

Une introduction au problème difficile de la conscience

Partie 1 : Du cogito cartésien aux zombies de Chalmers : une introduction historique au problème difficile de la conscience

Cette série de deux textes se veut une introduction accessible à l'un des problèmes centraux de la philosophie de l'esprit contemporaine : ce que David Chalmers a nommé le « problème difficile de la conscience ». Plutôt que d'en proposer immédiatement une définition abstraite, j'ai choisi d'en retracer la genèse. Le problème difficile s'inscrit dans une longue histoire des tentatives visant à comprendre la place de l'esprit et de la conscience dans le monde physique, et notamment, le rapport entre l'esprit et le cerveau.

Je partirai donc du dualisme cartésien et du problème de l'interaction entre l'esprit et le corps, avant de suivre les transformations successives de cette question : montée du matérialisme puis du physicalisme, développement de la physiologie et des neurosciences, principe de clôture causale, théorie de l'identité et fonctionnalisme. Cette trajectoire permettra de comprendre pourquoi, même après que l'explication physique et fonctionnelle du comportement a considérablement progressé, une difficulté semble demeurer : pourquoi ces processus s'accompagnent-ils d'une expérience subjective ?

L'objectif de ce premier article est donc à la fois historique et conceptuel : montrer comment le problème difficile a émergé et parvenir progressivement à sa formulation contemporaine. Un second article sera consacré aux principales réponses qui lui ont été proposées et à leurs coûts respectifs. Je ne prétendrai pas y adopter une neutralité artificielle : certaines solutions me paraissent plus convaincantes que d'autres, mais leur comparaison viendra seulement après que le problème lui-même aura été clairement posé.

I. Une porte d'entrée commode : Descartes

Dans ses *Méditations métaphysiques*, Descartes construit en plusieurs étapes un argument pour établir que l'esprit est une substance distincte de la matière. [Descartes (1641)]

Il part, dans la Méditation II, **du doute méthodique** : je peux douter de tout, y compris de l'existence de mon corps, mais je ne peux pas douter que je pense pendant que je doute. Cette certitude n'est pas le fruit d'un raisonnement déductif — Descartes insiste sur son caractère immédiat : « je pense, donc je suis » s'impose à l'esprit par intuition directe, sans passer par une prémisse générale préalable. C'est le cogito. Descartes en conclut qu'il est d'abord une chose qui pense — un esprit, un entendement — en évitant à ce stade le mot « âme », trop chargé du vocabulaire scolastique dont il cherche à se démarquer.

Cette première conclusion ne suffit pas encore à établir la distinction réelle entre pensée et matière. Pouvoir concevoir que mon corps n'existe pas ne garantit pas qu'il soit réellement possible que ma pensée existe séparée de lui : de la concevabilité à la possibilité réelle, il y a

un écart à combler. C'est l'enjeu des Méditations III à V, où Descartes établit l'existence d'un Dieu non trompeur : cette garantie assure que tout ce qui est conçu clairement et distinctement correspond à une possibilité réelle, puisque Dieu peut créer tout ce qui est ainsi conçu.

Cette garantie divine (outre le fait qu'elle repose sur la très contestée preuve ontologique de l'existence de Dieu) soulève cependant une difficulté célèbre, formulée notamment par Antoine Arnauld dans les Quatrièmes Objections : Descartes ne semble-t-il pas raisonner en cercle ? Pour démontrer l'existence d'un Dieu non trompeur, il s'appuie sur des raisonnements qu'il tient pour clairs et distincts ; mais c'est ensuite l'existence de ce Dieu non trompeur qui doit garantir que tout ce qu'il perçoit clairement et distinctement est vrai. La règle des idées claires et distinctes paraît donc être utilisée avant même que sa fiabilité ait été établie. Descartes répond qu'une perception claire et distincte s'impose avec certitude tant qu'elle est actuellement présente à l'esprit, et que la garantie divine n'est nécessaire que pour assurer la fiabilité durable de son souvenir. Cette réponse permet d'éviter la forme la plus grossière du cercle, mais elle reste discutée : l'argument en faveur de Dieu semble lui-même exiger une chaîne de raisonnements dont la fiabilité ne peut être maintenue en permanence par une intuition présente. [Arnauld (1641) ; Descartes (1641)]

C'est alors seulement, dans la Méditation VI, que Descartes peut conclure : puisqu'il conçoit clairement et distinctement son esprit comme entièrement séparable de son corps, et que cette conception claire et distincte garantit une possibilité réelle, la pensée et la matière sont deux substances réellement distinctes. C'est le **dualisme cartésien**.

