Neurosciences et droit pénal : le déterminisme peut-il sauver la conception utilitariste de la peine ?

Florian Cova
(Université de Genève)

I. Introduction

Cela fait maintenant plus d'un demi-siècle que le mot « neurosciences » est apparu et lesdites neurosciences ont connu depuis un développement ininterrompu. Les facettes de l'esprit humain qu'elles permettent d'explorer sont des plus diverses – l'attention, la mémoire, le langage[1], mais aussi les émotions ou la morale[2] – et on ne peut nier qu'elles aient déjà contribué à apporter une meilleure compréhension des opérations qui sous-tendent notre vie mentale. En résumé, les neurosciences sont en passe de nous permettre de mieux comprendre ce qu'est l'homme, ainsi que la façon dont nous pensons.

Le droit pénal a justement affaire aux hommes, à leurs actions (*actus reus*) et à leurs états mentaux (*mens rea*). Dans ce cas, ne peut-on pas s'attendre à ce que les progrès des neurosciences viennent profondément transformer le droit ? Non, répondent certains : sans nier que les neurosciences puissent venir apporter quelque chose au droit pénal, il n'en reste pas moins que tous leurs apports tombent dans des catégories déjà existantes et que la loi peut déjà prendre en compte. Autrement dit : les neurosciences ne chambouleront pas profondément le droit pénal, parce que celui-ci dispose déjà des concepts et des principes nécessaires pour prendre en compte leur apport[3].

---

[1] Voir respectivement : J.-P. Lachaux, *Le Cerveau Attentif*, Paris, Odile Jacob, 2011 ; E. Kandel, *À la Recherche de la Mémoire : Une Nouvelle Théorie de l'Esprit*, Paris, Odile Jacob, 2007 ; S. Pinker, *L'Instinct du Langage* (traduction de Marie-France Desjeux), Paris, Odile Jacob, 2008.

[2] Voir respectivement : A. Damasio, *L'Erreur de Descartes : la Raison des Emotions* (traduction de Marcel Blanc), Paris, Odile Jacob, 2010 ; J. Greene, « The secret joke of Kant's soul », in *Moral Psychology*, Vol. 3, MIT, MIT Press, 2007, pp. 35-79.

[3] Pour une défense de cette thèse, voir : S. J. Morse, « New Neuroscience, Old Problems », in *Neuroscience and the Law : Brain, Mind, and the Scales of Justice*, New York, Dana Press, 2004, pp. 157-198.

Pour mieux comprendre ce que veulent dire ceux qui défendent une telle position, prenons un exemple concret. Le 28 octobre 2009, on pouvait lire sur le site *Libération.fr* un article intitulé « Un juge italien découvre le gène du meurtre »[4]. Le fait marquant inspirant cet article était le suivant : pour la première fois, la génétique du comportement était prise en compte dans une cour de justice européenne. Plus précisément :

> « Condamné à 9 ans et deux mois de prison pour avoir poignardé un Colombien de 32 ans, à Udine, en mars 2007, Abdelmalek Bayout a vu sa peine réduite après s'être soumis à une analyse ADN "innovante". "On a découvert chez le sujet une série de gènes qui le prédisposeraient à faire preuve d'agressivité s'il venait à être provoqué ou à être exclu socialement", résume le site Internet du quotidien *Il Giornale*. Une prédisposition sociale et génétique au meurtre, l'héritage social, mais surtout, pour la première fois en Italie, le patrimoine génétique ont été reconnus par la cour d'appel de Trieste comme des circonstances atténuantes. Le soir du meurtre, la victime a agressé son meurtrier, le traitant notamment de "pédé". Des insultes qui, pour les juges, expliquent en partie la réaction disproportionnée de cet homme d'origine algérienne et musulman pratiquant ».

Contrairement à ce qu'annonce le titre de l'article, on est donc loin de l'appel à un quelconque « gène du meurtre ». L'idée semble plutôt être celle d'un gène qui favoriserait l'agressivité – voire plutôt l'impulsivité. La suite de l'article fournit plus de détails :

> « Selon une application totalement inédite de l'article 62 du code pénal italien, qui définit les circonstances atténuantes, les juges ont considéré que la réaction violente de l'accusé a été "déclenchée par le déracinement causé par la nécessité de concilier le respect de la propre foi islamique intégriste avec le mode de vie occidental". Mais, surtout, elle a été exacerbée par des éléments de son patrimoine génétique "qui, selon de nombreuses recherches internationales, augmentent de manière significative le risque de développer un comportement agressif impulsif", écrit le juge Pier Valerio Reinotti dans ses conclusions. Un héritage "sociobiologique" qui justifie alors, pour la Cour une réduction de peine d'un an ».

[4] M. Inizan, « Un juge italien découvre le gène du meurtre », publié le 28/10/2009 sur www.liberation.fr.

L'explication proposée par le juge du comportement de l'accusé est donc très éloignée du « tout génétique » annoncé par le titre de l'article. Elle est au contraire multifactorielle et combine :

1. *Un niveau sociologique :* Les croyances religieuses de l'accusé sont prises en compte.
2. *Un niveau génétique :* La constitution génétique de l'agent le rend plus susceptible de « céder » à la colère.
3. Les circonstances de l'acte.

À y regarder de plus près, l'argument utilisé par le juge est en fait un classique des tribunaux. Il consiste à dire que l'accusé n'est pas totalement responsable parce qu'il agissait de façon impulsive, sous le coup de la colère. L'argument est juste « amélioré » de deux façons : l'appel aux croyances religieuses de l'accusé permet d'expliquer pourquoi l'insulte en question était susceptible de le toucher profondément tandis que la génétique est utilisée pour sous-entendre que l'accusé était plus susceptible que la moyenne de céder à la colère. Nulle part n'est invoqué un « gène du meurtre » : le juge suppose juste que certaines personnes sont plus susceptibles que d'autres de céder à la colère et que cette susceptibilité peut être influencée par des facteurs génétiques. En résumé, Abdelmalek Bayout a vu sa peine réduite pour des raisons somme toute classiques : il a été « prouvé » qu'il avait agi de manière impulsive et n'était pas en pleine possession de ses moyens.

Cet exemple montre comment des formes de connaissances radicalement neuves peuvent avoir un intérêt pour le droit pénal tout en le laissant intact et inchangé : dans ce cas précis, la preuve génétique sert uniquement à montrer que le coupable avait agi de façon impulsive, une catégorie qui faisait déjà partie du droit pénal. Or, cet exemple est facilement transférable au cas des neurosciences : elles aussi peuvent apporter quelque chose au droit pénal en permettant de mieux comprendre l'état mental de l'agent (la *mens rea*), mais elles le feront uniquement en permettant de mieux déterminer dans quelle catégorie déjà existante tombe cet état mental.

Prenons un second exemple : aux États-Unis, le jugement de la Cour Suprême dans l'affaire *Roper contre Simmons* a conduit à l'abolition de la peine de mort pour toute personne âgée de moins de 18 ans. Or, l'utilisation d'arguments d'ordre neuroscientifique a joué un rôle reconnu dans la décision de la Cour Suprême. Ces arguments se basaient sur des données montrant que le cerveau adolescent n'était pas encore complètement mature,

ce qui pouvait fausser les décisions et rendre les adolescents plus prompts à agir de façon impulsive. Là encore, il s'agit d'un exemple dans lequel l'appel à des données neuroscientifiques joue un grand rôle, mais uniquement parce qu'elles permettent de montrer qu'une certaine catégorie d'actes tombe sous un principe déjà existant.

Ainsi, selon ces exemples, les neurosciences auraient un rôle important à jouer dans la procédure pénale, en tant que preuves de la *mens rea* (de l'état mental de l'agent), sans pour autant remettre en cause les principes actuels du droit pénal[5]. C'est cette conclusion que rejettent deux arguments qui concluent en sens inverse que les neurosciences vont ou devraient nous conduire à changer radicalement nos mœurs juridiques, arguments que cet article se donne pour but d'examiner.

II. Pourquoi les neurosciences vont (ou devraient) bouleverser le droit pénal : deux arguments

Comme nous venons de le voir, il existe des arguments solides en faveur de la thèse selon laquelle les neurosciences peuvent apporter quelque chose au droit pénal sans pour autant le bouleverser et remettre en causes ses principes, et l'usage actuel des neurosciences dans les cours de justice semble aller dans le sens de cette hypothèse. Pourtant, certains philosophes se sont opposés à cette conclusion. À l'appui de ce rejet, on peut citer deux arguments assez similaires, à cette exception près que le premier est *descriptif*, et suppose que les neurosciences *vont* bouleverser les principes fondamentaux du droit pénal, tandis que le second est *prescriptif* et affirme que les neurosciences *devraient* nous pousser à réformer les principes actuels du droit pénal.

2.1. L'argument descriptif de Greene et Cohen

Nous devons l'argument descriptif, qui est chronologiquement premier, aux philosophes Joshua Greene et Jonathan Cohen (ci-après G&C, pour simplifier)[6]. À la source de cet argument se trouve une opposition entre deux conceptions de la peine :

[5] Pour une revue des différentes façons dont les neurosciences peuvent contribuer de cette façon à l'étude de la *mens rea*, voir : E. Aharoni, C. Funk, W. Sinnott-Armstrong et M. Gazzaniga, « Can Neurological Evidence Help Courts to Assess Criminal Responsibility? Lessons from Law and Neuroscience », in *Annals of the New York Academy of Science*, 1124, 2008, pp. 145-160.

[6] J. Greene et J. Cohen, « For the Law, Neuroscience Changes Nothing and Everything », in *Philosophical Transactions of the Royal Society B: Biological Sciences*, 359/1451, 2004, p. 1775.

- Une conception *rétributive* de la peine, selon laquelle le but de la peine est de rendre au criminel ce qu'il mérite. Autrement dit : la peine n'est légitime que si le criminel est responsable d'un crime. C'est une conception tournée vers le passé : ce qui détermine la peine, c'est le mal qu'a déjà fait le criminel.
- Une conception *utilitariste* de la peine, selon laquelle le but de la peine est de prévenir de futurs crimes (par dissuasion ou réhabilitation). Autrement dit : la peine n'est légitime que si le mal qu'elle permet de prévenir est supérieur au mal qu'elle inflige. C'est une conception tournée vers le futur : ce qui détermine la peine, ce sont les conséquences qu'elle est susceptible d'avoir.

Selon G&C, nos systèmes pénaux reposent (implicitement ou non) sur une conception rétributive de la peine : par exemple, nous considérons que la peine doit être proportionnée au crime et que le facteur critique est la responsabilité de celui dont on doit décider la condamnation. Plus largement, G&C pensent que nos systèmes pénaux sont intrinsèquement rétributifs parce que nous avons pour la plupart d'entre nous des intuitions rétributives. C'est là la première prémisse de leur argument :

> (1) Nous avons des intuitions rétributives au sujet de l'attribution de peines.

Or, selon G&C, la conception rétributive de la peine est fortement liée à la notion de *responsabilité morale*. En effet, lorsque l'on dit qu'un criminel *mérite* d'être puni, cette notion de « mérite » transporte avec elle un certain nombre de suppositions, et en particulier celle selon laquelle le criminel en question est personnellement responsable de ce qu'il a fait, et a agi librement. Voici donc la seconde prémisse de leur argument :

> (2) La conception rétributive de la peine est liée à l'idée de responsabilité : selon elle, un agent ne saurait mériter de peine pour une action sans être responsable de cette action.

De plus, selon G&C, les notions de responsabilité morale et de liberté présupposées par la conception rétributive de la peine sont incompatibles avec le déterminisme, même entendu au « sens faible » (c'est-à-dire au sens

où chacune de nos pensées et de nos actions est entièrement causée par des facteurs qui nous sont extérieurs)[7] :

> (3) Notre conception intuitive de la liberté et de la responsabilité morale implique que celles-ci sont incompatibles avec le déterminisme.

Or, nous disent G&C, les neurosciences nous apprennent que nos pensées et notre comportement sont entièrement déterminés par des facteurs qui nous sont extérieurs :

> (4) Les neurosciences montrent (et vont finir par nous convaincre) que toutes nos pensées et tous nos actes sont déterminés.

Dans le cadre de cet article, nous admettrons cette prémisse sans la discuter. On peut bien sûr se demander si les neurosciences montrent vraiment que nos pensées et nos actes sont déterminés. Un argument général en faveur de cette conclusion consiste à dire que nous savons (par la physique) que les états physiques sont déterminés (au sens faible), que les neurosciences montrent que nos états mentaux sont des états physiques (des états du cerveau), et donc que les neurosciences montrent que nos états mentaux sont déterminés.

D'autres arguments, peut-être rhétoriquement plus puissants, mais logiquement moins satisfaisants, consistent à faire appel à certaines expériences neuroscientifiques surprenantes. La plus connue est la fameuse expérience de Libet[8]. Dans cette expérience, les participants avaient pour consigne d'accomplir un acte moteur simple (appuyer sur un bouton, plier un doigt, etc.). De plus, ils avaient devant eux une horloge sur laquelle tournait, en guise d'aiguille, un point. Les participants avaient pour

[7] Quelques précisions : par déterminisme « au sens fort », nous entendons la thèse selon laquelle la description de tout état du monde à un moment donné est une conséquence logique de la conjonction des lois qui gouvernent le monde et de la description d'un état antérieur de ce même monde. Autrement dit : au « sens fort », le déterminisme est la thèse selon laquelle tout ce qui se produit dans le monde est entièrement causé par l'état du monde qui précède *plus* la thèse selon laquelle une même cause a nécessairement les mêmes effets. Cette version du déterminisme entre en conflit avec certaines interprétations actuelles de la physique quantique, selon lesquelles les lois de la nature sont probabilistes, et donc pour lesquelles une même cause peut avoir des effets différents. Il n'en reste pas moins que, même selon ces interprétations, la physique quantique reste déterministe « au sens faible », étant donné qu'elle suppose que l'état actuel d'un système physique est entièrement dû, causé et explicable par l'état antérieur de ce système.

