De la neurocriminologie au neurodroit : le droit pénal à l'ère du sujet cérébral

Marie Penavayre

De la neurocriminologie au neurodroit : le droit pénal à l'ère du sujet cérébral

In H. KASSOUL & A. CUKIER (dir.), *Les Rencontres de Thémis et Sophia (2eme éd.), Nature de l'Homme, nature du droit, Colloque Droit et Philosophie*, Université de Poitiers, 2021

MARIE PENAVAYRE
Docteure en Épistémologie et histoire des sciences
Chercheuse associée à l'Institut des Maladies Neurodégénératives
Université de Bordeaux

Résumé : Depuis une dizaine d'années, on assiste à la construction d'un discours qui propose de se saisir des outils neuroscientifiques pour informer les questions fondamentales du droit pénal. Cet article propose de déconstruire sur le plan épistémologique le présupposé selon lequel le niveau de description neurobiologique permettrait de rendre compte du phénomène criminel. En cherchant à dépasser une lecture strictement anti-réductionniste ou anti-naturaliste, je montre pourquoi le cadre théorique et méthodologique des recherches expérimentales échoue à fournir une explication satisfaisante pour répondre aux questions de droit pénal.

Mots-clés : neurocriminologie ; neurodroit ; neuroessentialisme ; cerveau criminel

Au vu des progrès scientifiques majeurs réalisés dans la compréhension du fonctionnement cérébral, il n'est pas surprenant que l'on assiste parfois à un décalage entre les attentes sociales vis-à-vis des neurosciences et le pouvoir explicatif réel de leurs outils ou résultats[1]. La place prépondérante qui leur est accordée dans l'explication de nos comportements s'est notamment illustrée par la construction d'un neuro-discours sur le phénomène criminel, qui propose de se saisir des outils neuroscientifiques pour répondre aux questions fondamentales du droit pénal. Le débat qui a émergé en France suite à la révision de la loi de Bioéthique de 2011 – qui introduisait une disposition légale sur l'usage judiciaire des neurosciences – doit être clairement distingué des deux projets scientifiques qui caractérisent ce nouveau champ d'application aux États-Unis : le neurodroit d'une part, qui défend l'apport des techniques de neuroimagerie dans le cadre d'une procédure d'irresponsabilité pour cause psychiatrique[2], et la neurocriminologie d'autre part, qui propose de redéfinir le problème de la criminalité en termes neurobiologiques, en isolant les spécificités cérébrales des individus représentant une menace pour la société[3].Bien que le neurodroit et la neurocriminologie se distinguent du point de vue des finalités recherchées et du cadre méthodologique qui régit les études expérimentales, les deux approches se recoupent dans l'intérêt exclusif porté au cerveau dans l'étude du phénomène criminel. Partant de l'idée que le cerveau est nécessaire à l'exercice des capacités cognitives, le neurodroit défend le recours à l'imagerie cérébrale pour objectiver l'altération des capacités requises à l'attribution de responsabilité pénale (la capacité à contrôler son comportement et la capacité à apprécier la valeur morale de son acte). La neurocriminologie s'inscrit quant à elle dans une logique éliminativiste vis-à-vis des approches sociales ou

1 D. FOREST, *Neuropromesses : une enquête philosophique sur les frontières des neurosciences,* Ithaque, 2022.

2 O. JONES, R. MAROIS, M. J. FARAH et H. T. GREELY, « Law and Neuroscience » *Journal of Neuroscience* 33 (45), 2013, p. 17624-17630.

3 A. RAINE, M. S. BUCHSBAUM, J. STANLEY, S. LOTTENBERG, L. ABEL, et J. STODDARD. « Selective reductions in prefrontal glucose metabolism in murderers. » *Biological Psychiatry* 36 (6), 1994, p. 365-373.

sociologiques de la criminalité : son modèle explicatif propose de faire du cerveau l'objet d'étude privilégié pour comprendre le phénomène criminel, en excluant toute interprétation alternative à l'explication neurobiologique des comportements violents, impulsifs ou agressifs.

En somme, ces deux voies de recherches se structurent autour d'un programme de naturalisation des comportements, du crime, et des catégories juridiques elles-mêmes. Elles tendent à opérer une réduction du sujet de droit à un « sujet cérébral » objectivable et mesurable par les techniques de neuroimagerie. En France, les résistances exprimées à l'égard d'un tel projet tiennent au fait qu'il se double d'une ambition prescriptive : celle de fonder les catégories normatives du droit pénal (notamment l'évaluation de la responsabilité et du risque de récidive) sur des critères neurobiologiques. L' « argument neuro » et le postulat d'objectivité scientifique concourent à l'idée selon laquelle les techniques de neuroimagerie permettraient de combler les insuffisances de l'expertise psychiatrique en matière pénale, entretenant une grande confusion entre ce que ces techniques permettent d'expliquer ou seulement de décrire.

Aux États-Unis, la procédure accusatoire facilite le recours à des preuves scientifiques pour statuer sur la responsabilité ou la dangerosité d'un individu. Les attentes sociales et politiques grandissantes à l'égard des progrès neuroscientifiques font donc craindre l'avènement d'un « solutionnisme neuroscientifique » au sein même de la procédure pénale américaine, par la tentation de projeter sur les techniques de neuroimagerie une solution directe, concrète et objective aux difficultés posées par l'expertise psychiatrique, jugée trop subjective et source de contradictions. Mais peut-on réellement donner un sens à l'ambition d'utiliser les outils neuroscientifiques pour informer les questions fondamentales du droit pénal ? Les normes de responsabilité pénale sont-elles accessibles à l'investigation empirique ? Et surtout, quelle conception du droit pénal et du sujet de droit est mobilisée en arrière-plan de ce projet ?

Il est difficile de prendre la juste mesure du projet de neurodroit sans renvoyer dos-à-dos les limites de l'expertise psychiatrique et la critique d'un solutionnisme neuroscientifique déconnecté de la réalité des pratiques

judiciaires en matière pénale. De même, il est tentant d'opposer à la neurocriminologie un pluralisme explicatif qui viendrait invalider son présupposé neuro-réductionniste et sa logique éliminativiste vis-à-vis des approches externalistes (psycho-sociales, socio-économiques, etc.). La limite de cette position est qu'elle nous amènerait à produire un contre-discours inaudible du point de vue des partisans de ce projet, lesquels défendent la prééminence de l'explication neurobiologique à grands renforts d'études IRMf rapportant un lien entre le comportement criminel et la présence d'anomalies cérébrales[4]. Sans nécessairement renverser la chaîne de causalité en mobilisant d'autres stratégies explicatives, il existe de bonnes raisons de penser que le niveau de description neurobiologique ne permet pas de rendre compte du phénomène criminel, ni de répondre à la question de la responsabilité pénale. On peut d'ailleurs admettre que certains états cérébraux sont nécessaires à l'exercice des capacités cognitives, et s'opposer dans le même temps aux représentations neuroessentialistes qui invoquent la réalité empirique d'un « cerveau criminel » ou d'un « cerveau pénalement responsable ». Le cheminement argumentatif que nous proposons de suivre vise à questionner le présupposé fondateur du projet de neurocriminologie et de neurodroit : l'idée selon laquelle le sujet de droit est « d'essence cérébrale » et qu'il est dès lors possible d'expliquer les conditions d'émergence d'un comportement criminel en étudiant uniquement les états cérébraux de l'individu. Cette problématique nous invite à interroger deux thèses : 1/ la thèse selon laquelle le comportement criminel est imputable au seul cerveau ; et 2/ la thèse selon laquelle les conditions d'émergence des capacités cognitives sont internalisées dans des états cérébraux spécifiques. La première thèse est issue de ce qui est communément

4 On dénombre plusieurs centaines d'études proposant d'utiliser les techniques de neuroimagerie comme un instrument de connaissance du sujet criminel. Si les premières publications remontent au début des années 1940 avec l'émergence de l'électroencéphalographie (EEG), ce programme de recherche s'est surtout développé dans les années 1990 sous l'impulsion d'Adrian Raine, l'une des figures majeures de la neurocriminologie contemporaine. A. RAINE, P. H. VENABLES, et M. WILLIAMS. « Relationships between central and autonomic measures of arousal at age 15 years and criminality at age 24 Years. » Archives of General Psychiatry 47 (11), 1990, p. 1003–1007.

appelé le « cérébro-centrisme », c'est-à-dire la tendance à se référer exclusivement au cerveau dans l'étude des états mentaux ou des comportements. La seconde correspond à une conception internaliste des processus cognitifs, à savoir l'idée que les conditions de l'activité mentale sont internes au cerveau et qu'elles s'expriment sous la forme d'une potentialité préexistante.