Ce dualisme n'allait cependant pas de soi comme point de départ. Thomas Hobbes, contemporain de Descartes et auteur des Troisièmes Objections aux Méditations dès 1641, défendait déjà une **position matérialiste** opposée : pour lui, tout ce qui existe, y compris ce que nous appelons pensée ou volonté, n'est que corps en mouvement, sans qu'il soit besoin de postuler une seconde substance distincte de la matière. [Hobbes (1641)]

Ce dualisme pose cependant un problème, pointé par la princesse Élisabeth de Bohême dans sa correspondance avec Descartes : comment deux substances de nature différente peuvent-elles interagir causalement, de manière à ce que notre pensée soit affectée par ce qui arrive à notre corps, et que notre corps soit mû par notre pensée ? Descartes propose d'abord un site physiologique à cette interaction, **la glande pinéale**, mais cette réponse ne touche pas au cœur du problème conceptuel soulevé par Élisabeth. Le fait que la même glande vaguement identifiée ne dise en réalité rien du *mécanisme* par lequel une substance sans étendue pourrait mouvoir une substance étendue, oblige finalement Descartes à changer de terrain plutôt qu'à répondre frontalement à Élisabeth. Il concède qu'on ne peut concevoir l'union de l'âme et du corps avec les mêmes outils que ceux qui permettent de les distinguer : la distinction relève de l'entendement pur, tandis que l'union ne se laisse saisir que par l'usage de la vie et des sens — la faim, la douleur, le désir en sont l'expérience directe, non la démonstration. Il va jusqu'à suggérer qu'à trop vouloir méditer métaphysiquement sur cette union, on finit par obscurcir ce qu'on savait déjà confusément par le simple fait de vivre incarné. Ce déplacement n'est pas un pur aveu d'échec —

Descartes maintient que sa doctrine n'a rien à corriger, c'est l'exigence de clarté rationnelle qui serait mal dirigée sur cet objet. Mais il n'apporte à aucun moment de mécanisme causal, et Élisabeth ne se déclare jamais pleinement convaincue : l'échange se prolonge dans leur correspondance, jusqu'à irriguer plus tard les *Passions de l'âme*. La difficulté qu'elle a mise au jour est restée, aux yeux de la plupart des commentateurs, l'un des points où le dualisme interactionniste cartésien montre sa fragilité la plus nette. [Élisabeth de Bohême & Descartes (1643)]

II. Trois réponses au déficit de l'interactionisme

Ce problème de l'interaction, laissé ouvert par Descartes, va structurer une bonne part de la métaphysique du XVII^e siècle : chacun des grands continuateurs du cartésianisme va chercher à s'en débarrasser en repensant, à sa manière, le rapport entre pensée et matière.

Spinoza choisit la voie la plus radicale : il supprime le problème en supprimant sa condition de possibilité, à savoir la pluralité des substances. Dans l'*Éthique*, il n'existe qu'une seule substance, infinie, que Spinoza nomme indifféremment Dieu ou la Nature. Cette substance unique possède une infinité d'**attributs**¹, dont deux seulement nous sont connus : la pensée et l'étendue². Autrement dit, ce que nous appelons « mon corps » et « mon esprit » ne correspond pas, chez Spinoza, à deux choses différentes qui devraient communiquer entre elles. Il s'agit d'une seule et même réalité, considérée de deux manières différentes : comme réalité étendue lorsqu'on la décrit sous l'attribut de l'étendue, et comme réalité mentale lorsqu'on la décrit sous l'attribut de la pensée. D'où le fameux **parallélisme** : l'ordre et la connexion des idées sont les mêmes que l'ordre et la connexion des choses. Il n'y a jamais causalité de l'esprit sur le corps ni du corps sur l'esprit, parce qu'il n'y a jamais deux substances séparées à faire communiquer — seulement une même réalité lue sous deux angles. [Spinoza (1677)]

Cette position préfigure à la fois certaines formes de **monisme neutre** et ce qu'on appellera plus tard les **théories du « double aspect »**, deux familles dont la frontière n'est d'ailleurs pas toujours nette. Le terme monisme signifie ici qu'il n'existe, au niveau fondamental, qu'une seule réalité ou substance. Dans le monisme neutre, cette réalité fondamentale n'est en elle-même ni spécifiquement matérielle ni spécifiquement mentale : le mental et le physique correspondent alors à deux manières dont cette réalité peut se manifester, s'organiser ou être décrite.