[8] B. Libet, « Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action », in *Behavioral and Brain Sciences*, 8/4, 1985, pp. 529-539.

consigne de noter la position du point sur l'écran au moment où ils sentiraient qu'ils avaient « l'intention » d'accomplir l'action en question (par exemple : plier le doigt). Ainsi, après soustraction du délai mis par les participants pour noter la position du point, il était possible de déterminer à quel moment (en moyenne) ceux-ci avaient ressenti l'intention de plier le doigt. Les résultats obtenus par le psychologue Benjamin Libet montrent que, en moyenne, 200 millisecondes s'écoulaient entre l'intention ressentie et le moment où le doigt se pliait.

Ces résultats n'ont rien de très étonnant. Mais il y a plus : pendant la tâche, Libet enregistrait aussi l'activité cérébrale des participants au moyen d'un électroencéphalographe. L'aire cérébrale d'intérêt était le cortex moteur secondaire : en effet, lorsqu'un mouvement corporel est programmé, l'activité cérébrale augmente dans cette aire, rendant possible de déterminer le moment auquel un acte moteur est initialisé. Or, ce qu'a découvert Libet, c'est que le mouvement (plier le doigt) commençait à être programmé 500 millisecondes avant d'être réalisé, et donc que la programmation du mouvement prenait place 300 millisecondes *avant* que les participants ne ressentent l'intention de plier le doigt.

Autrement dit : lorsque nous ressentons l'intention d'accomplir une action, notre cerveau l'a déjà mise en route. Nous y reviendrons plus loin, mais nombreux sont ceux à penser que ces résultats montrent que nos intentions ne jouent pas de rôle causal dans nos actions, et donc que nous sommes le jouet de forces extérieures (en l'occurrence, notre cerveau).

Admettons donc que les neurosciences montrent que nos pensées et nos actions sont déterminées. De cette prémisse (4), et de (3) pris ensemble, il en résulte une première conclusion :

> (C1) Les neurosciences vont montrer que nous ne sommes ni libres, ni responsables de nos actes, au sens intuitif que nous donnons à ces termes.

De (C1) et de (2) pris ensemble, il en résulte que :

> (C2) Les neurosciences vont montrer que la conception rétributive de la peine, parce qu'elle repose sur une conception de la responsabilité qui est irréaliste, est inadéquate.

Finalement, de (C2) et (1) pris ensemble, G&C concluent :

> (C3) Puisque les neurosciences vont nous conduire à rejeter la conception rétributive de la peine, et que celle-ci est notre conception intuitive et guide nos principes juridiques actuels, alors les neurosciences vont grandement changer le droit pénal.

Mais G&C ne s'arrêtent pas là : ils ont aussi une idée de la conception qui viendra remplacer la conception rétributive de la peine. En effet, G&C semblent supposer la chose suivante :

> (5) Abandonner la conception rétributive de la peine nous conduira à adopter la seule conception de la peine compatible avec le déterminisme : la conception utilitariste de la peine.

Ce qui nous conduit à leur conclusion finale :

> (C4) Les neurosciences vont nous conduire à changer notre conception rétributive de la peine contre une conception utilitariste de la peine.

Ainsi, selon G&C, le progrès des neurosciences et leur prise en compte progressive nous conduiront à abandonner la conception rétributive de la peine pour une conception utilitariste de la peine.

2.2. L'argument prescriptif de Xavier Bébin

Dans un ouvrage postérieur à l'article de G&C, Xavier Bébin[9] (ci-après XB) reprend leur argument mais en propose une version normative : il ne s'agit plus de savoir pourquoi les neurosciences vont nous conduire à changer de conception de la peine, mais de savoir pourquoi les neurosciences montrent que nous *devons* changer de conception de la peine (pour embrasser la conception utilitariste de la peine).

XB accepte les mêmes prémisses (1), (2) et (4) que G&C : c'est-à-dire (1) que nous avons des intuitions rétributives au sujet de la peine, (2) que ces intuitions rétributives supposent la liberté et (4) que les neurosciences montrent que nos pensées et nos actes sont déterminés (au sens faible, ce qu'il appelle la « thèse causale »).

[9] X. Bébin, *Pourquoi punir ? L'approche utilitariste de la sanction pénale*, Paris, L'Harmattan, 2006. L'argument (qui n'est pas construit de cette façon mais que nous reconstituons pour faciliter la discussion) se trouve au Chapitre 4.

Les différences cruciales avec l'argument de G&C se trouvent au niveau des prémisses (3) et (5). En effet, XB se soucie peu de savoir ce que les gens pensent au sujet de la liberté, ce qui compte, c'est ce qui est vrai. Or, selon XB, la liberté est incompatible avec le déterminisme :

> (3*) La liberté et la responsabilité morale sont incompatibles avec le déterminisme.

. De cette prémisse (3*) et de (4) pris ensemble, il en résulte une première conclusion :

> (C1*) Les neurosciences montrent que nous ne sommes ni libres, ni responsables de nos actes.

De (C1*) et de (2) pris ensemble, il en résulte que :

> (C2*) Les neurosciences montrent que la conception rétributive de la peine, parce qu'elle repose sur une conception de la responsabilité qui est irréaliste, est inadéquate.

Finalement, de (C2*) et (1) pris ensemble, XB conclue :

> (C3*) Puisque les neurosciences nous conduisent à rejeter la conception rétributive de la peine, et que celle-ci est notre conception intuitive et guide nos principes juridiques actuels, alors nous devons grandement changer le droit pénal.

Mais, comme G&C, XB ne s'arrête pas là : lui aussi est un partisan de la conception utilitariste de la peine. C'est pourquoi il faut ajouter la prémisse suivante :

> (5*) Dans un cadre déterministe, la seule conception viable de la peine est la conception utilitariste

Ce qui nous conduit à la conclusion finale, parallèle mais non identique à celle de G&C :

> (C4*) Les découvertes des neurosciences ont pour conséquence qu'il est plus raisonnable de rejeter la

> conception rétributive de la peine pour adopter la conception utilitariste.

Autrement dit, si nous prenons sérieusement en compte les enseignements des neurosciences, il nous faut abandonner la conception rétributive de la peine pour adopter la conception utilitariste, et réformer le droit pénal en conséquence. Tant l'argument de G&C que celui de XB concluent à l'inévitabilité (descriptive ou normative) d'une réforme du droit pénal face aux données neuroscientifiques. Mais cette conclusion est-elle correcte ? C'est pour le déterminer qu'il nous faudra examiner de près chacune des étapes de ces arguments.

III. Sur la prémisse (1) : que nous avons des intuitions rétributivistes

Avant d'examiner en détail ces deux arguments, attardons-nous sur une prémisse purement descriptive qu'ils partagent tous deux, c'est-à-dire la première :

> (1) Nous avons des intuitions rétributives au sujet de l'attribution de peines.

Cette prémisse est probablement la plus solide de celles que nous allons examiner. En effet, un grand nombre de données semblent la confirmer. Soit l'opposition entre les deux grands types de conception de la peine, qui peut être développée de la manière suivante :

- *La conception rétributive de la peine :* On punit parce que le coupable « mérite » d'être puni. La peine est alors sa propre fin et constitue un devoir et un bien en soi. Pour citer Kant, grand représentant de cette position : la peine « ne peut jamais être considérée simplement comme un moyen de réaliser un autre bien, soit pour le criminel lui-même, soit pour la société civile, mais doit uniquement lui être infligée pour la seule raison qu'il a commis un crime »[10].

[10] Pour illustrer sa conception de la punition et bien la distinguer de la conception utilitariste, Kant propose l'expérience de pensée suivante : dans une société établie sur une île, un individu est condamné à mort. Suite à certaines circonstances, la société décide de quitter l'île pour aller vivre ailleurs. La question se pose alors de savoir s'il faut libérer le prisonnier et le laisser se débrouiller seul ou bien l'exécuter avant de partir. Pour Kant, la morale nous impose de l'exécuter.

- *La conception utilitariste de la peine:* On punit pour éviter de nouveaux crimes, soit de la part du coupable lui-même soit de la part d'autres personnes. La peine n'a aucune valeur en soi et constitue même un mal (bien que ce mal puisse être nécessaire). En reprenant la citation de Kant, on peut distinguer deux grandes variétés de peine utilitaristes : les peines *dissuasives*, qui visent à bénéficier à « la société civile » en dissuadant le criminel de récidiver ou d'autres personnes de perpétrer les mêmes exactions, et les peines *correctrices* qui visent à bénéficier au « criminel lui-même » en le rendant meilleur (mais elles bénéficient aussi, indirectement, à la société civile).

Ces deux conceptions de la peine ne conduisent pas à prendre en compte les mêmes facteurs lorsqu'il s'agit de déterminer le montant de la punition qui doit être infligée à un coupable. Selon la conception utilitariste de la peine, le montant de la peine est déterminé par son efficacité : il faut augmenter la peine jusqu'à ce qu'elle permette d'éviter des récidives. Ainsi, les facteurs pertinents seront :

1. *la probabilité que le coupable récidive*, car plus cette probabilité est haute, plus la peine devra être élevée pour dissuader le coupable de façon efficace,
2. *la détectabilité du type de méfait commis par le coupable*, car moins un méfait est détectable, plus la peine devra être élevée pour être dissuasive,
3. *le caractère public de la peine*, car en effet une peine rendue publique est plus efficace pour décourager les potentiels malfaiteurs qu'une peine privée dont personne n'a connaissance – il conviendra donc de préférer les peines publiques.

Selon la conception rétributive de la punition, le montant de la punition est déterminé par la quantité de mal infligé à autrui dont le coupable est responsable. De ce fait, les facteurs pertinents seront :

1. *la gravité du méfait :* plus un méfait sera grave, plus le mal commis sera élevé, plus la peine devra l'être, car elle doit être proportionnelle au mal infligé,
2. *les circonstances atténuantes :* la peine sera moins sévère si le

coupable a des circonstances atténuantes.[11]

Se demander quelle conception de la peine acceptent les gens au niveau intuitif revient ainsi à se demander quels sont les critères qu'ils prennent implicitement en compte lorsqu'ils jugent du bien-fondé (ou non) d'une peine. Or, les gens semblent ignorer les facteurs utilitaristes pour ne prendre en compte que les facteurs rétributifs. Ainsi, Carlsmith et ses collègues[12] ont donné à lire à leurs participants plusieurs variations sur des scénarios dans lesquels le personnage principal commettait un méfait. Chaque scénario existait dans 4 versions, car 2 facteurs étaient variés : soit le crime était grave, soit il ne l'était pas (facteur *rétributif*) et soit le crime était difficile à détecter, soit il ne l'était pas (facteur *utilitariste*). Les participants devaient ensuite assigner une peine en années de prison au personnage principal. Les résultats ont montré que les participants variaient la sévérité de la peine en fonction de la gravité du crime, mais pas en fonction de sa détectabilité. Dans d'autres expériences sur le même modèle, les facteurs à varier pouvaient être différents (publicité ou non de la peine pour les facteurs utilitaristes, présence ou non de circonstances atténuantes pour les facteurs rétributifs), mais les résultats restaient les mêmes : les participants modulaient la peine en fonction des facteurs rétributifs et pas en fonction des facteurs utilitaristes.

Une autre série d'expériences menées par Sunstein et ses collègues[13] conduit aux mêmes conclusions. Comme les expériences précitées, la première de leurs expériences montre que les participants ne prennent pas en compte la détectabilité du crime lorsqu'il s'agit d'attribuer une punition. Dans leur seconde expérience, ils ont soumis à 84 étudiants de la University of Chicago Law School un scénario dans lequel il était expliqué que, parce que le nombre d'agents du fisc par individu était beaucoup plus petit en

[11] Une conception rétributive de la peine se doit de prendre en compte les circonstances atténuantes. Cependant, cela ne signifie aucunement qu'une conception utilitariste ne peut pas aussi se soucier de la présence ou non de circonstances atténuantes : elle le pourrait, dans la mesure où la crainte constante d'être puni pour un mal dont on n'est pas responsable peut être considérée comme un mal à minimiser. Néanmoins, il n'est pas nécessaire logiquement qu'elle le fasse : une législation qui punirait aussi ceux qui ont des circonstances atténuantes serait tout aussi dissuasive. En conclusion, une expérience montrant que les gens se soucient des circonstances atténuantes ne permettrait pas d'en conclure qu'ils n'ont pas une conception utilitariste de la peine, mais une expérience montrant que les gens *ne* se soucient *pas* des circonstances atténuantes permettrait de rejeter l'hypothèse selon laquelle ceux-ci ont une conception rétributive de la peine.

[12] K. Carlsmith, J. Darley et P. Robinson, « Why do we punish? Deterrence and just deserts as motives for punishment », in *Journal of Personality and Social Psychology*, 83/2, 2002, pp. 284-299.

[13] C. Sunstein, D. Schkade et D. Kahneman, « Do people want optimal deterrence? », in *The Journal of Legal Studies*, 29/1, 2000, pp. 237-253.

Utah qu'en Californie, la fraude fiscale était moins risquée en Utah. La suite de ce scénario expliquait que, parce qu'il était impossible d'engager plus d'agents en Utah, le fisc avait décidé d'augmenter le montant des peines pour fraude fiscale en Utah tout en le laissant au même niveau en Californie, une décision logique selon la conception utilitariste de la peine, mais injuste selon la conception rétributive. La plupart des participants se sont opposés à cette décision.

De la même façon, certaines expériences menées par Baron et Ritov[14] montrent que les gens ne se soucient pas principalement des conséquences des peines qu'ils attribuent : ils punissent avant tout parce que le coupable mérite d'être puni. Dans un de leurs scénarios, une entreprise qui produit des vaccins se voit intenter un procès parce qu'un enfant est mort à cause de leur vaccin. Dans une première version du scénario, il était expliqué aux participants que punir l'entreprise (en leur demandant de payer une amende) aurait un effet « pervers » : l'entreprise arrêterait du même coup de produire ce vaccin. Dans une seconde version, ils étaient informés que la peine aurait un effet « positif » et que la punition conduirait l'entreprise à remplacer son vaccin par un vaccin plus sûr. Les participants devaient lire les deux versions du scénario puis dire si, selon eux, la peine devait être plus élevée dans un cas que dans l'autre ou égale dans les deux cas. La plupart des participants ont répondu que la peine devait être la même dans les deux cas. De la même façon, une autre paire de scénarios contrastait un cas où la peine était rendue publique avec un cas où elle ne l'était pas. Encore une fois, la plupart des participants n'ont fait aucune différence entre les deux cas.