Nous montrerons ainsi successivement pourquoi le niveau de description neurobiologique est insuffisant pour fournir une explication du comportement criminel, et dans quelle mesure il devient possible d'argumenter contre l'ambition de fonder les catégories normatives du droit pénal sur des critères neurobiologiques. Précisons d'ores et déjà que notre objectif n'est pas tant de contester la pertinence de l'approche cérébro-centrique – en montrant que les causes du comportement criminel sont situées « en dehors du cerveau » – mais plutôt de mettre en doute l'idée que les critères d'appréciation de la dangerosité ou de la responsabilité pénale sont accessibles à l'investigation neuroscientifique. Nous verrons en effet que la stratégie consistant à défendre une position « externaliste » vis-à-vis des processus psychologiques ne suffit pas à déconstruire le projet du neurodroit et de la neurocriminologie ; encore faut-il montrer pourquoi le cadre théorique et méthodologique de ces études échoue à fournir une explication causale et mécaniste des comportements étudiés. En clair, plutôt que d'adopter un rejet de principe à l'égard des potentielles applications judiciaires des neurosciences, on peut aussi faire le choix méthodologique de prendre au sérieux les ambitions du neurodroit et de la neurocriminologie, en proposant un pas de côté par rapport à une perspective anti-réductionniste ou anti-naturaliste.

Notre analyse porte donc à la fois sur la thèse épistémologique de l'internalisme – la thèse selon laquelle il est possible d'expliquer les conditions d'émergence de la criminalité en étudiant uniquement les états internes (neurobiologiques) du sujet criminel – et sur l'ambition prescriptive du neurodroit et de la neurocriminologie. Il convient d'ailleurs de lever une confusion éventuelle sur la cible de notre analyse critique : elle concerne bien les applications judiciaires des neurosciences, et non pas les neurosciences en elles-

mêmes. À notre sens, le projet du neurodroit et de la neurocriminologie soulèvent des problèmes métaphysiques et épistémologiques qui ne sont pas inhérents à la méthodologie des neurosciences en tant que telles, mais bien plutôt à l'exportation de leurs outils ou de leurs résultats hors de leur cadre d'élaboration. Les neurosciences ont bel et bien une valeur épistémique, mais elles n'épuisent pas les questions relatives à l'explication d'un comportement social complexe, et encore moins les questions normatives mobilisées dans le champ pénal.

***

I. Pourquoi le criminel n'est pas « d'essence cérébrale »

A. Étudier le cerveau seul ne suffit pas : l'argument externaliste

Les critiques dirigées contre l'intérêt exclusif porté au cerveau peuvent mobiliser plusieurs voies argumentatives, qui s'étendent de la philosophie pragmatiste ou externaliste jusqu'aux connaissances empiriques fournies par les sciences biologiques[5]. On peut par exemple s'appuyer sur des arguments tirés de la biologie de l'évolution ou de la biologie du développement et défendre une position « interactionniste », c'est-à-dire concevoir les phénomènes cognitifs comme le résultat d'une co-construction entre différents niveaux d'organisation (de la génétique jusqu'à l'environnement physique et social de l'individu). À ce titre, il parait trivial de noter que le cérébro-centrisme tel qu'il est défendu dans la neurocriminologie et le neurodroit est difficilement compatible avec la multiplication des études menées sur le système nerveux entérique, qui montrent que la cognition n'est pas concentrée dans le seul cerveau : privilégier

5 Pour une analyse détaillée des différentes critiques qui peuvent être formulées à l'encontre du cérébro-centrisme, nous renvoyons le lecteur au troisième chapitre de D. FOREST, *Neuroscepticisme*, Ithaque, 2014.

le cerveau seul dans l'étude des phénomènes cognitifs ne permet pas de rendre compte de la dynamique des interactions entre l'environnement et les processus physico-chimiques et biologiques, ni de comprendre le rôle joué par les réseaux neuronaux situés en dehors du cerveau.Deux arguments peuvent être mobilisés pour illustrer les limites d'une conception internaliste des processus mentaux : l'argument de l'externalisme sémantique tel qu'il a été initialement formulé par Hilary Putnam, et le modèle de l'esprit étendu, de la cognition incarnée ou située, selon lequel les processus cognitifs se distribuent dans d'autres parties du système nerveux, dans le corps et l'environnement de l'individu.

1. Le modèle internaliste à l'épreuve de l'externalisme sémantique

À bien des égards, la neurocriminologie et le neurodroit favorisent la représentation d'un individu réduit à un « cerveau dans une cuve », isolé du reste de son corps et de son environnement physique et social[6]. La rhétorique du « *my brain made me do it defense* » (ou « c'est pas moi, c'est mon cerveau ») – régulièrement mobilisée dans le champ du neurodroit – contribue à effacer les dimensions psycho-affectives et sociologiques du crime, en autorisant un déplacement de la responsabilité de l'individu vers son cerveau. De même, la neurocriminologie s'est formée autour du présupposé selon lequel le criminel est « d'essence cérébrale » : l'individu dangereux serait porteur d'une criminalité immanente, intrinsèque à son organisation cérébrale. Ce présupposé neuroessentialiste réduit *in fine* le sujet à ses déterminations internes, à des propriétés intrinsèques qui font sens à l'intérieur de son cerveau.Or, une telle conception de la nature humaine peut conduire à une

6 En privilégiant le cerveau seul dans l'étude des comportements, la neurocriminologie et le neurodroit marquent un retour à un nouveau dualisme « cerveau – reste du sujet » : le cérébro-centrisme conduit en définitive à substituer la substance pensante et immatérielle de Descartes (la *res cogitans*) par une entité interne, matérielle (le cerveau). Comme nous le montrerons à la fin de cette section, cette conception est problématique parce qu'elle se rend coupable d'une erreur de catégorie : on fait l'erreur d'attribuer à la partie (le cerveau) des propriétés qui relèvent d'un tout (l'individu dans son environnement).

forme d'éliminativisme vis-à-vis des états mentaux car elle tend à concevoir ces derniers comme des épiphénomènes, c'est-à-dire des évènements ne jouant aucun rôle dans la chaîne causale qui mène à la réalisation des comportements. De ce point de vue, l'idée selon laquelle le cerveau est une condition suffisante à l'exercice des processus cognitifs nous amènerait à céder définitivement à une explication en termes de causes et non de raisons d'agir. Elle conduirait à neutraliser la pertinence des états intentionnels – pourtant nécessaires à l'explication de nos comportements et à l'attribution de responsabilité pénale – pour les éliminer au profit de ses déterminations internes.

Certaines expériences de pensée, notamment l'expérience dite de la « Terre-Jumelle » ou la théorie du « cerveau dans une cuve » permettent d'appréhender les faiblesses d'une conception internaliste des processus cognitifs. Ces expériences de pensée ont été formulées par Hilary Putnam dans le cadre d'une critique de la théorie computationnelle de l'esprit, laquelle conçoit les états mentaux comme des états computationnels, c'est-à-dire comme un système de manipulation de symboles. De ce point de vue, les états computationnels possèdent un contenu conceptuel internalisé dans le cerveau, à l'image d'un « cerveau dans une cuve » dont les représentations mentales ne seraient que le produit de stimuli délivrés par un ordinateur[7].Putnam montre que les états mentaux (et en particulier les états intentionnels tels que les croyances, les désirs ou les intentions) n'ont pas de référence intrinsèque, c'est-à-dire qu'ils ne font pas sens à l'intérieur de notre cerveau, indépendant de l'environnement du sujet et de son activité : « les significations ne sont pas dans

7 L'hypothèse selon laquelle « je suis (en réalité) un cerveau dans une cuve » permet d'illustrer la théorie computationnelle de l'esprit. Partant de cette hypothèse sceptique radicale, notre monde extérieur, notre corps (et nos états mentaux eux-mêmes) ne seraient que le produit des informations délivrées par un ordinateur à notre cerveau. La croyance selon laquelle « je suis un cerveau dans une cuve » serait donc elle-même une construction intrinsèque à notre cerveau. Putnam montre que cette hypothèse est auto-réfutante : si j'étais un cerveau dans une cuve (c'est-à-dire en admettant le présupposé du fonctionnalisme computationnel), je ne pourrais jamais me référer à une véritable cuve ni à un véritable cerveau... je serais donc incapable de formuler et penser cette hypothèse. Les états intentionnels n'ont pas de référence intrinsèque car ils sont en partie déterminés par notre environnement.

la tête »[8]. Sans pour autant défendre l'autonomie explicative des raisons d'agir vis-à-vis du niveau de description neurobiologique[9], on peut en effet contester l'idée qu'elles possèdent une référence intrinsèque dans la mesure où elles sont en grande partie déterminées par l'environnement physique et social de l'individu. Il ne s'agit pas ici de défendre l'irréductibilité des propriétés intentionnelles des états mentaux, mais seulement de contester l'idée qu'elles sont intrinsèques au cerveau : elles ne font sens qu'en vertu des relations que l'individu entretient avec son environnement.

