animaux. Notre aversion envers ces traitements ainsi que les idées de droit des animaux n'ont sans doute pas été sélectionnés par l'environnement primitif de nos ancêtres chasseurs-cueilleurs. Néanmoins, les animaux domestiques sélectionnés artificiellement (et souvent inconsciemment) depuis des millénaires pour leur capacité à nous attendrir parviennent à activer nos modules tout aussi bien que des individus quelconques croisés dans la rue. On pourrait dire la même chose des petits des mammifères qui partagent de nombreuses caractéristiques faciales avec les enfants humains. Les modules cognitifs n'ont donc rien des tables de la loi exprimées en langage humain. Ils fonctionnent selon des paramètres physiques qui peuvent être aisément détournés par toutes sortes de stimulus, notamment culturels (Sperber et Hirschfeld, 2004).

Cependant, accorder autant d'importance aux émotions ne signifie pas réduire la morale à celles-ci. Ainsi, on connaît des psychopathes insensibles à autrui et capables de comprendre des normes morales (Blair, 1995). À l'inverse, les grands singes ont probablement eux aussi de l'empathie et une inhibition pour la violence intra-groupe (De Waal, 1996) sans pour autant montrer de comportements spécifiquement moraux comme la culpabilité et l'indignation qui font références à des normes partagées (Tomasello, 1999). Egan suggère seulement que nos jugements moraux sont le produit d'interactions entre des processus complexes comme la compréhension des normes et d'autres plus basiques, comme la sensibilité à la

souffrance des autres. Ces processus basiques peuvent certes être activés différemment, comme on l'a vu, par le contexte ou la façon dont est présenté le problème. Néanmoins modifier ces processus implique de modifier en retour tous les processus de plus haut niveau qui s'appuient sur eux. Pour reprendre le vocabulaire des économistes utilisé au début, changer les préférences modifie tout le processus de maximisation rationnel, et ce d'autant plus lorsque ces préférences sont fondamentales comme dans le cas de la morale et susceptibles d'influencer des pans très importants des projets d'un individu (plus que ne le serait un changement dans les préférences concernant l'appétence pour le fromage). En l'occurrence, le héros choisit à la fin de la nouvelle de retrouver cet état d'indifférence envers le genre humain que lui avait procuré transitoirement la nanomachine :

« Ce que je veux c'est ce que j'ai ressenti cette nuit-là, la conviction inébranlable que la vie d'Amy — et encore plus celle d'Anderson — n'avait tout simplement aucune importance. Pas plus que la mort d'une mouche ou d'une amibe. » (Axiomatique, 1997 ; p. 29)

Quel est le statut de ces « axiomes moraux » ? Nous avons vu qu'ils pouvaient prendre la forme d'une émotion d'empathie. Pour Egan, il s'agit évidemment de combattre la vieille idée platonicienne qui ferait exister la morale hors du cerveau humain et de promouvoir une vision naturaliste de la morale.

« Peut-être espérais-je prouver que mes convictions (...) étaient gravées sur je ne sais quelles tables métaphysiques, lesquelles planaient dans une dimension spirituelle qu'une simple machine ne pouvait atteindre. » (Axiomatique, 1997 ; p. 19)

Pour autant, on pourrait redonner une certaine « transcendance » à ces convictions morales devenues si contingentes qu'une simple machine peut les changer. Suggérer que nos convictions sont le produit de réseaux neuronaux spécifiques ne doit pas nous laisser croire que les hommes auraient pu être tout autres si ces réseaux avaient été différents. Bref, ce n'est pas parce que la morale est dans notre cerveau qu'elle résulte de la confirmation accidentelle de nos réseaux de neurones. Au contraire, elle pourrait dépendre très étroitement de lois universelles. En effet, comme cherche à le montrer la psychologie évolutionniste et plus généralement la théorie de l'évolution dont cette dernière s'inspire, ces convictions morales pourraient être des solutions sélectionnées par l'évolution pour répondre à des problèmes rencontrés par les êtres humains au cours de leur histoire.

Ces problèmes sont des problèmes généraux qui se posent à tout type d'êtres vivants évoluant en société (Dawkins, 1976). De même que la détection de la lumière pose un même problème en tout point de l'univers et est susceptible de provoquer l'apparition de solutions identiques (l'œil est ainsi apparu plusieurs fois au cours de l'évolution terrestre), de même la coopération et ses diverses solutions sont universelles. Ainsi, l'altruisme de parentèle (*i.e.*, aider les individus apparentés parce qu'ils véhiculent les mêmes gènes ; Hamilton, 1964) ou l'altruisme réciproque (*i.e.*, aider les individus susceptibles de nous aider dans le futur ; Trivers, 1971) sont des solutions susceptibles d'avoir été découvertes par d'autres êtres vivants évoluant par mutation/sélection¹. De fait, la coopération est un phénomène très répandu dans le

¹ Le groupe Science-Fiction-Philosophie a organisé le 13 décembre 2004 une journée « La vie comme phénomène universel : Science-fiction, exobiologie et écologie cognitive, philosophie de la biologie » sur le même format que la journée Greg Egan. Là encore, il nous semblait que la

monde animal et semble obéir partout aux mêmes logiques évolutives (Dugatkin, 1997).