Pour finir, citons une expérience de Carlsmith[15] permettant de révéler quels facteurs les participants considèrent comme étant les plus importants dans l'assignation d'une peine. Les participants recevaient un cas décrivant un criminel auquel ils devaient attribuer une peine. Seulement, ils ne disposaient pas de toutes les informations au sujet de ce cas. Neuf informations supplémentaires leur étaient accessibles, mais ils ne pouvaient en choisir que cinq. Par exemple, ils pouvaient choisir de connaître la gravité du crime commis, les intentions du criminel (facteurs *rétributifs*), la fréquence de ce type de crimes ou encore la probabilité d'une récidive (facteurs *utilitaristes*). L'analyse des informations choisies par les participants et l'ordre dans lequel celles-ci étaient choisies permettaient ainsi de déterminer quels facteurs ces participants considéraient comme le

[14] J. Baron et I. Ritov, « Intuitions about penalties and compensation in the context of tort law », in *Journal of Risk and Uncertainty*, 7/1, 1993, pp. 17-33.

[15] K. Carlsmith, « The roles of retribution and utility in determining punishment », in *Journal of Experimental Social Psychology*, 42/4, 2006, pp. 437-451.

plus important. Une fois de plus, les participants se sont avant tout intéressés aux informations portant sur des facteurs rétributifs.

Ces quelques résultats montrent que, comme le supposent G&C et XB, nous adoptons une conception rétributive de la peine au niveau intuitif. Cela ne nous empêche pas pour autant d'accepter aussi la conception utilitariste au niveau réflexif. En fait, une étude menée (une fois de plus) par Carlsmith[16] suggère qu'il existe un écart entre la conception de la peine que les gens adoptent « dans l'abstrait » (au niveau réflexif) et celle qu'ils adoptent quand ils sont confrontés à des cas particuliers (au niveau intuitif). Ainsi, Carlsmith a demandé dans un premier temps aux participants d'attribuer une peine à des agents dont les crimes étaient décrits dans de courts scénarios. D'un scénario à l'autre, divers facteurs étaient variés, certains étant pertinents d'un point de vue utilitariste et d'autres d'un point de vue rétributif. Dans un second temps, les participants étaient introduits aux deux conceptions de la peine et devaient dire (i) à quel point ils étaient d'accord avec chacune d'entre elles et (ii) si ces conceptions avaient fait partie de leurs motivations lorsqu'il s'agissait d'attribuer des peines dans la première partie de l'expérience. Les résultats obtenus par Carlsmith montrent que, bien que la plupart des participants déclarent pour la plupart être d'accord avec la conception utilitariste de la peine et être guidés par elle dans leurs attributions de peine, il n'en est en fait rien. Autrement dit : il n'y a pas de corrélation significative entre les conceptions de la peine que les participants déclarent adopter et les facteurs qui guident réellement leurs attributions de peine dans des cas particuliers.

IV. Examen critique de l'argument descriptif : les neurosciences vont-elle réellement nous conduire à abandonner la conception rétributive de la peine ?

Le bien-fondé de la prémisse (1) étant établi, il est temps d'examiner les autres prémisses pour évaluer la valeur des arguments. Nous allons dans un premier temps examiner l'argument descriptif de G&C, avant de passer à l'argument normatif de XB.

4.1. Examen de la prémisse (2) : la punition peut suivre un schéma rétributif sans faire appel au concept de mérite

Passons à l'examen de la prémisse (2) :

[16] K. Carlsmith, « On justifying punishment: The discrepancy between words and actions », in *Social Justice Research*, 21/2, 2008, pp.119-137.

> (2) La conception rétributive de la peine est liée à l'idée de responsabilité : selon elle, un agent ne saurait mériter de peine pour une action sans être responsable de cette action.

En un sens, cette prémisse est trivialement vraie : si on définit la conception rétributive de la peine comme une conception pour laquelle celui qui est puni *mérite* ce qui lui arrive, alors cette conception est évidemment liée à l'idée de responsabilité. Mériter une punition, c'est être responsable d'un certain mal.

Mais si maintenant, on caractérise la conception rétributive de la peine non par la notion de mérite, mais par ses caractéristiques extérieures (le fait qu'elle varie en fonction de la gravité du crime et qu'elle soit principalement tournée vers le passé et ce qui a été fait par le coupable), il semble possible de sauvegarder une grande partie de la conception rétributive de la peine tout en sacrifiant la notion de mérite.

En effet, il existe une autre notion, tout aussi importante, liée à la conception rétributive de la peine : celle de *compensation*. C'est l'idée selon laquelle la punition vise à rétablir l'équilibre, à compenser le mal qui a été fait, en la réparant par un autre mal. C'est une conception de la peine que l'on retrouve déjà chez Aristote (dans *l'Ethique à Nicomaque*, Livre V, section 7) mais qui semble aussi transparaître dans nos pratiques linguistiques. Par exemple, du criminel qui sort de prison, on dit qu'il a « payé sa dette à la société », alors même que, les travaux forcés étant abolis, son internement a coûté de l'argent à ladite société sans rien rapporter en retour. Plus généralement, il est courant de parler du châtiment comme un moyen de « faire payer » le coupable.

Cette notion de la peine comme réparation, comme retour à l'équilibre, se double souvent dans la conception populaire de l'idée selon laquelle la peine est un moyen d'apaiser la victime tourmentée. Nombre d'histoires de fantômes tournent autour du thème du mort dont l'âme ne sera apaisée que lorsque ses meurtriers seront punis. On entend aussi souvent dire que punir le coupable est une façon de « rendre la paix » à ses victimes.

Les deux notions s'impliquent d'ailleurs l'une l'autre : si la peine apporte une certaine satisfaction à la victime, alors on comprend comment elle peut constituer un « retour à l'équilibre » en apaisant et compensant le mal par le « plaisir » tiré de savoir le coupable puni. En ce sens, la peine est avant tout quelque chose qui est dû à celui qui a souffert, un devoir envers *lui*. Comme l'écrit Maurice Cusson, parlant des sociétés dans lesquelles les membres d'une famille ont pour obligation de venger les leurs :

> « L'obligation de venger un parent assassiné était considérée comme un devoir envers le disparu et comme un moyen de l'apaiser [...] En somme, la vengeance est une obligation qui s'enracine dans la solidarité familiale et le respect du mort. "Le devoir de vengeance est l'extension de l'aide mutuelle qu'on se doit entre parents" »[17].

On obtient ainsi une conception rétributive de la punition comme réparation et retour à l'ordre[18], conception qui peut survivre à la disparition du concept de mérite. Si ce qui compte c'est le retour à l'ordre antérieur, en « faisant payer » celui qui a fait du mal, alors peu importent la liberté et la responsabilité. On peut vouloir restaurer l'ordre même si les coupables ne sont pas responsables de leurs actes. Ainsi, il se pourrait qu'une forme « quasi-cosmique » de conception de la peine comme restauration de l'harmonie survive à l'idée de responsabilité. Autrement dit, montrer que les neurosciences rendent obsolète notre conception de la responsabilité morale ne suffit pas à montrer qu'elle mettent hors-jeu la conception rétributive de la punition.

4.2. Examen de la prémisse (3) : notre concept de mérite ne fait pas appel à une conception incompatibiliste de la liberté

Mais admettons la prémisse (2) pour le moment et passons maintenant à la prémisse suivante, soit la prémisse (3) :

> (3) Notre conception intuitive de la liberté et de la responsabilité morale implique que celles-ci sont incompatibles avec le déterminisme.

En termes techniques, on résumera cette prémisse de la façon suivante : nous avons des intuitions *incompatibilistes*. *L'incompatibilisme* est la thèse philosophique selon laquelle liberté et responsabilité morale sont incompatibles avec le déterminisme : si nos actions sont déterminées, c'est-à-dire causées entièrement par des facteurs qui nous sont extérieures, alors nous ne pouvons être libres et responsables de nos actes.

[17] M. Cusson, *Pourquoi punir ?*, Paris, Dalloz, 1987.
[18] Nicolas Baumard défend la thèse selon laquelle il s'agit là de la conception naturelle et intuitive de la peine. Voir : N. Baumard, « Punishment is not a group adaptation, Humans punish to restore fairness rather than to support group cooperation », in *Mind and Society*, 10/1, 2011, pp.1-26.

L'incompatibilisme s'oppose tout naturellement au *compatibilisme*, la thèse selon laquelle la liberté et la responsabilité sont compatibles, sous certaines conditions, avec le déterminisme.

G&C supposent que nous sommes « naturellement incompatibilistes », c'est-à-dire que notre conception intuitive de la liberté est incompatibiliste : nous tendons naturellement à penser que la responsabilité ne peut cohabiter avec le déterminisme. Mais est-ce réellement le cas ? Le sens commun pense-t-il réellement qu'un agent ne peut pas être responsable de ses actes dès lors que ces actes sont entièrement déterminés (ou causés) par des événements qui ont précédé sa naissance et se trouvent donc hors de son contrôle ?

- 4.2.1. Données suggérant que nous sommes naturellement compatibilistes

Le philosophe Eddy Nahmias a réalisé un certain nombre d'expériences destinées à trouver la réponse à cette question[19]. Dans une première expérience, Nahmias et ses collègues ont donné à lire à plusieurs participants le scénario suivant :

> *Superordinateur* – Imaginons qu'au siècle prochain nous découvrions toutes les lois de la nature et que nous mettions au point un superordinateur capable de déduire à partir de l'ensemble de ces lois de la nature et de l'état actuel du monde ce qui se produira à n'importe quel moment dans le futur. Ce superordinateur a la possibilité de voir tout ce qui se passe actuellement dans le monde et de prévoir avec un taux de succès de 100% ce qui s'y produira par la suite. Supposons donc qu'un tel superordinateur existe et qu'il enregistre l'état du monde à un moment donné, le 25 mars 2150, soit 20 ans avant la naissance de Jérémy Hall. De ces informations et des lois de la nature prises ensemble, le superordinateur en déduit que Jérémy braquera la Banque de la Fidélité à 6 heures du matin le 26 juin 2195. Comme toujours, le superordinateur voit juste et Jérémy braque la Banque de la Fidélité à 6 heures du matin le 26 juin 2195.

Les participants devaient répondre à l'une des deux questions suivantes : « Jérémy a-t-il agi librement (*of his own free will*)? » ou « Jérémy mérite-t-il d'être blâmé pour avoir braqué la banque? ». À la première question, 76% des participants ont répondu « oui », et 83% à la seconde. Dans une version

[19] E. Nahmias, S. Morris, T. Nadelhoffer et J. Turner, « Surveying freedom: Folk intuitions about free will and moral responsibility », in *Philosophical Psychology*, 18/5, 2005, pp. 561-584 ; E. Nahmias, S. Morris, T. Nadelhoffer et J. Turner, « Is incompatibilism intuitive? » in *Philosophy and Phenomenological Research*, 73/1, 2006, pp. 28-53.

moralement positive de ce même scénario, Jérémy sauvait un enfant en train de se noyer au lieu de braquer une banque : 68% des participants ont répondu que, dans ce cas, Jérémy avait agi librement (*of his own free will*) et 88% qu'il méritait des louanges. Enfin, dans une version neutre de ce scénario, 79% des participants ont répondu que Jérémy était allé faire du jogging librement. Ces réponses semblent indiquer que le sens commun a plutôt des intuitions compatibilistes.

Dans une seconde expérience du même type, le scénario proposé était le suivant :

> *Les Jumeaux Séparés* – Imaginez un monde dans lequel les croyances et les valeurs de chaque personne sont causées entièrement par la combinaison des gènes et de l'environnement de cette personne. Un jour, dans ce monde, deux jumeaux nommés Fred et Barney sont mis au monde puis abandonnés par leur mère. Fred est adopté par les Jerksons tandis que Barney est adopté par les Kindersons. Dans le cas de Fred, ses gènes combinés à l'éducation qu'il a reçue de l'égoïste famille Jerksons le conduisent à donner à l'argent plus de valeur qu'à tout le reste et à croire qu'il est permis de tout faire pour amasser de l'argent. Dans le cas de Barney, ses gènes (identiques à ceux de Fred) et l'éducation qu'il a reçue de l'adorable famille Kinderson le conduisent à placer l'honnêteté au-dessus de tout le reste et à croire qu'il faut toujours respecter le bien d'autrui. Fred et Barney sont tous deux des individus intelligents capables de délibérer au sujet de leurs actions.
>
> Un jour, Fred et Barney tombent chacun de leur côté sur un portefeuille contenant 1000$ et la carte d'identité du propriétaire du portefeuille. Chacun d'entre eux est sûr que personne ne peut le voir. Après délibération, Fred Jerkson, à cause de ses croyances et de ses valeurs, décide de garder l'argent. Après délibération, Barney Kinderson, à cause de ses croyances et de ses valeurs, renvoie le portefeuille à son propriétaire.
>
> Étant donné que, dans ce monde, les gènes et l'environnement de chaque personne causent entièrement les croyances et les valeurs de cette personne, alors il est vrai que, si Fred avait été adopté par les Kinderson, il aurait eu des croyances et des valeurs qui l'auraient conduit à renvoyer le portefeuille; et si Barney avait été adopté par les Jerksons, il aurait eu des croyances et des valeurs qui l'auraient conduit à garder le portefeuille.

Après avoir lu ces scénarios, 76% des participants ont répondu à la fois que Fred avait librement (*of his own free will*) gardé le portefeuille et que

Barney avait librement rendu le portefeuille. De plus, 60% des participants ont répondu que Fred méritait d'être blâmé pour son acte et 64% que Barney méritait d'être loué, ce qui va dans le sens du compatibilisme.