2. Vers une extension corporelle des processus cognitifs ?

Par ailleurs, la référence au seul cerveau dans l'explication de nos comportements s'oppose à un courant de pensée bien connu dans le champ de la philosophie critique des neurosciences, qui défend l'« extension corporelle » des processus cognitifs. La critique du cérébro-centrisme et de l'internalisme peut en effet puiser ses arguments dans la grande famille des théories dites de la « cognition incarnée » de l' « énactivisme » ou de l' « esprit étendu », que l'on retrouve sous diverses formes chez Francisco Varela[10], Mark Rowlands[11], ou

8 H. PUTNAM, « The Meaning of "Meaning" » In *Mind, Language and Reality. Philosophical Papers, 2.* Cambridge University Press, 1975, p. 227.

9 Ce problème renvoie à la distinction entre le niveau personnel (les explications par les raisons) et le niveau subpersonnel (les explications par les causes). La relation entre ces deux niveaux d'explication fait l'objet de nombreux débats en philosophie de l'esprit. Globalement, la question est de savoir si les propriétés intentionnelles des états mentaux peuvent être réduites à des causes de nature physique, et en définitive, à des explications neurobiologiques. Certains philosophes prônent une forme de séparatisme entre les deux niveaux, c'est-à-dire qu'ils défendent l'autonomie explicative du niveau personnel vis-à-vis du niveau subpersonnel. On peut au contraire défendre l'idée que le niveau d'explication personnel n'est pas irréductible et qu'il n'est pleinement satisfaisant que si l'on prend en compte le niveau subpersonnel. Voir notamment : D. DENNETT, *Content and Consciousness*, Routledge, 1969.

10 F. VARELA, *The embodied mind: Cognitive science and human experience*, MIT Press, 1991.

11 M. ROWLANDS, *The body in mind: Understanding cognitive processes*, Cambridge University Press, 1999.

encore chez Alva Noë[12]. Ces théories s'accordent à replacer le cerveau dans son contexte biologique et écologique, et à situer les processus cognitifs dans l'environnement conçu au sens large. La cognition s'étend en dehors du cerveau et se distribue dans d'autres parties du système nerveux, dans le corps et l'environnement de l'individu. Pour reprendre un exemple cité par le philosophe Alva Noë, il n'y a donc pas plus de sens à situer les phénomènes mentaux dans le (seul) cerveau que de localiser la vision dans l'œil ou le processus de digestion dans l'estomac : ce dernier, comme le cerveau, n'est qu'un organe parmi d'autres, une partie d'un système comprenant de multiples niveaux d'organisation, qui interagissent entre eux pour accomplir leurs fonctions[13].Alva Noë a consacré son ouvrage *Out of Our Heads* (2009) à la question de savoir « pourquoi vous n'êtes pas votre cerveau ». Sa thèse s'inscrit plus largement dans le cadre d'une critique de la théorie de l'identité du mental et du cérébral, de l'internalisme, et plus spécifiquement, de la possibilité empirique d'une explication neurobiologique de la conscience. Selon lui, les phénomènes mentaux (et en particulier les états conscients) désignent « quelque chose que nous faisons », et non pas « quelque chose qui se produit en nous ». L'idée selon laquelle la conscience est internalisée dans le cerveau n'est pas seulement dénuée de fondement empirique ; elle est en elle-même dépourvue de sens car « le cerveau n'a pas d'activité cognitive en propre »[14]. La thèse de Noë prend ainsi la forme d'un rejet de principe contre toute assimilation des processus cognitifs à des états cérébraux, et contre toute tentative visant à localiser les corrélats neuronaux de la conscience. Les neurosciences échouent selon lui à expliquer les propriétés fonctionnelles de la conscience (c'est-à-dire à identifier son rôle causal), car la seule référence au cerveau ne rend pas compte de la dynamique de l'interaction entre le reste du corps, l'activité sensorimotrice et l'environnement au sens large.La question des limites topologiques que l'on peut assigner à la cognition fait l'objet d'un vif

12 A. NOË, *Out of our heads: Why you are not your brain, and other lessons from the biology of consciousness*, Hill & Wang, 2009.
13 *Idem.*
14 *Op.cit*. p. 24.

débat, de même que la fécondité explicative des théories de l'énaction ou de l'esprit étendu. On peut en effet voir dans ces théories un anti-naturalisme de principe (ou du moins une certaine imperméabilité face aux connaissances neuroscientifiques) qui conduit, de fait, à minimiser le rôle joué par le cerveau dans l'exercice des processus cognitifs, et surtout à rendre caduque (ineffective) la recherche des fondements empiriques de leur extension corporelle. Selon les philosophes Fred Adams et Ken Aizawa, l'extension crânienne ou corporelle des processus cognitifs est une possibilité logique, mais elle n'a pas de réalité empirique[15]. Ils défendent ce qu'ils appellent un « intracrânialisme contingent » : à défaut d'identifier empiriquement les processus cognitifs qui ont lieu en dehors du corps, on est contraint d'admettre qu'ils se produisent à l'intérieur du cerveau. Sans nécessairement souscrire à cette thèse, on peut considérer que les théories de l'énaction et de l'esprit étendu ne suffisent pas à invalider le modèle internaliste tel qu'il est défendu par la neurocriminologie et le neurodroit. Tout au plus permettent-elles de récuser la thèse ontologique selon laquelle les processus psychologiques sont localisés exclusivement dans le cerveau des individus. Mais qu'en est-il de la thèse épistémologique, à savoir l'idée qu'il est possible d'expliquer les conditions d'émergence de la criminalité en étudiant uniquement les états internes du criminel ?

On ne saurait trop insister sur le fait qu'en privilégiant le cerveau seul, on s'interdit de saisir la complexité des relations qui l'unissent à l'environnement au sens large. Pour autant, défendre l'idée que les causes du comportement criminel sont situées en dehors du cerveau ne fait que contourner le problème. L'intérêt exclusif porté au cerveau dans l'étude du comportement criminel est contestable pour les raisons que nous avons mentionnées ci-dessus. Mais la stratégie consistant à défendre une position externaliste vis-à-vis des processus cognitifs ne suffit pas à invalider la thèse internaliste défendue par le neurodroit et la neurocriminologie, ni à déconstruire l'ambition de fonder les catégories normatives du droit pénal (notamment la question de la responsabilité) sur des critères neurobiologiques.

15 F. ADAMS et K. AIZAWA, « The bounds of cognition. » *Philosophical Psychology* 14 (1), 2001, p. 43-64.

Car ce projet est bien celui d'une « neuroessentialisation » du crime et des conditions requises à l'attribution de la responsabilité pénale. L'analyse critique d'un tel projet nous invite donc à dépasser le problème ontologique du cérébro-centrisme pour questionner la thèse épistémologique et normative : elle nécessite d'analyser les raisons pour lesquelles il est épistémologiquement erroné de faire du crime un problème *exclusivement* neurobiologique.

Les critiques que nous avons adressées à l'encontre du cérébro-centrisme ou de l'internalisme nous amènent *a minima* à défendre une conception dite « pluraliste » de l'explication des phénomènes mentaux. De ce point de vue, la fécondité de l'explication nécessite d'articuler le niveau de description neurobiologique avec ce que nous savons en dehors du cerveau. Elle exige l'instauration d'un dialogue avec des disciplines connexes qui étudient les mêmes mécanismes, mais à des niveaux de description différents[16]. Par exemple, l'affirmation selon laquelle la perte du contrôle de soi résulte d'un déficit d'activité au niveau de certaines régions cérébrales[17] n'est vérifiée que si elle est cohérente avec l'ensemble_des résultats obtenus par différentes méthodes d'investigation ou sources de connaissance. Le niveau de description neurobiologique ne nous dit rien sur le rôle fonctionnel d'une région cérébrale dans la perte du contrôle de soi, car il doit pouvoir être mis en relation avec les résultats obtenus sur d'autres niveaux de description (psychologiques, sociologiques, etc.). Ainsi, si l'on veut préciser le statut de la relation entre un déficit neurologique et une altération comportementale, il conviendrait de mettre en perspective les résultats issus de la neuroimagerie avec les modèles utilisés en psychologie cognitive. Qu'est-ce qui caractérise une perte de contrôle

16 Voir notamment L. FAUCHER et P. POIRIER, « Psychologie évolutionniste et théories interdomaines », *Dialogue: Canadian Philosophical Review* 40 (3), 2001, p. 453-486.