Malebranche, tout en restant à l'intérieur du cadre cartésien de deux substances distinctes, résout le problème en le déplaçant vers Dieu. Pour lui, aucune créature — ni le corps, ni l'esprit — n'est véritablement cause de quoi que ce soit : seul Dieu possède une vraie puissance causale. Quand ma volonté semble faire bouger mon bras, ou quand une brûlure semble produire une douleur, il s'agit en réalité d'« **occasions** » : mon état mental ou corporel est simplement l'occasion à laquelle Dieu produit, selon les lois qu'il a lui-même

¹ Chez Spinoza, un attribut est une manière fondamentale sous laquelle l'intellect saisit l'essence de la substance. La pensée et l'étendue ne sont donc pas deux parties de Dieu ou de la Nature, mais deux manières irréductibles sous lesquelles la même réalité peut être comprise.

² Dans ce contexte, l'étendue désigne le domaine corporel ou physique : ce qui occupe l'espace.

établies, l'effet correspondant dans l'autre substance. C'est la doctrine de l'**occasionalisme**. Elle permet d'éviter toute interaction directe entre les deux substances, au prix de faire de Dieu l'agent causal permanent et unique de tout ce qui arrive dans le monde. [Malebranche (1674-1675)]

Leibniz, enfin, refuse à la fois l'interactionnisme cartésien et l'occasionalisme de Malebranche, qu'il juge peu satisfaisant parce qu'il exige une intervention divine continue — ce qu'il appelle un « miracle perpétuel ». Sa métaphysique repose sur les **monades**, des substances simples, sans parties, et surtout « sans fenêtres » : rien n'entre ni ne sort d'une monade, chacune déploie ses états internes selon son propre principe interne, indépendamment de toutes les autres. Si l'esprit et le corps semblent malgré tout coordonnés — si je veux lever le bras et que le bras se lève —, ce n'est pas parce que l'un agit sur l'autre, mais parce que Dieu a réglé, une fois pour toutes au moment de la création, le déroulement de chaque monade de façon à ce qu'elles s'accordent parfaitement entre elles, comme deux horloges parfaitement synchronisées dès le départ et n'ayant plus jamais besoin d'être ajustées. C'est l'**harmonie préétablie** : un seul miracle initial remplace les interventions continues qu'exigeait l'occasionalisme. [Leibniz (1695, 1714)]

Leibniz formule par ailleurs une objection à ce qu'on pourrait appeler une explication **mécaniste** de l'esprit : l'idée selon laquelle la pensée et la perception pourraient être entièrement expliquées par l'organisation, les mouvements et les interactions des différentes parties d'un système matériel. Cette objection anticipe de manière frappante certains débats contemporains sur la conscience. Au paragraphe 17 de la *Monadologie*, il imagine une machine dont l'organisation serait capable de produire la pensée, la sensation et la perception. Supposons qu'on agrandisse cette machine en conservant ses proportions, jusqu'à pouvoir y entrer comme dans un moulin. En parcourant son intérieur, on n'y trouverait jamais que des pièces agissant les unes sur les autres, sans rien découvrir qui permette d'expliquer la perception elle-même. [Leibniz (1714)]

Cette conception mécaniste correspond d'ailleurs à une forme importante du matérialisme classique : on la trouve déjà chez Hobbes, puis plus tard, sous des formes différentes, chez La Mettrie et d'Holbach — même si Diderot s'en éloigne en défendant une conception plus dynamique de la matière. Nous retrouverons ces auteurs dans la section suivante.

Leibniz n'en conclut pas seulement que notre explication mécanique de la pensée et de la perception serait provisoirement incomplète. Il soutient que la perception est, par principe, inexplicable par les seules figures et les seuls mouvements, et qu'elle doit donc être cherchée dans une substance simple plutôt que dans un composé ou une machine. L'argument du moulin s'inscrit ainsi directement dans sa théorie des monades. Il préfigure néanmoins une intuition qui réapparaîtra sous une forme différente dans le problème difficile de la conscience : une description exhaustive des mécanismes et des interactions entre les parties semble encore laisser ouverte la question de savoir pourquoi, ou même comment, une expérience vécue en résulte.