Enfin, dans une troisième expérience, les participants devaient lire le scénario suivant :

> *L'Eternel Retour* – Imaginez un univers qui est recréé encore et encore, à partir à chaque fois des mêmes conditions initiales et avec les mêmes lois de la nature. Dans cet univers, les mêmes conditions et les mêmes lois de la nature produisent exactement les mêmes effets et donc, à chaque fois que cet univers est recréé, tout doit se produire exactement de la même façon. Par exemple, dans cet univers, une personne nommée Jill décide de voler un collier à un certain moment, et, chaque fois que l'univers est recréé, Jill décide de voler un collier à ce même moment.

Dans ce cas, 66% des participants ont jugé que Jill avait « librement » décidé de voler le collier et 77% qu'il était juste de la tenir moralement responsable pour sa décision de voler le collier. Dans l'ensemble, ces résultats vont directement à l'encontre de la prémisse 4 de G&C et suggèrent que nous avons plutôt des intuitions compatibilistes.

- 4.2.2. Données suggérant que nous sommes naturellement incompatibilistes

Néanmoins, un autre ensemble d'expériences semble cependant apporter un soutien plus solide à la thèse selon laquelle nous serions « naturellement » incompatibilistes. Dans une série d'expériences, Shaun Nichols et Joshua Knobe ont donné à leurs participants le scénario suivant, qui décrit deux univers différents[20] :

> Imaginez un univers (appelons-le l'Univers A) dans lequel tout ce qui se produit est entièrement causé par tout ce qui s'est produit auparavant. Cela est vrai depuis le tout premier commencement de cet univers, de telle sorte que ce qui s'est produit au tout début a causé ce qui s'est produit juste après et ainsi de suite jusqu'à aujourd'hui. Par exemple : un jour John décide de manger des frites. Comme tout le reste, cette décision est entièrement causée par ce qui s'est produit auparavant. Ainsi, si tout dans cet univers jusqu'à la décision de John est identique, alors il doit forcément arriver que John décide de manger des frites.

[20] S. Nichols et J. Knobe, « Moral responsibility and determinism: The cognitive science of folk intuitions », in *Nous*, 41/4, 2007, p.663.

> Imaginez maintenant un univers (appelons-le l'Univers B) dans lequel presque tout ce qui se produit est entièrement causé par tout ce qui s'est produit auparavant. La seule exception, ce sont les décisions humaines. Par exemple : un jour Marie décide de manger des frites. Puisque dans cet univers les décisions de chaque personne ne sont pas entièrement causées par ce qui a précédé, alors, même si tout est identique jusqu'à la décision de Marie, Marie n'est pas obligée de manger des frites. Elle aurait pu décider de manger quelque chose d'autre.
> La différence clé est donc que dans l'Univers A, toute décision est entièrement causée par ce qui a précédé la décision de telle sorte que, étant donné les événements passés, chaque décision devait forcément se produire de la façon dont elle s'est produite. Par contraste, dans l'Univers B, les décisions ne sont pas complètement causées par les événements passés et ne devaient pas forcément se produire de la façon dont elles se sont produites.

En premier lieu, les participants devaient dire lequel de ces deux univers ressemblait le plus, selon eux, à l'univers dans lequel nous vivons. Presque tous (c'est-à-dire 90%) ont répondu l'Univers B (soit l'univers *indéterministe*). Ensuite, les participants placés dans la *situation concrète* devaient répondre à la question suivante :

> Dans l'Univers A, un homme appelé Bill se sent attiré par sa secrétaire et décide que la seule façon d'être avec elle est de tuer sa femme et ses trois enfants. Il sait qu'il est impossible de s'échapper de sa maison en cas d'incendie. Avant de partir en voyage d'affaires, il installe dans sa cave un appareil qui met le feu à sa maison et carbonise sa famille. Bill est-il pleinement responsable du meurtre de sa femme et de ses enfants ?

Dans cette condition, 72% des participants ont répondu oui et ont donc donné une réponse *compatibiliste*, ce qui semble être cohérent avec les résultats obtenus par Nahmias et ses collègues. Mais considérons maintenant la *situation abstraite*. Dans cette situation, les participants devaient répondre à la question suivante :

> Dans l'Univers A, est-il possible pour une personne d'être pleinement responsable de ses actions ?

Dans cette situation, la plupart des participants (86%) ont donné la réponse *incompatibiliste* (« NON »). Il y a donc une contradiction entre les réponses

des participants dans la situation concrète et leur réponse dans la situation abstraite. Comment est-ce possible ? Une première solution serait que la situation concrète est trop longue et complexe, ce qui ferait oublier aux sujets que l'agent est dans un univers déterministe. Pour tester cette hypothèse, Nichols et Knobe ont donné à d'autres participants une version plus courte de la situation concrète :

> Dans l'Univers A, Bill tue sa femme et ses enfants à coups de poignard pour être avec sa secrétaire. Est-il possible pour Bill d'être pleinement responsable du meurtre de sa famille ?

Dans ce cas, 50% des participants ont donné la réponse compatibiliste, ce qui est moins que dans la première situation concrète, mais tout de même plus que dans la situation abstraite (une différence de 36%). Comment expliquer qu'environ un tiers des sujets semble changer de réponse entre ces deux conditions ?

On pourrait envisager l'hypothèse selon laquelle seule la situation concrète teste réellement les intuitions des gens, en leur demandant de réagir sur un cas particulier, tandis que la situation abstraite teste leur théorie explicite sur la responsabilité morale, qui ne s'accorde pas nécessairement avec leurs intuitions. Mais ce n'est pas la piste que favorisent Nichols et Knobe. Leur hypothèse est la suivante : tandis que les réponses compatibilistes sont le résultat de processus émotionnels, les réponses incompatibilistes sont le fruit de raisonnements abstraits. La situation concrète décrivant un meurtre atroce, elle aurait plus tendance à déclencher une réaction émotionnelle que la situation abstraite, entraînant ainsi plus de réponses compatibilistes. Pour tester cette hypothèse, Nichols et Knobe ont mis au point deux autres situations. La situation *émotionnellement faible* était la suivante :

> Comme il l'a déjà fait de nombreuses fois, Marc s'arrange pour tricher sur sa déclaration d'impôt. Est-il possible pour Marc d'être pleinement responsable du fait de tricher sur sa déclaration d'impôt ?

Tandis que la situation *émotionnellement forte* était la suivante :

> Comme il l'a déjà fait de nombreuses fois, Bill suit et viole une inconnue. Est-il possible pour Bill d'être pleinement responsable du fait de violer une inconnue ?

Dans chaque situation, une moitié des sujets était informée que l'action se déroulait dans l'Univers A et l'autre moitié que l'action se déroulait dans l'Univers B. La tableau ci-dessous décrit, pour chaque combinaison possible, le nombre de participants à avoir répondu oui.

Situation	*Univers A*	*Univers B*
Émotionnellement forte	64%	95%
Émotionnellement faible	23%	89%

Résultats de Nichols et Knobe (2007)

Lorsque l'action se situait dans l'Univers A, c'est-à-dire dans un univers déterministe, les participants tendaient vers l'incompatibilisme dans la situation émotionnellement faible et vers le compatibilisme dans la situation émotionnellement forte. Cela va dans le sens de l'hypothèse suggérée par Nichols et Knobe : les gens sont divisés entre des réponses émotionnelles compatibilistes et des réponses réflexives incompatibilistes.

Il serait donc trop simple de déclarer comme Nahmias que le sens commun est compatibiliste. À l'inverse, pour ceux qui considéreraient les réponses émotionnelles comme des biais, il semble plus correct de dire que le sens commun est fondamentalement incompatibiliste. Les résultats de Nichols et Knobe ont depuis été répliqués sur des participants issus de pays différents (Etats-Unis, Hong Kong, Inde, Colombie), sans variation notable[21].

- 4.2.3. Une théorie de l'erreur en faveur du compatibilisme

Mais est-il possible à l'hypothèse de Nichols et Knobe d'expliquer les résultats obtenus par Nahmias et ses collègues ? À première vue, il semble que oui : Nahmias et ses collègues ont obtenu des réponses compatibilistes parce que leurs scénarios étaient des cas de bonnes ou de mauvaises actions et étaient donc assez chargés d'un point de vue émotionnel pour susciter des réactions compatibilistes. Mais cette réponse se heurte à deux problèmes :

1. Le premier problème est que, dans l'un des scénarios utilisés par Nahmias et ses collègues, (***Superordinateur***), l'un des cas utilisés décrit une action neutre (aller faire du jogging) qui suscite pourtant des réponses compatibilistes. Ces résultats sont un problème pour la

[21] H. Sarkissian, A. Chatterjee, F. De Brigard, J. Knobe, S. Nichols et S. Sirker, « Is belief in free will a cultural universal? », in *Mind and Language*, 25/3, 2010, pp.346-358.

théorie de Nichols et Knobe, qui peine à les intégrer[22].

2. Le second problème provient des résultats que nous avons obtenus lors d'une expérience menée à l'hôpital de la Salpêtrière en collaboration avec Maxime Bertoux. Lors de cette expérience, nous avons soumis une description de l'Univers A suivi de la situation émotionnellement forte à 12 patients souffrant d'une variante comportementale de la démence frontotemporale, un trouble neurologique accompagné d'une forte diminution des réactions émotionnelles. Nous avons demandé à ces patients si l'agent dans la situation émotionnellement forte mais dans l'Univers A était responsable de son acte, devait être blâmé pour ce qu'il a fait et méritait une punition. En l'absence de réactions émotionnelles, et si l'hypothèse de Nichols et Knobe est correcte, nous aurions dû observer un pourcentage élevé de réponses incompatibilistes. Or, seul un patient sur 12 a donné des réponses incompatibilistes (et même ce patient oscillait entre compatibilisme et incompatibilisme selon les questions)[23].

Il faut donc trouver une autre façon d'expliquer l'écart apparent entre les résultats obtenus par Nahmias et ses collègues et ceux obtenus par Nichols et Knobe. La solution à ce problème, Nahmias l'a trouvée en étudiant un problème pourtant différent : celui de notre peur des neurosciences. Comme G&C et XB, beaucoup de gens (y compris un certain nombre de neuroscientifiques) ont l'impression que les neurosciences, en nous apprenant que nos décisions et nos actions sont le fruit de notre cerveau, prouvent que nous ne sommes pas libres et que la liberté n'est qu'une illusion. Mais pourquoi, s'ils sont compatibilistes ? Dans une expérience portant sur les différences entre déterminisme psychologique et déterminisme neurologique, Nahmias a utilisé la paire de scénarios suivante[24] :

> *Erta (Déterminisme Psychologique)* – Imaginez que dans un univers semblable à celui dans lequel nous habitons se trouve une planète nommée Erta qui ressemble beaucoup à la nôtre. La

[22] Voir : A. Feltz, E. Cokely et T. Nadelhoffer, « Natural compatibilsm versus natural incompatibilism: Back to the drawing board », in *Mind and Language*, 24/1, 2009, pp. 1-23.

[23] F. Cova, M. Bertoux, S. Bourgeois-Gironde et B. Dubois, « Judgments about moral responsibility and determinism in patients with behavioural variant of frontotemporal dementia : still compatibilists! », manuscrit, Institut Jean Nicod.

[24] E. Nahmias, « Folk fears about freedom and responsibility: Determinism vs. Reductionism », in *Journal of Cognition and Culture*, 6/1-2, 2006, pp. 215-237.

> géographie et la vie y sont très similaires à celles que l'on peut trouver sur Terre. On y trouve aussi une forme de vie avancée, les Ertains, qui parlent et se comportent comme nous, et nous ressemblent beaucoup. Néanmoins, la science des Ertains est beaucoup plus avancée que la nôtre. Plus précisément, les **psychologues** de la planète Erta ont découvert la façon précise dont fonctionne **l'esprit** des Ertains. Ces **psychologues** ont découvert que toute décision prise ou action accomplie par un Ertain est entièrement causée par **les pensées, désirs et projets qui se trouvent dans l'esprit de cet Ertain** et que ces **pensées, désirs et projets** sont entièrement causés par les événements qui ont précédé, y compris le patrimoine génétique et **l'éducation** de cet Ertain. Ainsi, à chaque fois qu'un Ertain agit, son action est entièrement causée par les **pensées, désirs et projets qu'il a à l'esprit** à ce moment donné, et ces **pensées, désirs et projets** sont entièrement causés par une chaîne d'événements antérieurs que l'on peut remonter jusqu'au patrimoine génétique et à **l'éducation** de cet Ertain.
>
> **Erta (Déterminisme Neurologique)** – [Même début que le scénario précédent] Néanmoins, la science des Ertains est beaucoup plus avancée que la nôtre. Plus précisément, les **neuroscientifiques** de la planète Erta ont découvert la façon précise dont fonctionne **le cerveau** des Ertains. Ces **neuroscientifiques** ont découvert que toute décision prise ou action accomplie par un Ertain est entièrement causée par les **réactions chimiques et les processus neurologiques qui se déroulent dans le cerveau de cet Ertain** et que **ces réactions chimiques et processus neurologiques** sont entièrement causés par les événements qui ont précédé, y compris le patrimoine génétique et l'**environnement physique** de cet Ertain. Ainsi, à chaque fois qu'un Ertain agit, son action est entièrement causée par les **réactions chimiques et les processus neurologiques qui se produisent dans son cerveau** à ce moment donné, et **ces réactions chimiques et processus neurologiques** sont entièrement causés par une chaîne d'événements antérieurs que l'on peut remonter jusqu'au patrimoine génétique et à **l'environnement physique** de cet Ertain.

Ces deux scénarios décrivent tous les deux des mondes déterministes. Mais, dans le premier monde, le déterminisme est un déterminisme *psychologique* (nos décisions et nos actions sont déterminées par nos *états mentaux*, qui sont eux-mêmes déterminés par notre patrimoine génétique et notre culture)

tandis que, dans le second scénario, il s'agit d'un déterminisme neurologique (nos décisions et nos actions sont déterminées par *l'activité de notre cerveau*).