17 Les résultats de la neurocriminologie montrent que les populations criminelles sont associées à des déficits d'activité au niveau des régions associées à la régulation émotionnelle ou à l'inhibition comportementale. Voir par exemple E. AHARONI, G.M. VINCENT, C.L. HARENSKI, V.D. CALHOUN, W. SINNOTT-ARMSTRONG, M.S. GAZZANIGA et K.A. KIEHL, « Neuroprediction of future rearrest. » *Proceedings of the National Academy of Sciences* 110 (15), 2013, p. 6223–6228.

de soi sur le plan comportemental ? Quels paradigmes expérimentaux permettent d'opérationnaliser cette incapacité, à la fois dans sa dimension motrice (l'incapacité à inhiber une action motrice) et cognitive (l'incapacité à envisager les conséquences de ses actions sur le long terme) ?Cette position ne nous engage donc pas à défendre l'idée que les processus mentaux sont inaccessibles à l'investigation neuroscientifique. Il ne s'agit pas pour nous de nier l'influence causale du cerveau dans l'exercice des facultés psychologiques, ni l'idée que ces derniers relèvent d'un traitement interne. Il s'agit plutôt d'une part, de contester le présupposé selon lequel *seuls* les états internes jouent un rôle dans l'exercice des facultés, et d'autre part, de montrer pourquoi le niveau de description neurobiologique ne permet pas de répondre aux questions fondamentales du droit pénal.

B. Les cerveaux ne commettent pas de crimes : l'argument de l'erreur méréologique

1. C'est l'individu qui agit, et non son cerveau

Afin de mieux saisir le problème posé par l'ambition de fonder les catégories normatives du droit pénal sur des critères neurobiologiques, il convient de placer notre analyse critique sur un second niveau. Dans la mesure où la neurocriminologie et le neurodroit s'intéressent à des capacités fonctionnelles qui incluent une large série de volitions (la capacité à former une intention d'action, à former des raisons d'agir, à contrôler son action, etc.), on peut se demander dans quelle mesure les états cérébraux du sujet nous informent sur l'exercice de ces processus psychologiques. Sans verser dans un behaviorisme naïf qui refuserait d'admettre que ces derniers désignent autre chose que des comportements, on peut soutenir qu'ils ne se traduisent pas seulement par la manifestation d'un état cérébral donné : considérer que les conditions de l'activité mentale sont internes au cerveau ne nous dit rien sur la manière dont les processus psychologiques sont réalisés, car ces derniers doivent pouvoir s'exprimer au travers d'un comportement.

Dans cette perspective, il peut être intéressant de s'attarder sur l'argument de l'erreur méréologique tel qu'il a été proposé par Max Bennett et Peter Hacker dans le cadre de leur analyse critique du projet des neurosciences[18]. Cet argument prend pour cible tout type de discours consistant à attribuer au cerveau des prédicats psychologiques, une erreur que Bennett et Hacker jugent constitutive du discours neuroscientifique. Selon eux, les résultats neuroscientifiques s'accompagnent d'un « neuro-discours » qui consiste à faire du cerveau un sujet, c'est-à-dire à lui allouer des propriétés ou des capacités que l'on attribue généralement aux individus[19]. Un des exemples cités par Bennett et Hacker est donné par un argument d'Antonio Damasio dans *L'erreur de Descartes* (1994), lorsqu'il proclame que « nos cerveaux peuvent prendre de bonnes décisions en quelques secondes »[20]. On peut également citer la neurophilosophe Patricia Churchland – l'une des grandes figures de la philosophie éliminativiste – qui affirme que « c'est le cerveau [...] qui ressent, qui pense, qui décide »[21].S'il est vrai que les neuroscientifiques cèdent parfois à ce type d'erreur grammaticale, il convient d'insister sur le fait que ce neuro-discours est surtout prégnant chez les réductionnistes éliminativistes. Il est d'ailleurs largement repris dans la littérature consacrée aux neurodisciplines telles que la neuroéducation ou le neuromarketing, lesquelles ont tout intérêt à montrer que c'est le cerveau (et le cerveau seul) qui « apprend », « résout des problèmes », « répond aux stimuli visuels », « prend la décision d'acheter », etc. Dans le champ de la neuro-criminologie, il est fréquent de rencontrer ce discours consistant à faire du cerveau un sujet, car il permet de réaffirmer la

18 M. BENNETT et P. HACKER, *Philosophical foundations of neuroscience*, Wiley, 2003.
19 Cet argument rappelle l'argument de l'erreur de catégorie de Gilbert Ryle, formulé dans le cadre de sa critique du dualisme cartésien. Ryle illustrait son argument par cet exemple : « un étudiant visite une université, avec sa bibliothèque, ses laboratoires, ses amphithéâtres, etc. avant de demander « mais où est l'université ? ». Cet étudiant fait une erreur méréologique en considérant que l'université appartient à la même catégorie que l'ensemble des éléments qui la constitue.
20 A. DAMASIO, *L'erreur de Descartes: la raison des émotions*, Odile Jacob, 1994, p. 69.
21 P. CHURCHLAND, *Brain-wise: studies in Neurophilosophy*. MIT Press, 2002, p. 2.

prééminence de l'explication neurobiologique sur les autres niveaux de descriptions (psychologiques et sociologiques).

Comme le soulignent Bennett et Hacker, ce type d'énoncé n'est pas faux sur le plan logique ; il n'a simplement aucune signification cognitive. Il est dénué de sens car le cerveau ne satisfait pas les critères d'attribution des prédicats psychologiques : le sujet du comportement n'est pas le cerveau mais l'individu lui-même, lequel doit être envisagé dans toutes ses dimensions et sa complexité. Autrement dit, en cherchant à allouer au cerveau des capacités psychologiques, on se rend coupable d'un sophisme méréologique puisque l'on attribue à la partie (le cerveau) des propriétés qui relèvent d'un tout (l'individu dans son environnement) : c'est l'individu qui pense et agit, et non pas son cerveau. Le cerveau ne réalise pas lui-même nos pensées ou nos comportements car il n'est qu'une partie de nous-mêmes. Il ne renferme pas en lui des intentions ou des raisons d'agir, mais seulement des activités électriques[22]. À titre d'exemple, la formation d'une intention d'agir dépend d'un large ensemble de capacités cognitives (la capacité à élaborer un plan d'action, à délibérer consciemment et rationnellement sur ses choix d'action, à exercer un contrôle sur ces derniers, etc.)[23], lesquelles ne peuvent être réduits à une seule description neurobiologique.

2. L'inadéquation ontologique entre le langage neuroscientifique et le langage juridique

L'argument de l'erreur de catégorie a été repris par les philosophes Michael S. Pardo et Dennis Patterson, dans un ouvrage consacré à l'analyse des fondements conceptuels et empiriques du projet du neurodroit[24]. Selon eux, cet argument constitue le principal rempart contre tous les projets d'intégration des neurosciences dans le champ pénal : les recherches empiriques sont dénuées de sens car elles se déploient dans la confusion méréologique la plus totale. Sans

22 A. NOË, 2004, *op.cit.*

23 E. PACHERIE, *Naturaliser l'intentionnalité*, Presses Universitaires de France, 1993.

24 M. S. PARDO et D. PATTERSON. *Minds, Brains, and Law: The Conceptual Foundations of Law and Neuroscience,* Oxford University Press. 2013.