Il est probable que nos normes morales soient reliées d'une manière ou d'une autre à ces solutions universelles aux problèmes évolutives de coopération. La capacité spécifiquement humaine à manipuler les normes crée une situation de coopération inédite (Gibbard, 1990). Les normes permettent aux êtres humains de se coordonner, et notamment, d'agir collectivement en vue d'un bien collectif et de contrôler le comportement individuel (Baumard, 2007). Dans ce nouvel environnement, un individu qui violerait une norme en tuant un congénère ou en partageant injustement les produits de la coopération pourrait être ostracisé et puni par son groupe (ce que l'on n'observe pas chez les primates). Au cours de l'évolution, une tendance à tenir compte du jugement collectif et du bien commun aurait pu être sélectionnée (dans une certaine mesure seulement, l'égoïsme et la tricherie sont parfois indétectables ou trop rentables comparés à la punition). Si certaines situations sont récurrentes (vol, viol, etc.) cette tendance à la culpabilité et à l'indignation pourrait devenir plus sensible à certains stimuli (violence, inégalité, etc.). De véritables intuitions morales se mettraient en place. Rien n'indique que ces situations récurrentes, liées à la rareté des ressources (nourriture, partenaires sexuels, places hiérarchiques) ne soient pas universelles. Bien sûr, on pourrait imaginer que les humains font face à cette nouvelle situation grâce au seul secours du calcul conscient. La psychologie évolutionniste montre pourtant qu'il est souvent plus intéressant d'utiliser des heuristiques, des règles par défaut, des systèmes biaisés plutôt que de réfléchir *on-line* ou de maximiser au plus près au risque de faire une erreur coûteuse (voler une fois de trop et se faire ostraciser).

Pour l'évolution, les préférences (pour la morale, pour ses enfants, pour la graisse), tout comme les moyens de les réaliser (systèmes de traitement de l'information, rationalité), sont des moyens au service d'une fin unique : augmenter la *fitness* de l'individu porteur des gènes qui programment les préférences. Derrière les préférences qui nous semblent parfois des axiomes dont découlent nos projets de vie, se cache un système de métapréférences — l'évolution — qui choisit celles-ci ou, pour le dire autrement, des axiomes plus fondamentaux dont nos préférences ne sont que les dérivés. Ainsi, derrière les axiomes psychologiques de nos comportements moraux se cachent les axiomes de la théorie des jeux qui s'appliquent partout dans l'univers aux situations de coopération. Ce sont eux qui expliquent pourquoi les individus munis des axiomes de comportements moraux tels que nous les connaissons ont mieux survécu dans les situations de coopération rencontrées au cours de l'histoire de notre espèce. Ce ne serait pas alors trahir Egan, qui inscrit souvent ses œuvres dans une perspective évolutionniste, que de souligner que nos intuitions morales sont à la fois dans le monde — permises par nos gènes, n'existant que dans nos cerveaux — et hors du monde — issues d'une logique évolutionniste universelle qui s'appliquent aux hommes, aux animaux et sans doute, comme la science-fiction le suggère, aux extra-terrestres !

science-fiction pouvait permettre de poser des questions scientifiques (en l'occurrence de théorie de l'évolution et de physiologie) d'une autre façon.

Références bibliographiques

- Anderson S.W., Bechara A., Damasio H., Tranel D., Damasio A.R. (1999). Impairment of social and moral behavior related to early damage in human prefrontal cortex. *Nature Neuroscience*. 2(11), 1032-1037.
- Atran S. (2002). In gods we trust: The evolutionary landscape of religion. Oxford: Oxford University Press.
- Baumard N. (2007). La morale n'est pas le social. Le point de vue de la psychologie. *Terrain*. 48, 49-72.
- Blair R.J.R. (1995). Is the psychopath morally insane? *Personality and individual differences*. 19, 741-752.
- Boyer P. (2001). Religion explained: The evolutionary origins of religious thought. New York: Basic Books.
- Damasio A. (1994). Descartes' error: Emotion, reason, and the human brain. New York: Grosset/Putnam.
- Dawkins R. (1976). The selfish gene. New York/Oxford: Oxford University.
- Dehaene S. (1996). La bosse des maths. Paris: O. Jacob.
- Dennett D. (1991). Consciousness explained. Boston: Little Brown.
- Dugatkin A. (1997). Cooperation among animals. Oxford: Oxford University Press.
- Egan G. (1994). Fidélité in Futurs, mode d'emploi. Paris: Pocket. Traduit de "Fidelity" (1991), Isaac Asimov's Science fiction magazine, September 1991.
- Egan G. (1997). Axiomatique in Axiomatique. DLM. Traduit de Axiomatic (1990), *Interzone 41*.
- Egan G. (1999a). Raisons d'être heureux. Étoiles Vives, 7. Traduit de « Reason to be cheerful » (1997), *Interzone*, 118.
- Egan G. (1999b). La cité des Permutants. Paris : Robert Laffont. Traduit de Permutation city (1994), London : Orion/Millennium.
- Fodor J.A. (1983). The modularity of mind: An essay on faculty psychology. Cambridge, MA: Bradford/MIT Books.
- Fukuyama F. (2002). Our posthuman future consequences of the biotechnology revolution. New York: Farrar, Strauss & Giroux.
- Gibbard A. (1990). Wise Choices. Apt Feelings. Harvard: Harvard University Press.
- Greene J., Haidt J. (2002). How (and where) does moral judgment work? *Trends in cognitive sciences*. 6(12), 517-523.
- Haidt J., Rozin P., McCauley C., Imada S. (1997). Body, psyche, and culture: The relationship of disgust to morality. *Psychology and developing societies*. 9, 107-131.
- Haidt J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological review*. 108, 814-834.
- Hamilton W.D. (1964). The genetical evolution of social behaviour. *Journal of theoretical biology*. 7(1), 1-16.
- Heintz C. (2004). Cognitive anthropology of science. *Journal of cognition and culture*. 4(3-4)
- Kahneman D., Tversky A. (2000). Choices, values, and frames. New York: Russell sage foundation; Cambridge: Cambridge university press.
- Kohlberg L. (1981). Essays on moral development. San Francisco: Harper & Row.
- Kuhn T. (1975). La structure des révolutions scientifiques. Paris : Flammarion.