Chaque participant recevait l'un des deux scénarios puis répondait aux deux questions suivantes :

- Lorsque les Ertains agissent, agissent-il librement ?
- Les Ertains méritent-ils d'être blâmés et loués pour leurs actions ?

Les réponses des participants peuvent être résumées de la façon suivante : alors que leurs réponses étaient hautement compatibilistes pour le cas *psychologique* (72% de réponses oui à la première question et 77% à la seconde question), elles étaient franchement incompatibilistes dans le cas *neurologique* (18% de réponses oui à la première question et 19% à la seconde question). Comment expliquer un tel renversement ?

Nahmias propose l'explication suivante : le compatibilisme dit que nous pouvons très bien être libres dans un monde déterministe du moment que nous agissons sur la base de désirs et de raisons qui découlent de nos valeurs – c'est-à-dire d'états mentaux – et cela quand bien même ces valeurs seraient elles-mêmes déterminées. Par contre, si des forces extérieures nous conduisent à agir sans que leur influence passe par nos valeurs, alors ces dernières sont mises hors circuit, et nous ne sommes pas libres (puisque nous n'agissons pas sur la base de nos valeurs). Ainsi, dans le cas *psychologique*, les Ertains agissent sur la base de leurs états mentaux, donc de leurs désirs, ce qui leur permet d'être libres (à supposer que certains de ces désirs émanent de leurs valeurs), même si ces désirs sont entièrement déterminés. En revanche, dans le cas *neurologique*, les Ertains agissent poussés par des mécanismes neurologiques, et non par leurs désirs et leurs valeurs, ce qui explique que, même si les participants sont compatibilistes, ils ne les considèrent pas comme libres.

Mais cette explication repose sur une prémisse qu'il faut mettre en lumière : elle suppose que les participants ont l'intuition que des processus neurologiques ne sauraient être des états mentaux, c'est-à-dire ne sauraient constituer des désirs et des valeurs (sans quoi, il serait parfaitement possible d'agir sur la base d'états mentaux en étant déterminés par son cerveau, puisque les états mentaux seraient des états intérieurs au cerveau). Autrement dit : les gens seraient *intuitivement dualistes* et auraient l'impression que des états cérébraux ne peuvent pas être des états mentaux, autrement dit que l'esprit est distinct du corps – et c'est ce dualisme qui expliquerait leurs réponses au cas *neurologique*, non le fait qu'ils aient des

intuitions incompatibilistes. Plus largement, la tendance générale à considérer que les explications neuroscientifiques minent notre liberté, présente même chez les neuroscientifiques, ne serait pas le fruit d'un incompatibilisme intuitif, mais d'un dualisme intuitif.

Mais comment cela peut-il nous aider à trancher entre les résultats de Nahmias et ses collègues et ceux de Nichols et Knobe ? En reprenant l'idée générale selon laquelle un compatibiliste peut donner des réponses incompatibilistes dès lors qu'il considère que le déterminisme, tel qu'il est présenté dans un cas, implique que nous n'agissons pas en fonction de nos désirs et de nos valeurs (ce qui est bien sûr une erreur : le déterminisme dit que nous pouvons agir en fonction de nos désirs et de nos valeurs, mais que ces désirs et ces valeurs sont déterminés par des facteurs extérieurs – la thèse selon laquelle nos états mentaux n'auraient aucune influence sur nos actions n'est pas le déterminisme mais porte le doux nom d'*épiphénoménisme*). L'hypothèse de Nahmias est qu'il y a quelque chose dans la façon dont Nichols et Knobe présentent leur scénario qui conduit les gens à mal comprendre le déterminisme en le confondant avec la thèse selon laquelle nos actions ne dépendent pas de nos désirs et de nos décisions.

Pour tester cette hypothèse, Nahmias et Murray[25] ont mis au point une expérience dans laquelle les participants recevaient soit le scénario de Nichols et Knobe (la description de l'Univers A et de l'Univers B), soit le cas de **l'Eternel Retour**. Dans chacun de ces deux groupes, la moitié des participants recevait des questions abstraites (ne portant sur aucun individu en particulier) tandis que l'autre moitié recevait des questions portant sur un individu particulier (Bill qui tue sa famille pour le scénario de Nichols et Knobe, Jill qui vole un collier pour **l'Eternel Retour**). Ces questions consistaient à demander aux participants si un agent vivant dans l'univers en question (ou l'agent particulier décrit dans le scénario) pouvait être moralement responsable de ses actes, pouvait agir librement et pouvait mériter d'être blâmé pour le mal qu'il faisait.

Plus important, après avoir répondu à ces questions, les participants devaient répondre à un certain nombre de questions permettant de déterminer s'ils avaient bien compris ce qu'était le déterminisme, ou s'il l'avait confondu avec autre chose. Par exemple :

1. Dans l'Univers décrit, les décisions d'une personne n'ont aucune influence sur ce qu'elle finit par faire. (VRAI ou FAUX ?)

[25] E. Nahmias et D. Murray, « Philosophy on free will: An error theory for incompatibilist intuitions », in *New Waves in Philosophy of Action*, New York, Palgrave Macmillan, 2010.

2. Dans l'Univers décrit, ce que veut une personne n'a aucune influence sur ce qu'elle finit par faire. (VRAI ou FAUX ?)
3. Dans l'Univers décrit, ce qu'une personne croit n'a aucune influence sur ce qu'elle finit par faire. (VRAI ou FAUX ?)

Une personne qui a vraiment compris le scénario doit répondre FAUX à ces trois questions. Celui qui répond VRAI confond le déterminisme (la thèse selon laquelle nous agissons selon des états mentaux et des décisions déterminés par des facteurs extérieurs) avec un cas particulier du déterminisme que nous appelons *épiphénoménisme* (la thèse selon laquelle nos états mentaux et nos décisions n'ont aucun impact sur nos actions).

Autre question :

4. Dans l'Univers décrit, tout ce qui se produit *devait* se produire, même si le passé avait été différent ? (VRAI ou FAUX ?)

Là encore, une personne qui a vraiment compris le scénario doit répondre FAUX : le déterminisme (au sens fort) est la thèse selon laquelle tout ce qui se produit devait se produire de cette façon, *étant donné l'état antérieur du monde*. Mais si l'on change cet état antérieur, alors ce qui s'ensuit change. La thèse selon laquelle les choses doivent se produire quoiqu'il arrive, même si le passé était largement différent, est le *fatalisme*. Il y a une différence énorme entre le fatalisme, pour lequel une chose qui a été « décrétée » arrivera de toute façon (par exemple : Œdipe aurait tué son père et épousé sa mère de toutes les façon, même s'il n'avait pas été abandonnée par ceux-ci) et le déterminisme, pour lequel une légère différence dans l'état initial du système peut entraîner d'énormes différences à long terme (par exemple : si Œdipe était arrivé un quart d'heure plus tard à ce croisement, il n'aurait pas rencontré son père et tout aurait été très différent).

Nahmias et Murray ont ainsi pu comparer les réponses des sujets aux questions portant sur la liberté des agents à celles sondant leur compréhension du déterminisme. Ils en ont tiré plusieurs conclusions intéressantes. Premièrement, ils ont observé une très forte corrélation inverse entre les réponses incompatibilistes aux trois premières questions et une bonne compréhension du déterminisme : plus les participants tendaient à confondre déterminisme « tout court » et déterminisme épiphénoméniste, plus ils tendaient à donner des réponses incompatibilistes. Deuxièmement, comme prévu par Nahmias et Murray, les gens avaient plus tendance à développer une mauvaise compréhension du déterminisme (et donc à donner des réponses incompatibilistes) dans le cas du scénario de Nichols et

Knobe que dans le cas de *l'Eternel Retour*. Troisièmement, les participants avaient plus tendance à développer une mauvaise compréhension du déterminisme dans les situations décrites de façon abstraite que dans les situations mettant en jeu des individus concrets, et donc plus tendance à donner des réponses incompatibilistes.

Ces résultats montrent ainsi que les « intuitions incompatibilistes » observées par Nichols et Knobe ne sont que des artefacts dus au fait que leur scénario et leur condition abstraite favorisent une confusion du déterminisme avec l'épiphénoménisme et le fatalisme. Plus généralement, ces résultats laissent penser que, si nombre de gens semblent à première vue incompatibiliste et penser que déterminisme et liberté sont incompatibles, c'est parce qu'ils ne comprennent pas bien le déterminisme. Au final, si l'on sonde les intuitions de ces mêmes personnes sur des cas qui leur permettent de bien saisir ce qu'est le déterminisme et quelles sont ses implications, on se rend compte que c'est le compatibilisme qui est intuitif, pas l'incompatibilisme : notre concept ordinaire de liberté (et de responsabilité morale) n'exige pas que nous agissions en-dehors de toute influence par des facteurs externes.

- 4.2.4. Implications pour l'argument descriptif

Ainsi, les expériences de Nahmias et ses collègues montrent que, comme G&C, les participants ont l'impression que les neurosciences menacent la liberté. Cependant, contrairement à ce que suppose leur argument, ce n'est pas parce que les gens adoptent une conception incompatibiliste de la liberté, mais parce qu'ils ont du mal à concevoir que nos états mentaux (désirs, intentions, croyances, etc.) puissent être des états matériels. C'est parce qu'ils sont dualistes, et non incompatibilistes, que les gens ont l'impression que les neurosciences constituent une menace pour la liberté.

Cela signifie que, si comme le souhaitent G&C, le discours des neurosciences se diffuse dans la société et que les gens le comprennent, alors les gens devraient de moins en moins penser que les neurosciences montrent qu'ils ne sont pas libres. En effet, si les gens viennent à embrasser pleinement le discours des neurosciences, ils devraient finir par abandonner leur dualisme intuitif pour accepter que les états mentaux puissent aussi être des états physiques dans le cerveau. Dans ce cas, ils ne confondront plus déterminisme et épiphénoménisme : ils comprendront que quand le cerveau agit, c'est *eux* qui agissent. Ils atteindront alors une vision du monde dans lequel ils sont des êtres qui agissent selon leurs intentions et leurs désirs, intentions et désirs complètement déterminés par leur passé (gènes, environnement, etc.). Or, les expériences de Nahmias et ses collègues

suggèrent que nous sommes en grande majorité prêts à dire que des agents qui agissent sur la base de leurs valeurs, quand bien même celles-ci seraient déterminées, sont parfaitement libres. Il semble donc qu'une meilleure compréhension et une meilleure diffusion des neurosciences conduiront en fait à une situation dans laquelle les gens auront arrêté de craindre que les neurosciences montrent qu'ils ne sont pas libres. Par exemple, ils seront prêts à envisager la possibilité que, dans les expériences de Libet, c'est leur cerveau, donc *eux*, qui déclenchent le mouvement avant d'avoir conscience d'en avoir l'intention. Dans ce cas, tout ce que montrent les expériences de Libet, c'est que nous n'avons pas immédiatement conscience de ce que *nous* faisons, ce qui est une observation intéressante (on peut d'ailleurs la faire dans d'autres domaines), mais ne menace en aucun cas notre liberté.

Bien sûr, il est possible qu'une mauvaise compréhension des neurosciences se répande dans la population, qui mélangerait déterminisme (nos actes sont déterminés pas notre cerveau) et dualisme (nous ne sommes pas notre cerveau, nos états mentaux ne sont pas matériels). Dans ce cas, la diffusion des neurosciences viendrait effectivement miner, comme le prédisent G&C, notre croyance dans notre caractère d'êtres libres et responsables d'eux-mêmes. Mais ce n'est probablement pas ce que G&C veulent dire : ils pensent qu'une *bonne* compréhension des neurosciences devrait nous conduire à abandonner notre concept traditionnel de responsabilité, pas une compréhension *erronée*.

4.3. Examen du passage de (C1) à (C2) : montrer que la responsabilité requise par la conception rétributive est incompatible avec le déterminisme ne suffit pas à rendre fausse cette conception

Un autre problème de l'argument de G&C est dans le passage de (C1) :

> (C1) Les neurosciences vont montrer que nous ne sommes ni libres, ni responsables de nos actes, au sens intuitif que nous donnons à ces termes.

À (C2) :

> (C2) Les neurosciences vont montrer que la conception rétributive de la peine, parce qu'elle repose sur une conception de la responsabilité qui est irréaliste, est inadéquate.

Ce passage semble assurer dans notre reconstruction de l'argument par la prémisse (2) :

> (2) La conception rétributive de la peine est liée à l'idée de responsabilité : selon elle, un agent ne saurait mériter de peine pour une action sans être responsable de cette action.

Autrement dit, si les neurosciences montrent que nous ne sommes pas libres, alors la conception rétributive de la peine s'effondre, parce qu'elle est *liée* à la notion de responsabilité. Mais, s'il est vrai que la conception rétributive de la peine est « liée » à l'idée de responsabilité, c'est dans le sens suivant :

> (a) *Conception rétributive de la peine* = Un agent ne mérite d'être puni que s'il est responsable de son acte.

Et *non* dans le sens suivant :

> (b) La *conception rétributive de la peine* n'est vraie que si nous pouvons être responsables de nos actes.

Dire que montrer que nous ne sommes pas libres rend fausse (ou inadéquate) la conception rétributive de la peine, c'est supposer (b) plutôt que (a). Mais c'est (a) qui est vraie, alors que (b) est fausse : la conception rétributive de la peine peut être vraie même si nous ne sommes jamais responsables de nos actes : elle a alors pour conséquence que nous ne méritons jamais d'être punis, ce qui est une conclusion radicale, mais ne la rend pas fausse. La thèse déterministe ne rend pas fausse la conception rétributive de la peine car elle est logiquement cohérente avec elle. Il n'y a pas de contradiction entre « seuls les gens responsables de leurs actes méritent d'être punis » et « personne n'est responsable ». Ainsi, même si les neurosciences montraient que le déterminisme est vrai, et même si la conception de la responsabilité à laquelle est liée la conception rétributive de la punition était incompatible avec le déterminisme, cela n'impliquerait pas automatiquement que nous rejetterions la conception rétributive de la responsabilité.