partager entièrement leurs conclusions – qui s'apparentent à notre sens à un rejet de principe à l'égard des neurosciences en tant que telles – certaines méritent d'être mentionnées ici car elles permettent de mieux cerner le problème posé par l'ambition de fonder les catégories normatives du droit pénal sur des catégories neurobiologiques.Leur analyse offre une perspective générale sur les erreurs conceptuelles et inférentielles qui sous-tendent la relation établie entre les trois niveaux de description mobilisés dans le champ du neurodroit : le niveau de description neurobiologique (notamment les informations obtenues par l'imagerie cérébrale pendant la réalisation d'une tâche cognitive), les concepts psychologiques relatifs aux capacités fonctionnelles des individus (la capacité de discernement, la capacité à contrôler son comportement...), et les concepts normatifs relevant du droit pénal (l'appréciation de la responsabilité ou de la dangerosité). Les recherches en neurodroit sont en effet dictées par l'ambition d'exporter dans le champ pénal les résultats neuroscientifiques qui portent sur des concepts mentaux ayant une pertinence vis-à-vis de la question de la responsabilité pénale (l'intention, le contrôle de soi, la connaissance morale, etc.). En clair, le neurodroit assoit sa légitimité scientifique sur le fait que le droit pénal mobilise un large éventail de concepts issus de la psychologie populaire, qui sont justement accessibles à l'investigation neuroscientifique. Pour Pardo et Patterson (2013), l'ambition d'intégrer les neurosciences dans le système pénal apparaît alors essentiellement comme un problème conceptuel ou catégoriel : celui de la traduction d'un concept empirique (relatif par exemple à un dysfonctionnement dans les régions impliquées dans le contrôle de soi) en un concept éminemment normatif (en l'occurrence une preuve de l'irresponsabilité pour cause psychiatrique)[25]. Autrement dit, selon eux, le problème inhérent au projet du neurodroit est celui de l'inadéquation ontologique entre le langage neuroscientifique et le langage juridique issu de la psychologie populaire. Les catégories du droit pénal sont inaccessibles à

25 Nous faisons ici référence à l'un des projets du neurodroit, qui vise à utiliser les données issues de la neuroimagerie pour informer la condition volitionnelle de l'*Insanity Defense* (ou défense d'aliénation mentale), c'est-à-dire apporter la preuve que le prévenu est incapable de contrôler son comportement.

l'investigation empirique car elles mobilisent des concepts normatifs qui ne peuvent être traduits en termes naturels.Pour reprendre un argument régulièrement cité par Pardo et Patterson, l'ambition d'intégrer les neurosciences dans le système pénal repose notamment sur l'idée qu'il serait possible d'inférer la possession d'une connaissance (qu'elle soit morale, juridique ou pratique) en enregistrant l'activité cérébrale d'un individu. Cette idée est largement partagée par les chercheurs en neurodroit car le problème de la responsabilité pénale pose un certain nombre de questions relatives à la notion de connaissance. La connaissance des règles morales ou juridiques est effectivement un prérequis à l'attribution de responsabilité pénale : l'évaluation de la responsabilité nécessite de déterminer si le prévenu est capable de former un jugement moral (et donc s'il a une connaissance morale), et s'il est capable de contrôler son comportement conformément à cette connaissance.

Du point de vue des chercheurs en neurodroit, l'imagerie cérébrale apparait alors comme un outil précieux pour évaluer la capacité d'un prévenu à apprécier la valeur morale de son acte, mais aussi pour préciser le degré de *mens rea* (c'est-à-dire estimer si le prévenu a agi sciemment et en connaissance de cause, ou par simple négligence ou imprudence)[26] ou pour déterminer s'il a connaissance de certaines informations relatives au crime pour lequel il est poursuivi. Cette dernière ambition s'est formée au début des années 1990, à partir des recherches menées par le neuroscientifique Lawrence Farwell sur ce qu'il a appelé lui-même le *Brain Fingerprinting Test* [27]. Ce test propose de

26 La notion de connaissance est particulièrement importante lorsqu'il s'agit d'examiner l'élément moral de la responsabilité pénale (la *mens rea*) et se prononcer ainsi sur la culpabilité ou non d'un prévenu. Dans les juridictions de *common law*, il existe en effet quatre degrés d'intention coupable selon le niveau de connaissance (ou d'incertitude) face aux risques associés à l'acte : *purpose* / *knowledge* / *recklessness* / *negligence*. Des études publiées dans le domaine du neurodroit ont montré que ces degrés d'intention étaient associés à des patterns spécifiques d'activité cérébrale, suggérant la possibilité d'utiliser l'IRMf pour déterminer si le prévenu avait ou non connaissance des risques encourus. Voir notamment : I. VILARES et al. "Predicting the knowledge–recklessness distinction in the human brain", *Proceedings of the National Academy of Sciences* 114 (12), 2017, p. 3222-3227.

27 L. FARWELL et E. DONCHIN. "The truth will out: interrogative polygraphy (lie detection) with event-related brain potentials", *Psychophysiology* 28 (5), 1991, p. 531-547.

reconnaître dans l'activité EEG d'un individu la « trace mémorielle » d'une information donnée, en enregistrant le profil d'activité de l'onde p300 : cette onde, qui survient 300 millisecondes après la présentation d'un stimulus, manifeste une réponse caractéristique lorsque ce dernier est familier. Aux États-Unis, ce test a été utilisé dans certaines procédures pénales, pour déterminer si le prévenu avait ou non connaissance des informations relatives au crime pour lequel il était poursuivi (l'arme du crime, la scène du crime, etc.) Pour ne citer qu'un seul exemple, ce test a été reconnu comme une preuve scientifique pour disculper un individu, Terry Harrington, qui avait été condamné pour homicide volontaire, le rapport d'expertise réalisé par Lawrence Farwell ayant largement contribué à sa libération en 2000, après 22 ans de détention[28]. Ce rapport révélait que « le cerveau de Harrington ne contenait pas d'informations relatives au crime » : l'onde p300 ne manifestait aucune réponse caractéristique d'un stimulus familier, mais présentait un profil d'activité spécifique lors de la présentation des informations relatives à son alibi. S'il n'est pas de notre propos ici de discuter de la fiabilité de cette technique ni de l'acceptabilité de la preuve elle-même, retenons qu'elle illustre parfaitement la confusion méréologique qui règne dans l'ambition d'intégrer les outils neuroscientifiques dans le système pénal. De toute évidence, le principe de fonctionnement du *Brain Fingerprinting Test* (BFT) repose sur l'idée que la connaissance d'une information est internalisée dans le cerveau sous la forme d'un engramme mémoriel. Ce présupposé rend légitime le neuro-discours consistant à faire du cerveau un sujet – « le cerveau de Harrington ne contient pas » les informations relatives au crime pour lequel il était poursuivi – et il justifie par la même occasion l'idée que le cerveau seul nous renseigne sur la question de savoir si l'individu a connaissance ou non des informations qui lui sont transmises : d'une donnée empirique, on infère la possession d'une connaissance (un attribut psychologique) qui nous renseignerait sur la culpabilité ou non du prévenu.

28 Précisons que l'usage de l'EEG s'inscrit ici dans un contexte très particulier, favorisé par la pression de l'opinion publique pour libérer le détenu, Terry Harrington, en raison d'une insuffisance de preuves.

Comme le montrent Pardo et Patterson (2013), le raisonnement consistant à inférer la possession d'une connaissance à partir des états cérébraux d'un individu relève d'une erreur méréologique, dans la mesure où les états neurologiques ne remplissent pas les critères d'attribution de la connaissance : la proposition selon laquelle « j'ai connaissance d'une certaine information » n'est pas satisfaite si l'on tient compte uniquement de mes états cérébraux. Les états internes de l'individu ne suffisent pas à lui attribuer une quelconque connaissance car ils ne rendent pas compte de la manière dont elle s'exprime sur le plan comportemental. Les critères d'attribution de la connaissance se rapportent en effet à des comportements (ce que l'on fait ou ce que l'on dit) et non à des états neurologiques. Ainsi selon Pardo et Patterson, le raisonnement qui sous-tend le principe de fonctionnement du BFT est dénué de sens car « la connaissance n'est pas localisée dans le cerveau ». Il repose sur une conception erronée de la connaissance, laquelle mobilise un ensemble de capacités cognitives (notamment la capacité à reconnaître une proposition comme vraie, la capacité à former des croyances, etc.), qui doivent pouvoir s'exprimer au travers d'un comportement.

Précisons d'ailleurs que les états neurologiques n'apportent aucune information sur les relations que la notion de connaissance entretient avec d'autres concepts tels que la croyance, le doute, la certitude, etc. Et quand bien même les données constitueraient un « indicateur neurophysiologique » de la possession d'une connaissance (dans sa définition la plus générale), il resterait encore à déterminer si elles nous disent quoi que ce soit sur la connaissance au sens où elle est mobilisée dans le droit pénal[29].