III. La réponse matérialiste au problème cartésien

Il ne faut pas lire la séquence qui suit comme un débat direct avec Leibniz. Rien n'indique que La Mettrie ou d'Holbach aient construit leur matérialisme comme une réponse explicite à l'argument du moulin. Leur adversaire principal reste plutôt le dualisme cartésien. Ils rejettent bien la conclusion de Leibniz en soutenant que la pensée et la sensation peuvent appartenir à un organisme entièrement matériel, mais ils ne semblent pas affronter directement la difficulté précise qu'il soulève : comment l'organisation et les interactions des parties d'une machine peuvent-elles rendre compte de l'existence même d'une perception vécue ? Sur ce point, le problème reste largement en suspens.

La Mettrie ouvre la voie avec *L'Homme Machine* (1747), en radicalisant jusqu'à son terme l'animal-machine que Descartes réservait aux bêtes : l'homme lui-même n'est qu'une machine organisée, et la pensée n'est rien d'autre qu'une fonction de cette organisation matérielle, sans qu'il soit besoin de postuler une âme distincte ni un lieu d'interaction. [La Mettrie (1747)]

Diderot part d'une position plus prudente — proche, dans ses premiers écrits, d'un déisme classique — avant de basculer vers un matérialisme abouti, notamment dans le *Rêve de d'Alembert* : il y défend l'idée que la sensibilité n'est pas le privilège d'une âme immatérielle mais une propriété générale de la matière elle-même, tantôt inerte, tantôt active selon son organisation — une solution moins mécaniste que celle de La Mettrie, plus proche d'un vitalisme matérialiste. [Diderot (1769)]

D'Holbach, enfin, avec le *Système de la nature* (1770), donne à ce mouvement sa formulation la plus systématique : la nature y est décrite comme un tout matériel, régi par ses seules lois internes, sans besoin d'aucune cause spirituelle ni divine pour en rendre compte. Le problème de l'interaction âme-corps, qui obsédait les héritiers directs de Descartes, se trouve ainsi non pas résolu mais dissous : il n'y a plus, pour ces auteurs, qu'une seule substance à expliquer. [d'Holbach (1770)]

IV. Le dix-neuvième siècle : trois contraintes empiriques

Le XIXe siècle ne tranche pas directement le débat entre dualistes et matérialistes, mais il accumule des contraintes empiriques qui rendent la position dualiste de plus en plus difficile à tenir, et qui préparent le terrain sur lequel se formulera, au siècle suivant, l'idée d'une **clôture causale du physique**, qui sera expliquée ultérieurement.

La première contrainte vient du **darwinisme**. Si l'esprit humain s'inscrit dans un continuum graduel de complexité mentale, de l'organisme le plus simple jusqu'à l'homme, sans rupture nette dans la lignée évolutive, il devient difficile de justifier le moment précis où une âme immatérielle serait venue s'ajouter à la matière. Ce n'est pas un argument causal comme les deux suivants : le darwinisme attaque moins le mécanisme de l'interaction esprit-corps que le motif métaphysique qui justifierait, chez l'espèce humaine seule, l'irruption d'une substance d'un genre radicalement différent. [Darwin (1871)]

La deuxième contrainte vient du développement de la **mécanique classique** et des **lois de conservation**, formulées et systématisées au cours du XIX^e siècle notamment par Mayer, Joule et Helmholtz. La conservation de l'énergie impose que, dans un système physique isolé, la quantité totale d'énergie demeure constante. À première vue, une intervention d'une substance immatérielle dans le fonctionnement du corps semble donc problématique : si l'esprit produisait un mouvement en ajoutant une cause extérieure à la dynamique physique, on pourrait s'attendre à constater une modification inexplicée du bilan énergétique.

La conservation de la quantité de mouvement ajoute une difficulté supplémentaire. Supposer que l'esprit ne crée aucune énergie, mais se contente d'en modifier la distribution ou d'orienter autrement un mouvement, ne suffit pas nécessairement à sauver l'interactionnisme : une telle déviation pourrait elle-même modifier l'impulsion totale du système. À ces deux contraintes s'ajoute l'idéal **déterministe** de la mécanique classique, exprimé de manière célèbre par Laplace : si l'état physique complet du monde à un instant et les lois qui le gouvernent déterminent entièrement son évolution future, il semble ne rester aucun événement physique dont une volonté immatérielle pourrait modifier l'issue sans introduire une rupture dans cette évolution.