Néanmoins, G&C pourraient donc répondre qu'une conception de la peine qui nous conduirait à ne punir personne aurait des conséquences pratiques désastreuses, ce qui nous forcerait tout de même à l'abandonner, pour des raisons pratiques (et non théoriques). Acceptons donc cette

réponse: si nous ne sommes pas libres, alors la conception rétributive de la peine ne nous permet pas de punir les criminels, et est inutile. Parce que nous avons le souci de maintenir l'ordre public, il nous faudra bien abandonner la conception rétributive de la punition.

Une autre réponse que pourraient faire G&C (et qui est plus proche de l'esprit de leur argument d'origine) est que, une fois les questions de mérite écartées, notre seule préoccupation sera le bien-être de la population et sa maximisation. Ce souci devrait aussi nous conduire à abandonner la conception rétributive de la punition, qui, comme nous l'avons vu, ne vise pas à minimiser la souffrance.

À ces deux réponses, on pourra tout de même opposer l'existence *d'attitudes réactives* : même si nous en venons à croire réflexivement que nous ne sommes pas libres, il est probable que cela ne modifiera pas nos intuitions profondes selon lesquelles nous sommes responsables de nos actes, intuitions qui sont à l'origine de nombreuses réactions émotionnelles, comme la colère. Autrement dit, même si les neurosciences finissent par nous convaincre, à un niveau réflexif, que nous ne sommes pas libres, il n'en reste pas moins que nous continuerons à réagir comme si nous étions libres.

Mais, malgré les apparences, cette critique ne touche pas l'argument de G&C. G&C disent eux-mêmes qu'ils ne s'attendent pas à ce que nos attitudes réactives changent. Cependant, disent-ils, ce qui compte pour le droit, ce ne sont pas nos réactions, mais les discussions réflexives entre experts, dans lesquelles on peut espérer que les attitudes réactives seront mises de côté. Pour citer XB :

> « Il nous est certes impossible, dans la vie de tous les jours, d'éviter d'agir comme si autrui était doté d'un libre-arbitre, et il serait sûrement desséchant humainement de penser les actions de nos proches comme totalement déterminées. Mais la réflexion proprement politique ne subit pas un tel écueil, comme le montre l'analogie proposée par Greene et Cohen : bien que la physique nous enseigne que l'espace est courbe, il nous paraît parfaitement impossible de nous représenter un espace non plat. Cette incapacité n'est nullement un problème pour la vie de tous les jours ; ce n'est que lorsque nous souhaitons lancer des fusées que nous devons raisonner en termes d'univers courbe. De même, notre tendance innée à considérer que les actions d'autrui ne sont pas déterminées n'est pas un problème pour notre vie quotidienne ; en revanche, lorsqu'il s'agit de décider d'envoyer quelqu'un en prison, nous pouvons difficilement ne pas tirer toutes les implications de la thèse causale ».

Autrement dit, G&C font juste l'hypothèse que les neurosciences nous conduiront à adhérer réflexivement à l'idée selon laquelle nous ne sommes pas libres, et que cela suffira pour changer le droit. L'objection des attitudes réactives ne touche donc pas leur argument. Il n'en reste pas moins qu'il existe des raisons de douter de l'hypothèse selon laquelle la mise en valeur de l'incompatibilité de la conception rétributive de la responsabilité avec le déterminisme du monde nous conduira à rejeter la conception rétributive de la peine : d'autres issues sont possibles.

4.4. Examen de la prémisse (5) : Accepter le déterminisme ne nous conduira pas nécessairement à adhérer à la conception utilitariste de la peine

Finalement, même si accepter la vérité de la thèse déterministe nous conduisait à rejeter la conception rétributive de la peine, rien ne nous dit que nous serions pour autant enclin à adhérer à la conception utilitariste de la peine. En fait, la prémisse (5) est une prémisse empirique susceptible d'être évaluée sur la base de données expérimentales, et les résultats dont nous disposons sont loin de donner une image claire et unifiée des choix pénaux de ceux qui croient dans la vérité du déterminisme.

Les premiers résultats chronologiquement disponibles semblent confirmer la prémisse (5) de G&C. Nettler[26] a soumis à un millier de participants deux questionnaires. Le premier permettait de mesurer dans quelle mesure les participants considéraient que ce qui leur arrive dans la vie détermine le comportement des personnes. Le second mesurait leur tendance à punir les auteurs de délits plutôt qu'à les réhabiliter. Ce qu'a trouvé Nettler, c'est que les participants qui avaient tendance à ne pas croire que nos comportements étaient déterminés avaient effectivement plus tendance à attribuer des punitions aux coupables, et des punitions dont le but principal était de « faire payer » le coupable plutôt que de le réhabiliter.

Néanmoins, ces résultats ne doivent pas être adoptés sans réserve car ils sont loin d'avoir été reproduits. Viney et ses collègues[27] ont donné à un ensemble de participants un questionnaire mesurant leur degré de croyance au libre-arbitre et un questionnaire mesurant leur tendance à attribuer des punitions. Ils ont observé une corrélation inverse entre les deux scores : autrement dit, selon leurs résultats, les personnes niant l'existence de la

[26] G. Nettler, « Cruelty, dignity, and determinism », in *American Sociological Review*, 24/3, 1959, pp. 375-384.
[27] W. Viney, D. Waldman et J. Barchilon, « Attitudes towards punishment in relation to beliefs in free will and determinism », in *Human Relations*, 35/11, 1982, p. 939.

liberté et acceptant le déterminisme attribuaient des punitions plus lourdes que les défenseurs de la liberté. Qui plus est, certaines questions du second questionnaire permettaient de mesurer si les participants considéraient (explicitement) la punition plutôt comme une forme de rétribution ou plutôt comme une forme de réhabilitation. Sur ce point, ils n'ont observé aucune différence entre les deux types de participants.

Le problème de ces deux expériences, c'est qu'elle ne différencie pas entre deux raisons de « faire payer » les coupables : parce qu'il le mérite ou pour protéger la société ? Réhabiliter le coupable ne constitue pas, comme nous l'avons vu, toute la conception utilitariste de la peine. Heureusement pour nous, Viney et ses collègues[28] ont donné à un autre ensemble de participants 4 scénarios décrivant 4 types différents de comportements mauvais. Pour chaque scénario, les participants devaient attribuer une punition à l'agent (en années de prison) puis justifier cette attribution en choisissant parmi les 4 justifications suivantes :

- Pour réformer ou réhabiliter le condamné.
- Pour protéger la société.
- Le condamné mérite d'être puni.
- Pour le faire souffrir ou pour qu'il paye pour son crime.

Les deux premières réponses correspondent à une vision utilitariste de la peine et les deux dernières à une conception rétributive. Les participants remplissaient aussi un questionnaire évaluant leur croyance à la liberté et au déterminisme. Au final, Viney et ses collègues n'ont trouvé aucune différence entre les personnes qui considéraient que le monde est déterminé et que nous ne sommes pas libres et ceux qui considéraient que nous étions libres et responsables de nos actes. Les deux groupes attribuaient des punitions de même ampleur, et, selon les scénarios, utilisaient les deux types de justification.

Ainsi, il n'est pas évident que l'abandon général de la croyance à la liberté entraînerait une réforme du système pénal dans un sens favorable à l'utilitarisme en favorisant la punition comme réhabilitation sur la punition comme rétribution. Comme nous l'avions vu en critiquant la prémisse (2), il reste possible d'adhérer à une vision rétributive de la punition tout en acceptant le fait que nous ne sommes pas libres. D'autres études sont nécessaires avant de pouvoir se prononcer sur ce sujet.

[28] W. Viney, P. Parker-Martin et S. Dotten, « Beliefs in free will and determinism and lack of relation to punishment rationale and magnitude », in *Journal of General Psychology*, 115/1, 1988, pp.15-23.

En conclusion, il y a de nombreuses raisons de douter de l'argument descriptif de G&C : premièrement, il n'est pas exclu que la conception rétributive de la peine puisse avoir un sens sans référence à la notion de responsabilité ; deuxièmement, il n'est pas clair que notre concept ordinaire de responsabilité soit incompatible avec le déterminisme ; troisièmement, même si le déterminisme est incompatible avec la responsabilité morale, et même si la conception rétributive de la peine est liée à la notion de responsabilité, cela ne montre pas qu'elle est fausse, et donc ne rend pas inévitable son abandon ; finalement, même si le déterminisme nous conduit à abandonner la conception rétributive de la peine, rien ne nous garantit que ce soit pour adopter la conception utilitariste de la peine.

V. Examen critique de l'argument normatif : les enseignements des neurosciences ont-il pour conséquence que nous devons abandonner la conception rétributive de la peine ?

Passons maintenant à l'examen de l'argument normatif. L'argument normatif a en effet le mérite d'échapper à certaines difficultés rencontrées par l'argument descriptif. Prenons l'exemple de la prémisse (2), selon laquelle la conception rétributive de la peine est liée à l'idée de responsabilité. Nous avons objecté à cette prémisse qu'il pourrait bien y avoir un sens dans lequel cette conception pourrait rester compréhensible sans la notion de mérite : par exemple, dans une conception « cosmique » de la peine, selon laquelle la punition a pour but de restaurer un équilibre. Néanmoins, si cette objection peut toucher l'argument descriptif (parce qu'elle suggère qu'il y a possibilité psychologique que les gens restent attachés à la conception rétributive de la peine sur d'autres bases que la croyance au mérite), le tenant de l'argument normatif peut répondre qu'une telle conception serait normativement très discutable, parce qu'elle reposerait sur des entités et une métaphysique qu'il serait difficile de justifier (que faut-il entendre au juste par « harmonie » ou « équilibre » ?). Si, au mieux, on allège métaphysiquement cette thèse pour entendre par harmonie la restauration de la confiance inter-individuelle au sein de la société, par la garantie que ses membres devront « payer » en cas de crime, alors cette compensation peut s'intégrer parfaitement à une approche purement utilitariste – il s'agit bien de maximiser le bonheur des membres de cette société[29].

[29] Nous remercions un relecteur anonyme de *Klesis* pour ces remarques que nous reprenons ici.

5.1. Examen de la prémisse (3) : notre concept de mérite ne requiert pas une conception incompatibiliste de la liberté*

De la même façon, étant donné que la prémisse (3*) de l'argument normatif est légèrement différente de celle de l'argument descriptif, en ce qu'elle ne repose pas sur ce que pensent les gens, elle ne devrait pas être sensible aux résultats suggérant que les gens sont naturellement compatibilistes. Pour rappel :

> (3*) La liberté et la responsabilité morale sont incompatibles avec le déterminisme.

Ces résultats ont pourtant un léger impact : dans son argumentaire, XB semble considérer les théories compatibilistes, qui soutiennent que (3*) est faux, comme allant directement à l'encontre de l'évidence. Pour le citer :

> « Les théories compatibilistes ne sont pas convaincantes parce qu'elles contournent l'implication fondamentale de la thèse causale, à savoir qu'il apparaît illusoire d'attribuer la responsabilité morale à un individu qui ne pouvait pas s'empêcher de commettre son acte. »

Autrement dit, ce que XB reproche aux théories compatibilistes, c'est de nier l'évidence : que le déterminisme rend les agents non responsables de leurs actes. Mais est-ce vraiment une évidence ? Comme on vient de le voir, il semble que, malgré les apparences, c'est l'inverse qui est intuitif pour beaucoup de gens : une fois écartés les malentendus dus à leur dualisme intuitif ou à leur confusion du déterminisme avec l'épiphénoménisme, nombreuses sont les personnes à considérer que la liberté est compatible avec le déterminisme, et donc que la prémisse selon laquelle le déterminisme serait incompatible avec la responsabilité n'a rien d'évident.

À cela on peut ajouter que si l'on consulte le *PhilPapers Survey*, un questionnaire qui a circulé sur Internet à l'échelle mondiale et qui interrogeait les philosophes professionnels sur leurs positions[30], on voit que 59% des répondants ont déclaré adhérer au (ou incliner vers le) compatibilisme, alors que seulement 28,6% d'entre eux ont déclaré adhérer à (ou incliner vers) une position incompatibiliste (libertarisme ou déterminisme « dur »). Autrement dit, le compatibilisme semblait une position bien plus acceptée que l'incompatibilisme. De plus, le

[30] Voir : http://philpapers.org/surveys/.

compatibilisme est loin d'être une position neuve. À l'époque moderne, il était représenté par des auteurs comme Locke, Leibniz, Hume ou Hobbes. Il était même déjà présent au tout début des débats sur les rapports entre nécessité ou liberté, lorsque les épicuriens soutenaient que nécessité et liberté étaient incompatibles (ils étaient incompatibilistes) et que les stoïciens (en bon compatibilistes) répondaient que nous pouvions être libres même dans un monde soumis à un plan divin programmé à l'avance[31].

Tout cela montre que la prémisse (3*) n'est pas aussi évidente que le pense XB et que la charge de la preuve n'incombe pas aux compatibilistes : les incompatibilistes aussi doivent avancer des arguments. Bien évidemment, XB propose un certain nombre d'arguments contre la thèse compatibiliste. Ce sont ces arguments que nous allons maintenant discuter.

- 5.1.1. La théorie de l'erreur de XB

Une première série d'arguments utilisés par XB consiste à tenter de miner le compatibilisme en montrant que l'adhésion à cette thèse provient de sources psychologiques peu fiables. Ainsi, la plupart des tenants du compatibilisme seraient coupables de confusion en ne faisant que repousser le problème. Pour citer XB :

> « Si le caractère intentionnel ou réfléchi de l'action paraît intuitivement fonder la responsabilité, c'est *uniquement* parce que nous tendons spontanément à penser les intentions, les désirs de second ordre et les délibérations rationnelles comme des sources de choix non causés, *comme des causes premières* ».