29 On peut d'ailleurs ajouter que le raisonnement consistant à inférer l'exercice d'un état psychologique à partir de l'activité cérébrale correspond à une inférence inverse, qui est très largement discutée dans le champ des neurosciences et de la philosophie des neurosciences. Ce raisonnement est limité par la spécificité fonctionnelle des régions considérées : une telle inférence est valide si et seulement si la région R est activée uniquement lorsque la fonction F est engagée. On peut donc douter de sa validité en ce qui concerne l'attribution d'une connaissance, dans la mesure où l'on s'intéresse à des fonctions cognitives complexes et fortement intégrées, qui engagent de multiples réseaux d'interactions. Voir à ce sujet R. POLDRACK, "Can cognitive processes be inferred from neuroimaging data." *Trends in Cognitive Sciences* 10, 2006, p. 59-63.

Sans mettre en doute le bien-fondé de l'analyse proposée par Pardo et Patterson, on peut s'interroger cependant sur la portée de l'argument de l'erreur méréologique pour rejeter définitivement l'ambition de redéfinir les catégories normatives du droit pénal à la lumière des données neuroscientifiques. On peut en effet questionner la pertinence d'une analyse centrée sur les problèmes conceptuels ou catégoriels posés par ce projet, et qui fait fi des conditions expérimentales dans lesquelles les recherches sont conduites.

Force est de constater que certaines recherches publiées dans le domaine du neurodroit livrent une réflexion très fine sur la nécessité de proposer une clarification des concepts mobilisés dans le champ, afin de garantir la validité conceptuelle de leurs protocoles expérimentaux et d'éviter d'éventuelles erreurs inférentielles[30]. Cet effort de clarification conceptuelle s'exprime notamment au travers d'une réflexion théorique sur les critères de définition du contrôle de soi tel qu'il est défini dans le droit pénal américain d'une part, et tel qu'il peut être opérationnalisé sur le plan expérimental d'autre part. Ces recherches visent à construire un paradigme expérimental qui recouvre les aspects volitionnels et cognitifs du contrôle de soi, c'est-à-dire à la fois la capacité à contrôler ses pulsions et la capacité à contrôler son comportement conformément à une connaissance morale ou juridique.

En concentrant notre critique autour du cérébro-centrisme ou de l'inadéquation ontologique entre le langage neuroscientifique et le langage juridique, on fait donc l'impasse sur les questions proprement empiriques concernant les modalités d'opérationnalisation des comportements ou des processus mentaux. Comment le contrôle de soi est-il conceptualisé et opérationnalisé dans les recherches expérimentales ? Quels choix méthodologiques président à la construction de ces modèles ? En clair, si l'on veut questionner l'apport potentiel des neurosciences dans le système judiciaire,

30 Voir notamment J. BUCKHOLTZ, V. F. REYNA et C. SLOBOGIN, "A neuro-legal lingua franca: Bridging law and neuroscience on the issue of self-control." *Mental Health, Law and Policy Journal ; Vanderbilt Public Law Research Paper* No 16, 2016.

il convient de dépasser toute hostilité de principe pour interroger le cadre théorique et méthodologique qui régit les études publiées dans le champ.

***

II. Les raisons de dépasser le rejet de principe

A. Les insuffisances de l'argument de l'erreur méréologique

L'argument de l'erreur méréologique nous a permis d'apporter quelques éléments de réponse pour invalider le modèle internaliste défendu par la neurocriminologie et le neurodroit, et contester l'idée que les états internes produisent à eux seuls les phénomènes psychologiques. On peut pourtant se demander si cet argument suffit à rejeter le projet d'intégration des neurosciences dans le champ criminologique et judiciaire. Depuis la publication de leur *Philosophical foundations of neuroscience* par Bennett et Hacker[31] les neuroscientifiques sont nombreux à contester la portée de cet argument, le neuro-discours auquel ils s'adonnent parfois n'étant selon eux qu'une simple « analogie métaphorique et anecdotique »[32]. On peut supposer que les chercheurs en neurodroit ou en neurocriminologie partagent cette position, et que l'erreur de catégorie n'est selon eux qu'un faux problème issu d'une « philosophie en fauteuil » qui ne tient pas compte des spécificités méthodologiques de leurs recherches.

Plus encore, on peut regretter que la radicalité de cette thèse masque les enjeux réels posés par leur projet, précisément parce qu'elle n'est pas suffisamment informée sur le plan empirique. La question de la connaissance (et des divers usages qui sont envisagés dans le champ du neurodroit) est largement traitée dans l'ouvrage de Pardo et Patterson. Elle éclipse pourtant la question plus fondamentale encore de l'apport des neurosciences pour informer les conditions volitionnelles et cognitives de l'*Insanity Defense*, c'est-à-dire

31 M. BENNETT et P. HACKER, 2003, *Op.cit.*
32 N. LEVY, "Is neurolaw conceptually confused?" *The Journal of Ethics* 18 (2), 2014, p. 171-185.

déterminer si le prévenu possède les capacités requises pour être tenu pénalement responsable, à savoir : la capacité à contrôler son comportement et la capacité à apprécier la valeur morale de son acte.

Pardo et Patterson estiment que les neurosciences n'apportent aucune information vis-à-vis des concepts psychologiques mobilisés dans le champ pénal, car le principe méréologique des neurosciences interdit toute identification du sujet (de ses comportements ou de ses états psychologiques) à son cerveau ou à ses régions cérébrales : l'erreur méréologique suffit selon eux à invalider le projet du neurodroit. S'il est vrai que cette confusion soulève de sérieux doutes quant à sa validité épistémologique, nous voudrions ici nuancer la thèse selon laquelle les neurosciences n'apportent aucune information pertinente vis-à-vis du droit pénal. Pour le dire simplement, nous souscrivons à l'argument, mais nous estimons qu'il est insuffisant pour tirer une conclusion définitive sur le projet d'intégrer les neurosciences dans le champ pénal. Deux raisons peuvent être invoquées.

1. Un problème épistémologique

La première est que la thèse de Pardo et Patterson se trompe de cible. Elle s'inscrit dans la continuité des réflexions menées dans le champ de la philosophie critique des neurosciences, et relève en ce sens davantage d'un rejet de principe de l'apport des neurosciences vis-à-vis des concepts mentaux, que d'une analyse approfondie des arguments empiriques avancés par les chercheurs en neurodroit. En outre, si le projet du neurodroit est bien marqué par des confusions conceptuelles, celles-ci ne sont pas suffisantes pour conclure que *les neurosciences* en général n'apportent aucune information pertinente vis-à-vis du droit pénal. Le problème ne concerne pas les neurosciences en tant que telles, mais la solidité du cadre théorique et méthodologique dans lequel les recherches en neurodroit sont conduites.

D'ailleurs, et sans nécessairement chercher à sauver ce dernier, on peut noter que leur thèse sous-estime la complexité des études menées dans le champ, lesquelles ne peuvent se réduire à la recherche d'une localisation cérébrale des concepts juridiques issus de la psychologie populaire (l'intention, le contrôle de

soi, la connaissance morale, etc.). Le principal enjeu posé par le neurodroit est de déterminer dans quelle mesure les données neuroscientifiques peuvent informer les capacités requises pour être tenu pénalement responsable de son acte. Le problème est donc bel et bien épistémologique : il ne s'agit pas tant de savoir si les propriétés cérébrales peuvent être identifiées à des propriétés psychologiques, mais de déterminer si les données neurobiologiques peuvent nous dire quoi que ce soit sur les facultés nécessaires à l'attribution de responsabilité pénale.

2. Un problème empirique

Cette remarque nous amène à la seconde raison pour laquelle l'argument de l'erreur méréologique ne suffit pas à invalider le projet du neurodroit. La question de savoir si les données neuroscientifiques peuvent résoudre des problèmes normatifs n'est pas seulement conceptuelle ou catégorielle ; elle est aussi et surtout empirique. Il y a bien une erreur de catégorie dans le raisonnement consistant à inférer la possession d'un état psychologique à partir des états cérébraux de l'individu, puisque l'on confond les propriétés de la partie (les propriétés cérébrales) avec les propriétés du tout (les propriétés psychologiques ou comportementales). Mais le problème n'est pas tant l'inadéquation ontologique entre le langage neuroscientifique et le langage psychologique ou juridique, que la confusion entretenue dans l'interprétation des données empiriques, entre les états psychologiques et leurs conditions d'exercice[33]. En cherchant à inférer la possession d'une connaissance (ou d'une intention) à partir des états cérébraux de l'individu, on assimile les capacités psychologiques des individus avec les conditions d'exercice de ces capacités, sans donner accès aux mécanismes neurobiologiques impliqués dans cette relation. Comment, dès lors, pourrait-on statuer sur la relation entre un état cérébral donné et l'altération des capacités requises à l'attribution de responsabilité pénale ?