Il faut toutefois éviter de confondre ces différentes contraintes avec le principe plus fort de **clôture causale du physique**, selon lequel **tout effet physique possède une cause physique suffisante**. Les lois de conservation et le déterminisme classique ont certes été conçus comme des principes très généraux du domaine physique, mais ils ne disent pas, par leur seul contenu, que toute cause d'un événement physique doit elle-même être physique. Un dualiste interactionniste peut précisément contester que les processus cérébraux constituent un système régi de manière causalement complète par ces seules lois lorsqu'une cause mentale intervient. Il peut soutenir qu'à certains moments de l'interaction entre esprit et cerveau, l'évolution physique ne se laisse pas entièrement déterminer par l'état physique antérieur, voire que les relations de conservation habituellement observées ne s'appliquent pas sans exception.

La physique classique n'a donc pas réfuté logiquement le dualisme interactionniste. Elle en a cependant considérablement alourdi la charge explicative. Plus les processus physiques et physiologiques sont décrits avec précision sans que soit observé un point où leur évolution exige l'intervention d'une cause non physique, plus le dualiste doit préciser où cette intervention se produit, comment elle modifie la dynamique cérébrale et pourquoi ses effets n'ont pas encore été mis en évidence. La possibilité logique d'une exception demeure, mais elle devient progressivement une hypothèse empirique particulièrement difficile à établir plutôt qu'une idée qui découle naturellement de nos meilleures explications scientifiques sur le fonctionnement du monde.

Le principe de clôture causale généralise précisément cette évolution : il affirme qu'il n'existe aucun tel point d'ouverture, autrement dit que les causes physiques suffisent à compléter l'explication de tout événement physique. C'est pourquoi son adoption constitue

une contrainte beaucoup plus directe pour le dualisme interactionniste que ne le sont, prises isolément, les lois de conservation ou le déterminisme classique.

Cette dernière contrainte perdra toutefois une partie de sa force avec l'avènement de la mécanique quantique, dont certaines interprétations introduisent des résultats seulement probabilistes. Cette ouverture ne suffira pas à résoudre le problème de l'interaction : remplacer une évolution déterminée par plusieurs issues possibles ne permet pas encore d'expliquer comment une volonté non physique sélectionnerait l'une d'entre elles. Nous y reviendrons.

La troisième contrainte vient de la physiologie expérimentale, puis de la neurologie naissante. Vers 1850, Hermann von Helmholtz parvient à mesurer la vitesse de propagation de l'influx nerveux. Le résultat ne réfute pas directement le dualisme : rien n'interdit encore à son défenseur de supposer qu'une interaction avec l'esprit intervient à un point particulier de la chaîne nerveuse. Mais il contribue à un déplacement décisif. Des phénomènes autrefois associés à l'action immédiate d'un principe vital ou mental — la transmission d'une sensation, la préparation d'un mouvement, le délai entre stimulation et réponse — deviennent des processus physiques quantifiables, inscrits dans une temporalité physiologique mesurable.

Les travaux de Paul Broca, à partir de 1861, puis de Carl Wernicke, dans les années 1870, renforcent cette contrainte en établissant des relations systématiques entre certaines lésions cérébrales et des déficits linguistiques déterminés. Une lésion de régions frontales de l'hémisphère gauche peut ainsi compromettre gravement la production articulée du langage, tandis que d'autres lésions affectent davantage sa compréhension. Ces résultats ne démontrent ni que chaque faculté mentale possède un siège unique et strictement délimité, ni que l'esprit se réduit au cerveau. Ils montrent cependant que des capacités mentales précises dépendent de l'intégrité et de l'organisation de structures cérébrales déterminées. Le dualiste ne peut donc plus se contenter d'affirmer une dépendance générale de l'esprit à l'égard du corps : il doit expliquer pourquoi l'altération locale d'un tissu physique entraîne des déficits mentaux relativement spécifiques.

C'est le physiologiste et philosophe Thomas Huxley qui, en 1874, tire l'une des conséquences les plus radicales de cette **naturalisation** progressive du système nerveux. Dans « On the Hypothesis that Animals Are Automata, and Its History », il défend l'hypothèse selon laquelle les animaux, êtres humains compris, sont des **automates conscients** : leurs mouvements et leurs comportements peuvent être entièrement produits par les mécanismes nerveux, tandis que leurs expériences conscientes accompagnent ces processus sans exercer sur eux d'influence causale propre. Huxley compare ainsi la conscience au sifflet à vapeur d'une locomotive, qui signale le fonctionnement de la machine sans contribuer à son mouvement. [Huxley (1874)]

Cette position sera plus tard désignée sous le nom d'**épiphénoménalisme** : les événements cérébraux causent les expériences conscientes, mais celles-ci ne produisent aucun effet physique en retour. Il ne faut cependant pas en conclure que Huxley réintroduit nécessairement une âme ou une substance mentale séparée. Sa thèse porte avant tout sur

l'inefficacité causale de la conscience, sans fournir une théorie pleinement développée de son statut ontologique³. Elle constitue néanmoins une position charnière : Huxley accepte que la conscience existe réellement et qu'elle soit liée au fonctionnement du cerveau, tout en abandonnant l'idée qu'elle joue un rôle dans la production du comportement.