- Lorenz K. (1969). L'agression : une histoire naturelle du mal. Paris : Flammarion.
- Nozick R. (1974). Anarchy, state and utopia. New York: Basic Books.
- Piaget J. (1932). Le jugement moral chez l'enfant. Paris : Alcan.
- Popper K. (1975 [1935]). La logique de la découverte scientifique. Paris : Payot.
- Popper K. (1985). Conjectures et réfutations. Paris : Payot.
- Pinker S. (1997). How the mind works. New York and London: W W Norton & Co Inc.
- Rawls J. (1971). A theory of Justice. Harvard: Harvard University Press.
- Smith A. (2003 [1759]). Théorie des sentiments moraux. Paris : PUF.
- Sperber D. (2002). In defense of massive modularity. In Dupoux, E. (ed.), *Language, brain and cognitive development: Essays in honor of Jacques Mehler*. Cambridge, Mass. MIT Press. 47-45.
- Sperber D., Hirschfeld L.A. (2004). The cognitive foundations of cultural stability and diversity. *Trends in cognitive sciences*. 8(1), 40-46.
- Sperber D. (2005). Modularity and relevance: How can a massively modular mind be flexible and context-sensitive? In Carruthers P., Laurence S., Stich S. (eds.), *The innate mind: Structure and content*. Oxford: Oxford University Press. 53-69.
- Shweder R., Much N.C., Mahapatra M., Park L. (1997). The "big three" of morality (autonomy, community, divinity) and the "big three" explanations. In Brandt A.M., Rozin P. (eds.), *Morality and health*. New York and London: Routledge. 119-172.
- Tomasello M. (1999). The cultural origins of human cognition. Cambridge: Harvard UP.
- Tooby J., Cosmides L. (1992). The psychological foundations of culture. In Barkow J., Tooby J., Cosmides L. (eds), *The adapted mind: Evolutionary psychology and the generation of culture*. Oxford: Oxford University Press. 19-136.
- Trivers R. (1971). The evolution of reciprocal altruism. *Quarterly review of biology*. 46, 35-37.
- de Waal F. (1996). Good natured: The origins of right and wrong in humans and other animals. Cambridge, MA: Harvard UP.
- Wilson E.O. (1975). Sociobiology. Boston: Harvard University Press.

Nicolas Baumard

Nicolas Baumard est doctorant en Sciences Cognitives à l'École des Hautes Études en Sciences Sociales. Sa thèse, sous la direction de Dan Sperber (Institut Jean-Nicod, CNRS), porte sur les bases naturelles de la morale, en particulier à travers les approches de la psychologie évolutionniste et de l'anthropologie cognitive. Nicolas Baumard a étudié les sciences sociales, la philosophie, la biologie et les sciences cognitives aux universités de Nantes, Aix-Marseille I, Paris VI Pierre-et-Marie-Curie et Paris IV Sorbonne ainsi qu'à l'École des Hautes Études en Sciences Sociales. Il a également été *visiting scholar* au centre *Culture and cognition* (départements de psychologie et anthropologie) de l'université du Michigan (Ann Arbor). Il anime le séminaire et le site Alphapsy consacrés aux interactions entre évolution, cognition et culture au département d'études cognitives de l'École Normale Supérieure. Outre son travail de thèse, il s'intéresse notamment aux dimensions philosophiques et scientifiques de la science-fiction et est l'un des animateurs du groupe Science-fiction-philosophie (École Normale Supérieure).

<http://www.institutnicod.org/notices.php ?user=Baumard>