33 Voir à ce sujet V. DESCOMBES, *La denrée mentale*, Editions de Minuit, 1995.

La question de l'apport de la neuroimagerie dans l'appréciation de la responsabilité pénale ne nécessite pas seulement d'établir un lien entre l'acte criminel et la présence d'une anomalie cérébrale (une lésion, une atrophie, un dysfonctionnement…). Il s'agit plutôt de déterminer si cette anomalie fournit une explication causale et mécaniste des incapacités fonctionnelles de l'individu (altération du discernement ou du contrôle de soi)[34] : dans quelle mesure ces capacités peuvent être affectées, et en vertu de quoi les données neuroscientifiques peuvent nous renseigner sur l'altération de ces capacités. Autrement dit, si l'on veut questionner la valeur de preuve de la neuroimagerie dans le cadre d'une procédure d'irresponsabilité pénale, il ne suffit pas d'identifier les conditions épistémologiques pour lesquelles un dysfonctionnement cérébral peut justifier une atténuation de la responsabilité. Il faut également déterminer si les données neuroscientifiques permettent de reconstruire la chaîne causale des évènements qui produisent l'altération des capacités requises à l'attribution de responsabilité.

À titre d'exemple, une explication mécaniste d'une perte du contrôle de soi supposerait d'identifier l'ensemble des mécanismes neurobiologiques qui jouent un rôle causal dans chacune des dimensions (motrices, cognitives, morales…) du comportement : l'incapacité à inhiber ses pulsions, à envisager les conséquences de ses actions sur le long terme, à contrôler son comportement conformément à des raisons morales, etc. L'objectif serait d'identifier *quel* mécanisme produit le comportement, mais aussi *par quel moyen* il est physiquement réalisé dans le cerveau[35]. Autrement dit, si l'on veut fournir une explication mécaniste d'une perte du contrôle de soi, il faudrait que l'on soit

34 S'il existe un important débat concernant le type d'explication le plus opérant en neurosciences, il est généralement admis qu'elles font appel à des explications de type mécaniste. Une explication est mécaniste lorsqu'elle permet de rendre compte du comportement d'un système complexe en identifiant les fonctions exécutées par ses parties, les modalités de leurs interactions et les mécanismes impliqués dans leur réalisation. Voir notamment W. BECHTEL et R.C. RICHARDSON, *Discovering complexity: Decomposition and localization as strategies in scientific research*, Princeton University Press, 1993.

35 Voir à ce sujet W.C. SALMON, *Scientific explanation and the causal structure of the world*, Princeton University Press, 1984.

capable d'accéder au fonctionnement interne du mécanisme par lequel la cause produit son effet.

De là apparaissent trois problèmes épistémologiques majeurs, qui nous invitent à interroger plus précisément le statut de la preuve neuroscientifique : 1/ les résultats de la neuroimagerie apportent-ils une explication mécaniste des incapacités fonctionnelles des individus, notamment d'une perte du contrôle de soi ? (le problème de l'explication) ; 2/ le paradigme expérimental utilisé pour mesurer cette capacité permet-il de rendre compte du comportement tel qu'il est exercé dans un environnement social complexe ? (le problème de l'opérationnalisation) ; et 3/ les résultats peuvent-ils nous dire quoi que ce soit sur le concept tel qu'il est mobilisé dans le droit pénal, c'est-à-dire sur un concept éminemment normatif qui s'intègre dans un environnement social complexe et qui dépend des catégories juridiques dans lesquelles il est invoqué ? (le problème de la translation).

Par leur complémentarité mutuelle, ces trois problèmes permettent de mieux appréhender la question de l'apport de la neuroimagerie vis-à-vis des questions fondamentales du droit pénal. Ils invitent en effet à préciser le statut épistémologique de la relation entre les trois niveaux de description qui sont au cœur de cette problématique : les informations obtenues par l'imagerie cérébrale (le niveau neurobiologique), la preuve d'une altération du contrôle de soi ou du discernement (le niveau psychologique ou comportemental), et l'appréciation de la responsabilité et de la dangerosité de l'individu (le niveau juridique et normatif).

B. Des impasses conceptuelles et empiriques

L'ambition d'intégrer les outils de neuroimagerie dans le système pénal américain repose sur plusieurs centaines d'études utilisant l'IRMf pour identifier les régions cérébrales impliquées dans la capacité à contrôler son comportement ou la capacité à apprécier la valeur morale de son acte. Quelles

que soient les finalités poursuivies – utiliser la neuroimagerie pour informer la question de la responsabilité pénale (dans le cas du neurodroit) ou la question du risque de récidive (dans le cas de la neurocriminologie) – le problème de la portée explicative des données est donc indissociable de la question de la pertinence des tâches cognitives utilisées dans les protocoles expérimentaux pour mesurer les comportements cibles (généralement, le contrôle de soi et la capacité de discernement).

Du point de vue des chercheurs en neurodroit ou en neurocriminologie, la construction du paradigme expérimental repose sur deux hypothèses : d'une part, l'idée que les tâches cognitives permettent de mesurer les capacités fonctionnelles des individus, et d'autre part, que les performances individuelles sont équivalentes aux comportements tels qu'ils sont exercés dans un environnement social complexe. La question de la pertinence explicative des tâches utilisés peut alors être abordée sur un plan conceptuel et empirique, en interrogeant à la fois la validité conceptuelle du modèle – en examinant les critères de définition du comportement que l'on cherche à isoler – et la validité de la mesure, c'est-à-dire en déterminant si le modèle expérimental est approprié pour mesurer ce comportement, et s'il le représente sous tous ses aspects.

Les contraintes épistémologiques qui pèsent sur l'opérationnalisation d'un comportement social complexe nous invitent également à analyser la portée explicative et normative des résultats : en admettant que les données empiriques nous renseignent sur le comportement cible (le contrôle de soi ou le discernement), permettent-elles de tirer des conclusions normatives sur la question de la responsabilité pénale ou du risque de récidive de l'individu ?

1. La question de la validité conceptuelle et de la validité de la mesure

À partir de ces différents niveaux d'analyse, on peut questionner les modalités d'opérationnalisation du contrôle de soi, telles qu'elles sont mises en œuvre dans les recherches en neurocriminologie et en neurodroit pour évaluer

la capacité d'un prévenu à exercer un contrôle sur ses actions. La plupart des études ont recours à la *Go/NoGo task*, qui permet d'évaluer la capacité d'un individu à inhiber une action motrice[36]. Rares sont celles qui utilisent des tâches cognitives permettant de mesurer la dimension prospective du contrôle de soi comme la *Delay Discounting Task*, qui invite le sujet à faire le choix entre une petite récompense immédiate ou une grosse récompense tardive, ou la *Iowa gambling task*, qui permet d'évaluer la capacité à évaluer la balance bénéfice/risque de ses actions[37].

Si ces paradigmes expérimentaux posent eux-mêmes un certain nombre de difficultés épistémologiques, on peut considérer qu'ils sont nettement plus appropriés pour opérationnaliser le comportement cible, lequel est porteur d'une dimension intrinsèquement sociale et prospective. La *Go/NoGo task* mesure en effet uniquement la cause proximale de l'action (la capacité à inhiber une « pulsion » motrice) et ne rend pas compte de la dynamique en amont de son initiation, ni de la dynamique intersubjective du comportement. De fait, ce modèle sous-spécifie le comportement tel qu'il est exercé dans l'environnement naturel, car les chercheurs font l'impasse sur les questions relatives à la définition de leur objet expérimental : ils échouent à isoler les critères de définition du concept qu'ils cherchent à opérationnaliser, dépouillant le contrôle de soi de sa dimension prospective, dynamique, morale, socio-affective, etc. Les régions cérébrales isolées pendant la réalisation de la *Go/NoGo task* ne correspondent donc pas au comportement cible. Elles peuvent nous donner des informations sur les corrélats neurobiologiques impliqués dans la capacité à inhiber une action motrice, mais rien ne garantit qu'elles sont mises en jeu dans un environnement social complexe ou lors du processus de planification de l'action sur le long terme.

De ce point de vue, la possibilité d'utiliser ce paradigme expérimental pour mesurer un comportement aussi complexe que le contrôle de soi paraît

36 Lors de cette tâche, le participant est invité à effectuer une action motrice (par exemple, appuyer sur un bouton) en réponse à un certain stimulus (condition *Go*), et à inhiber cette action en réponse à un autre stimulus (condition *NoGo*).