Un dernier jalon mérite d'être signalé, parce qu'il touche directement à ce qui deviendra le problème difficile de la conscience. En 1872, le physiologiste Emil du Bois-Reymond prononce un discours intitulé « Sur les limites de la connaissance de la nature ». Il y soutient que, même si la physiologie parvenait un jour à décrire intégralement les processus physiques du cerveau, elle ne pourrait pas pour autant expliquer comment une expérience consciente peut surgir de mouvements de matière. La difficulté ne tient donc pas seulement à l'insuffisance provisoire de nos connaissances : selon lui, aucun agencement de particules ni aucune description mécanique ne permet de rendre intelligible le passage entre l'activité nerveuse et la sensation vécue. [du Bois-Reymond (1872)]

Du Bois-Reymond résume cette limite par la formule devenue célèbre : Ignorabimus — « nous ne saurons pas ». Sa position est cependant plus prudente qu'un simple dualisme. Physiologiste mécaniste et adversaire du vitalisme, il considère que les phénomènes mentaux possèdent des conditions matérielles et doivent être étudiés scientifiquement. Mais il refuse d'en conclure que leur nature est entièrement réductible à celle de la matière, dont l'essence ultime lui paraît elle-même inconnue. Il défend ainsi moins une ontologie précise qu'un scepticisme métaphysique : les sciences naturelles peuvent découvrir les corrélations et les mécanismes physiques associés à la conscience sans nécessairement rendre intelligible l'existence même de l'expérience subjective. Une formulation très proche du futur fossé explicatif apparaît donc dès 1872, plus d'un siècle avant que Chalmers ne lui donne son nom actuel.

Cette attitude préfigure ce que l'on appellera beaucoup plus tard le « **mystérianisme** » : l'idée que certaines questions, et notamment celle du rapport entre cerveau et expérience consciente, pourraient être en principe hors de portée des capacités cognitives humaines. Les défenseurs contemporains de cette position ne soutiennent pas nécessairement que le problème est insoluble en lui-même, mais plutôt que notre esprit pourrait ne pas disposer des ressources conceptuelles nécessaires pour le résoudre. [McGinn (1989)]

Cette difficulté nourrit également une voie intermédiaire entre le dualisme traditionnel et le matérialisme strictement réductionniste : l'**émergentisme** britannique. Dès 1843, John Stuart Mill distingue les effets simplement composés, que l'on peut comprendre en additionnant les effets de leurs causes prises séparément, des effets qu'il appelle « hétéropathiques », qui font apparaître des propriétés ou des régularités nouvelles. Par exemple, les propriétés de l'eau ne ressemblent pas à une simple addition de celles de l'hydrogène et de l'oxygène qui la composent : leur combinaison produit un système possédant des caractéristiques nouvelles. George Henry Lewes introduit ensuite le terme d'**émergence** pour désigner ce type de nouveauté, avant que Samuel Alexander, C. Lloyd

³ L'ontologie correspond à ce qui existe, à ce qui est réel. Ici, Huxley ne discute pas de ce qu'est réellement la conscience.

Morgan et C. D. Broad ne systématisent cette position au début du XX^e siècle. [Mill (1843) ; Lewes (1875) ; Alexander (1920) ; Morgan (1923) ; Broad (1925)]

Appliqué à l'esprit, l'émergentisme affirme que la conscience n'est pas une substance ajoutée de l'extérieur à la matière : elle apparaît lorsque l'organisation physique atteint un certain degré de **complexité**. Mais les propriétés conscientes ne seraient pas pour autant déductibles des seules lois gouvernant les constituants physiques élémentaires. Cette stratégie permet de préserver la dépendance de l'esprit à l'égard du cerveau sans accepter une réduction complète du premier au second. Elle anticipe ainsi une famille de positions qui réapparaîtra régulièrement dans les débats contemporains : la conscience serait entièrement naturelle et produite par le physique, tout en constituant un niveau de réalité doté de propriétés ou de lois propres. ⁴

V. Du vitalisme aux neurosciences modernes : la naturalisation du vivant et de l'esprit

Cette section suit volontairement un fil thématique plutôt qu'une chronologie stricte. Du tournant du XX^e siècle jusqu'à sa seconde moitié, plusieurs programmes expérimentaux convergent pour rendre le vivant, le système nerveux et une part croissante des fonctions mentales accessibles à des explications physico-chimiques et physiologiques — non sans un dernier sursaut **vitaliste** ⁵ au passage. Cette progression ne doit cependant pas être confondue avec une conversion générale des chercheurs au matérialisme : plusieurs d'entre eux contribueront à étendre le domaine de l'explication naturaliste tout en conservant des positions métaphysiques dualistes ou non réductionnistes.