Dans cette phrase, XB vise en même temps plusieurs conceptions compatibilistes. L'une d'elles, que nous prendrons à titre d'exemple, est celle de Frankfurt, selon laquelle sont libres les actions qui découlent de désirs de premier ordre (c'est-à-dire de désirs portant sur des objets extérieurs) eux-mêmes approuvés par des désirs de second ordre (c'est-à-dire de désirs portant sur ces désirs). Ainsi, pour reprendre un célèbre

[31] Pour les stoïciens, tout ce qui existe dans le monde (y compris les âmes humaines) est de nature spatio-temporelle : tout est corps (à quelques exceptions près que nous laisserons de côté). Le monde en son entier n'est qu'un gigantesque corps : Dieu. Or, tout ce qui se produit dans le monde est le fruit de la nécessité divine : tout arrive de la façon dont cela devait arriver. De plus, le monde (Dieu) a une vie cyclique. À chaque fin de cycle, le monde s'embrase puis renaît et vit un cycle totalement identique, dans lequel chaque individu revit la même vie. Pour résumer, selon les stoïciens : (i) tout arrive selon le plan divin et (ii) tout s'est déjà produit de la même façon et se reproduira à l'identique. On est clairement dans une perspective déterministe. Pourtant, les stoïciens sont considérés comme les chantres de la liberté : selon eux, même enchaîné tout au fond d'une prison, vous êtes toujours libres.

exemple donné par Frankfurt lui-même[32], un drogué qui succombe à son désir de prendre de la drogue (désir de premier ordre) alors qu'il avait le désir de ne pas succomber à ce désir (désir de second ordre) n'est pas libre. En revanche, le drogué qui prend de la drogue et est parfaitement en accord avec cela (son désir de second ordre suit son désir de second ordre) est libre.

XB rejette ce genre de théories parce qu'il pense que l'introduction des désirs de second ordre constitue juste une façon de repousser le problème en concédant que les désirs de premier ordre sont causés mais en réintroduisant un réquisit indéterministe au niveau des désirs de second ordre. Pour le citer encore une fois :

> « Mettre l'accent sur les capacités mentales de l'agent semble ainsi correspondre à nos intuitions quotidiennes : lorsque nous nous trouvons en situation de choix, et que nous délibérons rationnellement et consciemment, nous avons la conviction que notre futur n'est pas joué d'avance, et qu'il ne dépend pas de notre "personnalité" initiale. »

Le problème avec cette critique, c'est qu'elle rate complètement l'intuition initiale de nombreuses conceptions compatibilistes, y compris celle de Frankfurt[33]. Depuis au moins Hume, l'une des intuitions compatibilistes majeures est justement que l'action libre est celle qui provient de notre personnalité initiale. Le critère de Frankfurt, celui des désirs de second ordre, ne vise pas à réintroduire le libre choix et l'indéterminisme au niveau des désirs de second ordre, mais à formuler un critère de ce que Frankfurt appelle dans d'autres articles *l'identification*[34] : un agent est libre quand il s'identifie aux désirs qui le motivent, c'est-à-dire quand il s'y reconnaît. Le fait que ses désirs de second ordre s'accordent avec ses motifs est juste un signe fiable d'une telle identification. Autrement dit, l'intuition fondamentale qui guide la conception de Frankfurt est l'idée qu'un agent est

[32] H. Frankfurt, « Freedom of the will and the concept of a person », in *The Journal of Philosophy*, 68/1, 1971, pp. 5-20.

[33] Un autre problème de cette critique est qu'elle semble confondre deux choses. XB dit que nous avons « l'intuition quotidienne » que « notre futur n'est pas joué d'avance ». Imaginons que ce soit le cas. Et alors ? Le fait que nous ayons l'intuition que notre futur n'est pas joué à l'avance (et que le déterminisme est faux) ne signifie pas que nous avons l'intuition que la responsabilité morale est incompatible avec le déterminisme. Ainsi, dans les expériences de Nichols et Knobe, les participants qui recevaient la *situation concrète* répondaient à la fois que notre monde ressemblait plus au monde indéterministe *et* qu'un agent dans le monde déterministe pouvait être responsable de ses actes.

[34] H. Frankfurt, « Identification and wholeheartedness », in *Responsibility, Character and Emotions: New Essays in Moral Psychology*, Cambridge, Cambridge University Press, pp. 159-176.

libre quand il agit sur la base de valeurs qui sont vraiment celles auxquelles il adhère, et pas celle d'un désir auxquels ces valeurs s'opposent.

L'intuition fondamentale à la base de la théorie de Frankfurt a son origine dans la vie quotidienne : alors que certains de nos désirs nous paraissent naturellement découler de ce que nous aimons, nous avons parfois des désirs qui s'opposent à nos valeurs. Je peux avoir l'envie de voler de l'argent à un ennemi, tout en jugeant (selon mes valeurs) que ce serait mal et peu honnête. Je peux penser que la fidélité est une des plus hautes vertus et pourtant désirer une autre personne que mon conjoint. Cette structure duale se retrouve déjà chez Aristote, qui nomme intempérant celui qui a des désirs qui s'opposent à ce qu'il veut « raisonnablement ».

Autrement dit, selon ces théories compatibilistes, ce qui rend libre l'agent, c'est le fait d'agir selon ses propres valeurs, et pas sur la base de désirs passagers qui seraient incompatibles avec ces valeurs[35]. Et bien sûr, le fait que ces valeurs soient complètement déterminées par nos gènes et notre environnement ne change rien. Dans le *cas des Jumeaux Séparés*, que nous avons rencontré plus haut, Fred agit librement parce qu'il agit en accord avec ses valeurs (l'argent vaut plus que l'honnêteté) – quand bien même ces valeurs seraient le fruit de son éducation.

Ainsi, la genèse psychologique du compatibilisme proposée par XB, selon laquelle la moitié des philosophes s'illusionneraient en repoussant le problème, repose sur une mauvaise compréhension de l'intuition fondamentale de la plupart des théories compatibilistes : l'intuition selon laquelle être libre, c'est agir selon ses propres valeurs, qu'importe leur origine[36]. Pour citer Bergson (qui était par ailleurs incompatibiliste, mais pour des raisons bien particulières)[37] :

---

[35] Il convient de noter que, dans toute cette discussion, nous entendons par liberté le fondement de notre capacité à être responsable moralement de nos actes. Certains philosophes dissocient liberté et responsabilité morale. Ainsi, Spinoza fonde sa propre notion de liberté en la détachant complètement de la question de la responsabilité morale, qu'il estime être impossible. À l'inverse, les philosophes qui se disent « semi-compatibilistes » considèrent que la liberté est incompatible avec le déterminisme mais que nous pouvons être moralement responsables dans un monde déterministe. Mais, dans le cas présent, nous entendons par « compatibilistes » ces philosophes qui ne cherchent pas à créer leur propre notion de responsabilité mais qui défendent la thèse selon laquelle nos notions ordinaires de liberté et de responsabilité morale sont liées et compatibles avec le déterminisme. Ainsi, quand Frankfurt dit que nous pouvons être libre dans un univers déterministe, il parle de la liberté nécessaire pour la responsabilité morale et affirme donc du même coup que nous pouvons être moralement responsables dans un univers déterministe. (Merci encore une fois à un relecteur anonyme pour nous avoir poussé à faire ces précisions.)

[36] Notons que, dans sa tentative pour faire la genèse psychologique du compatibilisme, XB s'attarde sur les expériences de Heider et Simmel montrant que nous avons spontanément tendance à attribuer des états mentaux (et en particulier des buts et des intentions) à des formes géométriques (voir : F. Heider et M. Simmel, « An experimental study of apparent

> « Bref, nous sommes libres quand nos actes émanent de notre personnalité entière, quand ils l'expriment, quand ils ont avec elle cette indéfinissable ressemblance qu'on trouve parfois entre l'œuvre et l'artiste. En vain on alléguera que nous cédons alors à l'influence toute-puissante de notre caractère. Notre caractère, c'est encore nous ; et parce qu'on s'est plu à scinder la personne en deux parties pour considérer tour à tour, par un effort d'abstraction, le moi qui sent ou pense et le moi qui agit, il y aurait quelque puérilité à conclure que l'un des deux moi pèse sur l'autre. Le même reproche s'adressera à ceux qui demandent si nous sommes libres de modifier notre caractère. Certes, notre caractère se modifie insensiblement tous les jours, et notre liberté en souffrirait, si ces acquisitions nouvelles venaient à se greffer sur notre moi et non pas à se fondre en lui. Mais, dès que cette fusion aura lieu, on devra dire que le changement survenu dans notre caractère est bien nôtre, que nous nous le sommes approprié. En un mot, si l'on convient d'appeler libre tout acte qui émane du moi, et du moi seulement, l'acte qui porte la marque de notre personne est véritablement libre, car notre moi seul en revendiquera la paternité ».

- 5.1.2. L'argument de la manipulation

Il est d'autant plus curieux d'accuser Frankfurt de considérer que la liberté provient de désirs de second ordre qui seraient des causes premières que celui-ci est célèbre pour avoir « avalé le morceau » (*bitten the bullet*) dans une discussion avec Locke[38]. Alors que Locke lui demandait si, selon lui, un agent dont les désirs de second ordre seraient implantés en lui par un démon serait toujours libre en agissant selon ces désirs, Frankfurt a répondu positivement. Cette réponse est contre-intuitive, mais d'autres théories

---

behavior », in *American Journal of Psychology*, 57, 1944, pp. 243-259). Selon XB, cette tendance serait à la source des intuitions compatibilistes. Néanmoins, cette hypothèse repose sur une double confusion. Premièrement, la tendance à attribuer des états mentaux n'est pas nécessairement tendance à attribuer de la responsabilité morale : nous attribuons des états mentaux aux animaux et aux enfants sans pour autant leur attribuer de la responsabilité morale. Deuxièmement, une tendance à considérer les gens comme responsable n'est pas la même chose qu'une tendance à penser que cette responsabilité est compatible avec le déterminisme. On comprend donc mal comment la tendance à attribuer des états mentaux à tout ce qui présente des signes d' « agentivité »est censée expliquer l'adhésion à la thèse selon laquelle la responsabilité morale est compatible avec le déterminisme.

[37] H. Bergson, *Essai sur les Données Immédiates de la Conscience*, Paris, Presses Universitaires de France, 1997.

[38] Pas le célèbre philosophe, mais un autre du même nom. Voir : D. Locke et H. Frankfurt, « Three concepts of free action », in *Proceedings of the Aristotelian Society Supplementary Volume*, 49, 1992, pp. 95-125.

compatibilistes ne sont pas forcées de l'admettre : après tout, des désirs de second ordre implantés ne sont pas représentatifs des véritables valeurs de l'agent.

Cet exemple nous amène au dernier argument de XB, un argument positif cette fois, qu'il tire du texte de G&C. Cet exemple met en scène un homme nommé Monsieur Marionnette (Mr. Puppet) :

> « Imaginons qu'un groupe de scientifiques aient sélectionné avec attention les gènes et l'environnement d'un individu ('Mr. Puppet') de manière à ce qu'il commette un crime dans certaines circonstances. Selon les auteurs, on ne saurait intuitivement tenir Mr. Puppet pour responsable de son crime, dans la mesure où celui-ci était causé par des forces qui n'étaient pas sous son contrôle. »[39]

L'expérience de pensée de Mr. Puppets semble constituer un contre-exemple aux théories compatibilistes que nous venons d'esquisser : lorsque Mr. Puppet agit (commet le crime en question), il agit bien selon les valeurs qui sont les siennes. Le compatibiliste devrait alors déduire qu'il est responsable de ses actes. Or, nous dit XB, nous avons l'intuition *qu'il n'est pas* responsable de ses actes. Donc le compatibilisme a tort.

Une première façon de rejeter l'argument consiste à dire que l'intuition selon laquelle Mr. Puppet n'est pas libre est loin d'être partagée par tous. Le cas n'a pas été testé directement, mais on peut citer une expérience utilisant un cas assez proche. Dans cette expérience[40], les participants recevaient le cas suivant :

> *JoJo* – Jojo est le fils préféré de Jo le Premier, un dictateur méchant et sadique qui règne sur un petit pays sous-développé entièrement coupé du monde extérieur. Parce que son père a des sentiments particuliers à son égard, Jojo suit une éducation spéciale et on lui permet d'accompagner et d'observer son père dans ses tâches quotidiennes. Ce traitement a pour effet que Jojo prend comme son père comme modèle et développe des valeurs très

[39] L'argument de G&C et de XB consiste à dire (i) que Mr. Puppet n'est pas libre dans ce cas et (ii) qu'il n'y a aucune différence significative entre ce cas particulier et celui d'une personne normale soumis au déterminisme. Ce procédé, qui consiste à rapprocher les cas de déterminisme des cas de manipulation, est une version simplifiée du fameux Argument de la Manipulation de Pereboom. Il existe plusieurs façons de résister à la comparaison en disant qu'il existe des différences significatives entre les cas de déterminisme et de manipulation, mais nous ne les développerons pas ici. Voir : D. Pereboom, « Determinism *al dente* », in *Nous*, 29/1, 1995, pp. 21-45.

[40] D. Faraci et D. Shoemaker, « Insanity, deep selves, and moral responsibility: the case of JoJo », in *Review of Philosophy and Psychology*, 1/3, 2010, pp. 1-14.

> proches de celle de son père. Une fois adulte, Jojo se comporte de manière similaire à son père : il lui arrive couramment d'envoyer des gens en prison ou de les faire tuer ou torturer sur un coup de tête. Il n'est pas obligé de faire ces choses. Quand il lui arrive de prendre du recul et de se demander : "Est-ce que je veux vraiment être ce genre de personne?", sa réponse est "Oui", car cette vie exprime ses valeurs et ses idéaux les plus profonds.

Dans ce cas, il était demandé aux participants de dire à quel point, selon eux, Jojo méritait d'être blâmé pour ses actions (sur une échelle allant de 1 à 7 avec « 1 = pas du tout », « 4 = un petit peu » et « 7 = complètement »). La moyenne des réponses fut de 4, 77. Or, le milieu de l'échelle étant à 4, une moyenne de 4,77 signifie que la majorité des participants considéraient Jojo comme responsable de ses actes. Il n'est donc pas si évident que Mr. Puppet n'est pas libre de ses actes.