37 Voir notamment W. MISCHEL et E.B. EBBESEN, "Attention in delay of gratification." *Journal of Personality and Social Psychology* 16 (2), 1970, p. 329-337.

doublement injustifiée : non seulement parce que les conditions de validité conceptuelle et de validité de la mesure ne sont pas remplies, mais aussi parce que le manque de spécificité fonctionnelle des régions isolées limite considérablement la portée explicative des données de neuroimagerie.

2. La question de la portée explicative et normative des résultats

Finalement, on pourrait se satisfaire du modèle *Go/NoGo* s'il avait pour seule ambition d'isoler les régions cérébrales impliquées dans l'inhibition d'une action motrice. Cependant, dès lors que les résultats sont mobilisés pour informer le contrôle de soi tel qu'il est défini dans le droit pénal, on se heurte inévitablement à un problème translationnel : le paradigme expérimental utilisé dans les études n'apporte aucune information sur la définition juridique du contrôle de soi, lequel est envisagé à la fois dans sa dimension volitionnelle, cognitive et morale. L'attribution de responsabilité pénale suppose en effet que l'individu soit capable d'inhiber une pulsion irrésistible, mais aussi d'envisager les conséquences de ses actions sur le long terme, de délibérer consciemment et rationnellement sur ses choix d'action, de contrôler son comportement conformément à la loi ou à des raisons morales, etc.[38] On peut dès lors douter que l'identification des régions cérébrales corresponde bien au comportement cible, précisément parce que les variables contextuelles, socio-affectives et morales du comportement sont exclues des conditions expérimentales.

Le problème central est donc le suivant : les contraintes épistémologiques propres à l'opérationnalisation d'un comportement complexe effacent les aspects intersubjectifs et normatifs du comportement

38 Notons que le système pénal américain repose sur une conception dite « capacitarienne » de la responsabilité, héritée de la « *capacity-responsability theory* » de Herbert Hart : l'attribution de la responsabilité repose sur la possession (et l'intégrité) de certaines capacités, notamment la capacité à contrôler son comportement et à apprécier la valeur morale de son acte. H.L.A. HART, *Punishment and Responsibility: Essays in the Philosophy of Law*, Oxford University Press, 1968.

cible (qui lui sont pourtant inhérents), alors même que l'évaluation des capacités requises à une agentivité responsable ne peut être appréhendée qu'en vertu de ce contexte. Autrement dit, on ne peut pas s'assurer que les données issues de la neuroimagerie nous renseignent sur la capacité à contrôler son comportement, dans la mesure où la sélection des régions d'intérêt repose sur un paradigme expérimental artificiel, décontextualisé, qui sous-spécifie le comportement cible.

On pourrait d'ailleurs invoquer les mêmes arguments en ce qui concerne les modalités d'opérationnalisation de la capacité de discernement. Le paradigme expérimental repose généralement sur la présentation d'un dilemme sacrificiel inspiré du « dilemme du tramway », initialement formulé par Philippa Foot en 1967[39] puis décliné dans différentes versions par Judith Thomson[40]. Le participant doit indiquer s'il est selon lui moralement approprié de sacrifier une personne pour sauver un plus grand nombre, dans différentes circonstances, notamment dans des situations mettant en jeu des facteurs émotionnels. La version classique du dilemme se présente ainsi :

> « Vous êtes au volant d'un tramway hors de contrôle qui approche un aiguillage. Sur la voie de gauche se trouve un groupe de cinq ouvriers de maintenance. Sur la voie de droite se trouve un seul ouvrier de maintenance. Si vous ne faites rien, le trolley poursuivra sa route vers la gauche, causant la mort des cinq ouvriers. La seule manière d'éviter la mort de ces ouvriers est d'actionner un levier sur le tableau de bord qui déviera le trolley sur la voie de droite, causant la mort de l'ouvrier isolé. Est-il selon vous moralement approprié d'actionner le levier afin d'éviter la mort des cinq ouvriers ? »

Ici encore, il se pose la question de savoir : 1/ si ce modèle est pertinent vis-à-vis du concept psychologique que l'on cherche à étudier (la capacité à

39 P. FOOT. "The problem of abortion and the doctrine of double effect", *Oxford Review* 5, 1967.
40 J. J. THOMSON, "The Trolley Problem", *The Yale Law Journal*, 94(6), 1975.

évaluer la valeur morale de son acte) ; 2/ si les données recueillies pendant la résolution de ces dilemmes garantissent une explication d'une altération ou d'une abolition du discernement ; et 3/ si elles permettent de tirer des conclusions normatives utiles à l'évaluation du risque de récidive ou de la responsabilité pénale. Or, la mesure de l'activité cérébrale pendant la résolution d'un dilemme sacrificiel ne permet pas de rendre compte de la composante cognitive et volitionnelle de nos jugements moraux, c'est-à-dire à la fois la capacité à mobiliser ses connaissances morales et juridiques, et à contrôler son comportement conformément à cette connaissance[41]. En outre, même si l'on admet que les jugements moraux (au même titre que le contrôle de soi) peuvent faire l'objet d'une investigation neuroscientifique, on est forcé d'admettre que l'ambition de la neurocriminologie et du neurodroit n'est pas à la hauteur de leurs moyens théoriques, méthodologiques et empiriques.

En analysant les modalités d'opérationnalisation des comportements, telles qu'elles sont mises en œuvres dans les recherches en neurocriminologie et en neurodroit, on peut ainsi multiplier les arguments pour invalider l'ambition de fonder les catégories normatives du droit pénal sur des critères neurobiologiques. Le problème est d'abord conceptuel, car les chercheurs échouent à identifier les critères de définition du concept qu'ils cherchent à opérationnaliser (le contrôle de soi, la connaissance morale, l'intention d'agir, etc.). Les conclusions empiriques s'en trouvent affaiblies car les confusions conceptuelles peuvent introduire un *gap* entre l'objet qui est opérationnalisé et celui qui est réellement mesuré. En conséquence, et sans compter les erreurs inférentielles qui peuvent émerger de l'interprétation des données, la possibilité de tirer des conclusions normatives sur la seule base de ces données paraît doublement injustifiée. Non seulement en raison des confusions conceptuelles et empiriques qui sous-tendent la construction des modèles expérimentaux, mais aussi parce que les modalités d'opérationnalisation génèrent un biais définitionnel entre le comportement opérationnalisé, le comportement tel qu'il

41 M. PENAVAYRE, C. BRUN et T. BORAUD, "Neurobiologie des jugements moraux, avancée épistémique ou voie sans issue ?" *Intellectica*, No. 70, 2019, p. 63-82.

est mis en jeu dans un environnement social complexe, et le comportement auquel on se réfère dans un cadre juridique spécifique.

***

Conclusion

Nous avons présenté un certain nombre d'arguments tirés de la philosophie externaliste et de la philosophie critique des neurosciences, pour s'opposer à la thèse neuroessentialiste défendue par la neurocriminologie et le neurodroit. Si le cerveau joue un rôle causal incontestable dans l'exercice de nos processus cognitifs, cela ne signifie pas pour autant que le niveau de description neurobiologique permet de rendre compte de l'ensemble de nos manifestations comportementales ou de nos raisons d'agir. À la lumière des différentes objections que nous avons formulées à l'encontre de ce présupposé, on peut donc légitimement mettre en doute la validité du raisonnement qui consiste à fonder l'appréciation de la dangerosité ou de la responsabilité sur des critères neurobiologiques : ce saut explicatif n'est pas justifié du point de vue épistémologique.

De toute évidence, les débats qui entourent l'usage judiciaire des neurosciences entretiennent une certaine confusion du point de vue des enjeux réels posés par ces nouveaux programmes de recherche. Ces derniers nous invitent à questionner le statut épistémologique de la preuve neuroscientifique, notamment en analysant la portée explicative des données issues de la neuroimagerie. Mais ils soulignent également l'impérieuse nécessité de mener une réflexion conceptuelle et méthodologique sur les modalités d'opérationnalisation des comportements, telles qu'elles sont mises en œuvre dans les recherches expérimentales. En clair, si l'on veut prendre la juste mesure des enjeux de la neurocriminologie et du neurodroit, il n'est pas suffisant d'opposer à ces recherches un anti-naturalisme de principe, lequel viendrait épuiser la question de l'apport de la neuroimagerie dans l'inaccessibilité empirique des catégories juridiques. Encore faudrait-il caractériser le statut

épistémologique de la relation entre un état cérébral donné et l'altération des capacités requises à l'attribution de responsabilité : tant que l'on ne précisera pas le statut de cette relation sur le plan causal, mécaniste et explicatif, la neuroimagerie ne sera d'aucun secours pour rendre compte du phénomène criminel et répondre aux questions fondamentales du droit pénal.