La première avancée concerne la structure même du système nerveux. Camillo Golgi met au point en 1873 une technique de coloration qui permet de visualiser individuellement les cellules nerveuses, mais il interprète ses propres résultats en défendant une théorie réticulaire, selon laquelle le système nerveux formerait un réseau continu. Santiago Ramón y Cajal, utilisant la même technique, établit au contraire la doctrine du neurone : chaque cellule nerveuse constitue une unité anatomique distincte, séparée des autres et communiquant avec elles sans continuité physique directe. Golgi et Cajal recevront pourtant ensemble le prix Nobel en 1906, chacun continuant alors de défendre son interprétation.

Cette découverte prépare le travail de Charles Sherrington, qui introduit à la fin du XIX^e siècle le terme de « synapse » pour désigner la zone fonctionnelle de communication entre deux neurones. Dans *The Integrative Action of the Nervous System* (1906), il montre comment des réflexes élémentaires peuvent être coordonnés et intégrés par l'organisation du système nerveux : l'explication physiologique de cette coordination peut être recherchée dans les propriétés des circuits et des interactions synaptiques eux-mêmes. [Sherrington

⁴ Voir également mon article [Vila Valls (2026)] « Émergence, réduction et real patterns », *The Naturalistic Outlook*, <https://thenaturalisticoutlook.com/2026/07/19/emergence-reduction-et-real-patterns/>

⁵ Le vitalisme désigne l'idée que les phénomènes propres au vivant ne peuvent pas être entièrement expliqués par les seules lois physico-chimiques et qu'ils exigent l'intervention d'un principe spécifique au vivant, parfois conçu comme une force ou un principe organisateur. Sous des formes diverses, cette manière de penser a été très répandue avant que la physiologie, la biochimie et la biologie ne parviennent progressivement à expliquer un nombre croissant de phénomènes vitaux sans recourir à un tel principe.

(1906)] Il ne faut toutefois pas en déduire que Sherrington était matérialiste. Dans ses réflexions ultérieures sur le rapport entre l'esprit et le cerveau, notamment dans *Man on His Nature* (1940), il demeure réticent à identifier purement et simplement le mental au nerveux et conserve une conception dualiste ou, au minimum, fortement non réductionniste. [Sherrington (1940)] Son cas illustre donc une distinction essentielle pour la suite : un programme scientifique peut naturaliser une fonction sans que son auteur en conclue que toute réalité mentale se réduit au physique.

Le mouvement de naturalisation n'est pas linéaire. Au même moment, l'embryologiste Hans Driesch relance le vitalisme sur des bases scientifiques sérieuses. Ses expériences sur les oursins, dès 1891, montrent qu'en séparant les deux premières cellules d'un embryon, chacune peut donner un organisme entier et normal — un résultat que les modèles mécanistes disponibles expliquent alors difficilement. Driesch en conclut, dans ses Gifford Lectures de 1907-1908, à l'existence d'une « entéléchie » : un principe organisateur non physique propre au vivant, censé guider le développement de l'organisme sans se réduire aux mécanismes physico-chimiques qui le composent. [Driesch (1908)]

C'est sur le terrain de la chimie et de l'expérimentation que ce vitalisme renaissant rencontre ses adversaires les plus directs. Eduard Buchner montre dès 1897 que la fermentation alcoolique se produit avec un simple extrait acellulaire de levure, sans qu'aucune cellule vivante intacte soit nécessaire. Jacques Loeb prolonge ce programme anti-vitaliste avec ses expériences de parthénogenèse artificielle chez l'oursin, à partir de 1899 : par des manipulations physico-chimiques, il déclenche le développement d'œufs non fécondés jusqu'à la formation de larves normales. Dans *The Mechanistic Conception of Life* (1912), il défend explicitement l'objectif d'une explication physico-chimique des phénomènes vitaux. [Buchner (1897) ; Loeb (1899, 1912)] Sa démarche est également associée à ce que l'historien Philip Pauly a appelé un « *engineering standpoint* » : comprendre le vivant, c'est aussi acquérir sur ses mécanismes un degré de contrôle expérimental comparable à celui que l'ingénieur recherche dans les systèmes artificiels. [Pauly (1987)]