Mais disons que ce cas suscite l'intuition selon laquelle Mr. Puppet n'est pas responsable de ses actes. Dans ce cas, il existe au moins trois raisons de ne pas faire confiance à cette intuition.

1. La première raison est que, dans le cas de Mr. Puppet, il existe d'autres personnes responsables du crime de Mr. Puppet : les hommes qui l'ont manipulé. Il se pourrait que leur rôle soit si saillant que nous ayons tendance à leur attribuer toute la responsabilité, diminuant ainsi celle de Mr. Puppet. Autrement dit, nos intuitions seraient victimes d'une illusion due au phénomène de diffusion de responsabilité : si nous avons déjà des responsables (les manipulateurs), pas besoin d'en chercher un autre (Mr. Puppet).

2. Cet effet pourrait même être augmenté par la présence d'un autre facteur : le fait que nous plaignons Mr. Puppet pour son sort (avoir été élevé par des manipulateurs et transformé en criminel). Selon une explication proposée par Shaun Nichols[41], cela nous conduirait à considérer Mr. Puppet comme une victime, ce qui « bloquerait » les mécanismes psychologiques impliqués dans l'attribution de responsabilité, et nous conduirait à le considérer comme non responsable. Autrement dit, nous serions victimes d'une sorte de biais : le statut de victime de Mr. Puppet nous empêcherait de nous concentrer sur son statut de responsable.

3. Finalement, une troisième raison, compatible avec les deux précédentes, consiste à dire que nous ne considérons pas que Mr. Puppet est libre parce que nous pensons en fait que l'éducation qu'il a reçue a étouffé ses véritables valeurs, et donc qu'en commettant son crime, il n'agit pas

[41] S. Nichols, « After incompatibilism: A naturalistic defense of the reactive attitudes », in *Philosophical Perspectives*, 21/1, 2007, pp.405-428.

selon ses valeurs. En effet, il semble que nombre de personnes considèrent que, au fond, même un criminel veut le bien. Dans une série d'expériences, Joshua Knobe et Erica Roedder[42] se sont intéressés au concept ordinaire de « accorder de la valeur à » (en anglais : *valuing*). Selon eux, ce concept contient une composante normative : un agent A ne peut être dit « accorder de la valeur à » un objet O que si O est moralement bon. Entrons dans le détail de l'une des expériences sur lesquelles Knobe et Roedder s'appuient pour défendre cette thèse. Les participants recevaient l'un des deux textes suivants :

> *Georges, Antiraciste Malgré Lui* – Georges appartient à une culture dans laquelle la plupart des gens sont extrêmement racistes. Georges pense que le point de vue des membres de sa culture est plutôt correct. Autrement dit, il croit qu'il a le devoir de favoriser les intérêts des membres de sa race au détriment de ceux des membres d'autres races. Néanmoins, Georges se sent souvent attiré par le point de vue opposé. Lorsqu'il fait du mal aux membres d'autres races, il ressent très souvent un sentiment de culpabilité. Et il lui arrive parfois même d'agir selon ces sentiments et de favoriser l'égalité des races. Georges voudrait bien pouvoir changer cet aspect de sa personnalité. Il aimerait ne plus se sentir attiré par l'idée de l'égalité des races et pouvoir agir uniquement dans l'intérêt de sa propre race.
>
> *Georges, Raciste Malgré Lui* – Georges appartient à une culture dans laquelle la plupart des gens croient à l'égalité des races. Georges pense que le point de vue des membres de sa culture est plutôt correct. Autrement dit, il croit qu'il a le devoir de favoriser les intérêts de tous, quelle que soit leur race. Néanmoins, Georges se sent souvent attiré par le point de vue opposé. Lorsqu'il aide des gens d'autres races au détriment de membres de la sienne, il ressent très souvent un sentiment de culpabilité. Et il lui arrive parfois même d'agir selon ces sentiments et de favoriser la discrimination raciale. Georges voudrait bien pouvoir changer cet aspect de sa personnalité. Il aimerait ne plus se sentir attiré par l'idée de discrimination raciale et pouvoir agir uniquement dans l'intérêt de tous, sans considération de race.

Les participants ayant reçu le premier scénario devaient dire à quel point ils étaient d'accord avec la phrase : « en dépit de ce qu'il croit consciemment,

[42] J. Knobe et E. Preston-Roedder, « The ordinary concept of valuing », in *Philosophical Issues*, 19/1, 2009, pp. 131-147.

Georges attribue en fait de la valeur à l'égalité entre races ». Les participants ayant reçu le second scénario devaient dire eux à quel point ils étaient d'accord avec la phrase : « en dépit de ce qu'il croit consciemment, Georges attribue en fait de la valeur à la discrimination raciale ». Les deux groupes répondaient en utilisant une échelle de -3 (désaccord) à 3 (accord). Pour le premier scénario, la moyenne fut de 0,83, et de −1,14 pour le second. Knobe et Roedder expliquent cette différence en supposant que l'idée que O est moralement bon fait partie du concept populaire « d'attribuer de la valeur à O ».

Il existe néanmoins une autre façon d'expliquer ces résultats. Chad Gonnerman[43] a repris les deux scénarios de Knobe et Roedder mais a cette fois demandé aux participants à quel point ils étaient d'accord avec les phrases « tout au fond de lui, ce que Georges veut réellement, c'est l'égalité entre races / la discrimination raciale ». Dans ce cas, on peut aussi observer une différence : la moyenne des réponses est de 0, 32 pour le premier scénario (« égalité entre races ») et de −1,05 pour le second (« discrimination raciale »). Un autre groupe de participants devait quant à lui évaluer son accord avec les phrases « malgré ce qu'il dit la plupart du temps, ce que, tout au fond de lui, Georges pense être la meilleure chose à faire, c'est promouvoir l'égalité entre races / la discrimination raciale ».Encore une fois, une différence a pu être observée : 1, 192 (« égalité entre races ») contre −0, 286 (« discrimination raciale »).

L'explication de Knobe et Roedder semble donc insuffisante, parce que limitée à « accorder de la valeur à ». Tentons donc une autre hypothèse. Dans les deux cas, il est probable que les participants tendent à se représenter la situation de Georges comme un conflit entre les aspirations de son « soi profond » (ses véritables valeurs) et celles de son « soi superficiel » (ses désirs qui ne correspondent pas à ses véritables valeurs). Leurs réponses aux questions dépendront des valeurs qu'ils imputent au « soi profond » et de celles qu'ils rejettent dans le « soi superficiel ». Par exemple, dans le cas du premier scénario, on peut tout autant se représenter un Georges « profondément » antiraciste mais « superficiellement » raciste du fait de sa culture qu'un Georges « profondément » raciste suite à son éducation en lutte avec des aspirations antiracistes qui ne sont que les effets « superficiels » de tendances biologiques à l'empathie. Ce que révèlent néanmoins les résultats de Knobe, Roedder et Gonnerman, c'est qu'un nombre non négligeable de participants ont tendance à attribuer plus facilement au « soi profond » des tendances moralement bonnes et à rejeter

[43] C. Gonnerman, « Reading conflicted minds: An empirical follow-up to Knobe and Roedder », in *Philosophical Psychology*, 21/2, 2008, pp. 193-205.

dans le « soi superficiel » des tendances qu'ils jugent moralement mauvaises. Autrement dit : un certain nombre de personnes considèrent que, au fond, la plupart des gens sont bons, et que la méchanceté n'est souvent que le produit d'influences qui viennent étouffer la bonté qui est en nous.

Cette hypothèse peut être maintenant transposée au cas de Mr. Puppet : il se peut que nombre d'entre nous attribuent, malgré ce que dit le scénario, des valeurs moralement bonnes à Mr. Puppet – ce qui a pour conséquence que l'éducation spéciale qu'il reçoit ne constitue pas ses véritables valeurs, mais vient en fait les contredire, de telle sorte qu'en commettant le crime, Mr. Puppet n'agit pas selon ses valeurs. Autrement dit, notre intuition selon laquelle Mr. Puppet n'est pas libre serait due à une sorte d' « optimisme anthropologique », et pas du tout au fait que Mr. Puppet est déterminé. Cette hypothèse fait d'ailleurs une prédiction : que nous aurons beaucoup moins l'intuition qu'un Mr. Puppet « positif » n'est pas libre. Imaginons en effet qu'un individu reçoive une éducation spéciale destinée à le rendre heureux et extrêmement bon. Dirions-nous que ce « bon » Mr. Puppet ne sera pas libre et ne méritera pas notre approbation lorsqu'il aidera un aveugle à traverser la rue ?

5.2. Nouvel examen du passage de (C1) à (C2) : encore une fois, montrer que la responsabilité requise par la conception rétributive est incompatible avec le déterminisme ne suffit pas à rendre fausse cette conception

Ainsi, les arguments de XB ne semblent pas suffisants pour rejeter le compatibilisme. De plus, l'argument normatif souffre du même problème que l'argument descriptif : il suppose à tort que montrer que le déterminisme est vrai et que notre conception de la responsabilité morale est incompatible avec le déterminisme suffit à rejeter la conception rétributive de la peine comme inadéquate. Mais, comme on l'a vu, c'est loin d'être le cas : la conception rétributive de la peine peut être vraie (et avoir un sens) même si la responsabilité morale est impossible dans notre monde soumis au déterminisme – elle a juste pour conséquence que personne n'est responsable. C'est d'ailleurs la conclusion qu'adoptent un certain nombre de philosophes partisans du *déterminisme dur*.

Il faut donc un argument supplémentaire pour montrer que la vérité du déterminisme rend nécessaire le rejet de la conception rétributive de la peine, un critère permettant de montrer la supériorité de la conception utilitariste sur la conception rétributive. Or, ce critère ne peut provenir du fait que la conception utilitariste de la peine est vraie alors que la conception rétributive est fausse : selon XB, parce qu'il n'existe pas de faits moraux,

alors aucune théorie morale (ou conception de la peine) n'est vraie (mais, au contraire, toutes les théories morales sont également fausses)[44]. De plus, à première vue, aucune des deux conceptions ne semble souffrir de contradictions internes ou avec les données scientifiques suffisantes pour la rejeter : comme nous l'avons dit, la conception rétributive de la peine est cohérente avec la vérité du déterminisme. Bien sûr, elle a pour conséquence que personne ne mérite d'être puni. Mais, à moins d'être prêt à vouloir rejeter une théorie pour des raisons pratiques, on voit mal en quoi cela serait une raison pour l'abandonner[45].

Ainsi, l'argument de XB repose sur une prémisse philosophiquement discutable (et discutée) mais, en plus, il n'atteint pas la conclusion qu'il vise : la conception rétributive de la peine n'est pas incompatible avec le déterminisme, c'est le fait d'être responsable qui l'est. Or, la conception rétributive de la peine peut survivre à l'inexistence de la responsabilité : elle ne suppose pas son existence (juste que personne ne mérite d'être puni s'il n'est pas responsable).

VI. Conclusion

G&C et XB ont proposé deux arguments assez proches mais différents : l'un concluant que les neurosciences nous conduiront *effectivement* à abandonner la conception rétributive de la peine pour adopter la conception utilitariste de la peine, l'autre concluant que les neurosciences nous montrent que nous *devons* abandonner la conception rétributive pour adopter la conception utilitariste. Dans cet article, j'ai essayé de montrer que ces arguments reposent sur des prémisses douteuses et que le passage de ces prémisses à la conclusion n'est pas toujours garanti. Ainsi, l'argument de G&C repose sur des prémisses empiriques douteuses :

[44] XB adopte ainsi ce que l'on appelle en méta-éthique la « théorie de l'erreur », selon laquelle tous les énoncés moraux sont faux parce que (i) tous les énoncés moraux font référence à des faits moraux objectifs et (ii) de tels faits moraux n'existent pas. Pour une défense de cette thèse, voir : J.L. Mackie, *Ethics: Inventing Right and Wrong*, Londres, Penguin, 1990. Pour une discussion empirique de la proposition (i), voir : F. Cova et J. Ravat, « Sens commun et objectivisme moral : objectivisme « global » ou objectivisme « local » ? Une introduction par l'exemple à la philosophie expérimentale », in *Klesis*, 9, 2008, pp.180-202.

[45] Bien sûr, on peut proposer plein d'autres arguments visant à montrer en quoi la conception rétributive de la peine souffre de problèmes et de contradictions internes, et c'est notamment ce que fait XB dans les autres sections de son ouvrage déjà cité. Mais, même si ces arguments sont au final suffisants pour nous pousser à rejeter la conception rétributive de la peine, il n'en reste pas moins que la vérité du déterminisme n'aura joué aucun rôle dans ce rejet. Si les arguments montrant des contradictions internes sont suffisants, ils n'ont pas besoin de l'appel au déterminisme, qui, comme nous avons insisté, ne prouve rien. Si donc nous finissons par accepter qu'il faut modifier le droit pénal, ce ne sera pas parce que les neurosciences ont prouvé que le déterminisme est vrai.

les études empiriques ne suggèrent (i) ni que nous considérons la liberté comme incompatible avec le déterminisme, (ii) ni que croire que nous ne sommes pas libres nous conduirait à adopter une conception utilitariste de la punition. L'argument de XB, lui, repose sur une prémisse philosophique très controversée (la liberté est incompatible avec le déterminisme) et ne suffit pas à montrer la supériorité de l'approche utilitariste de la peine sur l'approche rétributive. Au final, il n'est donc pas clair que les neurosciences nous conduisent ou nous forcent logiquement à réformer en profondeur le droit pénal.

Remerciements

Je tiens à remercier Patrick Ducray et deux relecteurs anonymes pour leurs commentaires sur des versions antérieures de ce manuscrit, ainsi que Xavier Bébin pour son livre dont la lecture a été très stimulante. Ce travail a été financé par une bourse de l'Agence Nationale pour la Recherche (ANR Blanche: SoCoDev).