Ce programme prend une dimension nouvelle avec l'essor de la biologie moléculaire. En 1944, Oswald Avery, Colin MacLeod et Maclyn McCarty montrent que le « principe transformant » capable de transmettre certains caractères héréditaires chez le pneumocoque est constitué d'ADN. L'hérédité commence ainsi à être reliée à une substance chimique déterminée plutôt qu'à un principe biologique irréductible. [Avery, MacLeod & McCarty (1944)]

En 1953, James Watson et Francis Crick proposent la structure en double hélice de l'ADN, en s'appuyant notamment sur les données de diffraction des rayons X obtenues à King's College par Rosalind Franklin et Raymond Gosling, ainsi que par Maurice Wilkins et ses collaborateurs. La complémentarité des bases fournit immédiatement l'esquisse d'un mécanisme de copie : si les deux brins se séparent, chacun peut servir de modèle à la reconstruction de son complément. Un phénomène central du vivant — la transmission de l'information héréditaire — devient ainsi intelligible à partir de la structure et des

interactions d'une molécule. [Watson & Crick (1953) ; Franklin & Gosling (1953) ; Wilkins, Stokes & Wilson (1953)]

La continuité avec le programme de Loeb est surtout intellectuelle plutôt que généalogique : la biologie moléculaire réalise de manière spectaculaire l'ambition de rendre les phénomènes vitaux manipulables et explicables par leurs mécanismes. Jacques Monod incarne particulièrement bien ce tournant. Avec François Jacob, il contribue au début des années 1960 à élucider des mécanismes moléculaires de régulation de l'expression des gènes. [Jacob & Monod (1961)] Puis, dans *Le Hasard et la Nécessité* (1970), il donne à cette biologie une interprétation philosophique explicitement naturaliste : l'apparente finalité des organismes — leur « téléonomie » — doit être comprise à partir de mécanismes moléculaires, de l'hérédité et de la sélection naturelle, non comme la manifestation d'un projet extérieur ou d'un principe vital. [Monod (1970)]

Une transformation parallèle se produit dans les sciences du cerveau. Les travaux d'Edgar Adrian, rendus possibles par l'amélioration des techniques d'amplification électrique, permettent dans les années 1920 de suivre l'activité de fibres nerveuses individuelles. Ils montrent notamment qu'un neurone produit des impulsions électriques élémentaires d'intensité assez semblable : un stimulus plus intense ne produit donc pas simplement une impulsion beaucoup plus forte. Il peut être représenté par un plus grand nombre d'impulsions émises par seconde, ainsi que par l'activation d'un plus grand nombre de fibres nerveuses. On peut, très grossièrement, comparer cela au code Morse : ce n'est pas l'intensité d'un « bip » isolé qui porte l'information, mais la manière dont les signaux sont distribués dans le temps. Dans le système nerveux, l'information ne dépend donc pas simplement de la force de chaque impulsion, mais notamment de leur fréquence et du nombre de fibres qui participent à la réponse. Le système nerveux commence ainsi à être décrit non seulement comme un réseau anatomique, mais comme un dispositif capable de transmettre et de coder physiquement de l'information. [Adrian (1932)]

En 1952, Alan Hodgkin et Andrew Huxley franchissent une étape supplémentaire en expliquant le mécanisme physique de ces impulsions nerveuses, appelées potentiels d'action. La membrane qui entoure le neurone sépare des milieux contenant différentes concentrations d'ions — notamment du sodium et du potassium, c'est-à-dire des particules électriquement chargées. Lorsqu'un neurone s'active, certains passages de sa membrane s'ouvrent et se ferment successivement, permettant à ces ions de circuler. Ces déplacements de charges électriques produisent la variation rapide de tension qui constitue le potentiel d'action et permettent sa propagation le long du neurone. L'influx dont Helmholtz avait mesuré la vitesse un siècle plus tôt reçoit ainsi une description électrochimique précise. [Hodgkin & Huxley (1952)]

Avec David Hubel et Torsten Wiesel, l'explication physiologique atteint ensuite le traitement de l'information perceptive lui-même. Leurs travaux sur le cortex visuel montrent que certains neurones répondent sélectivement à des caractéristiques déterminées des stimuli, notamment l'orientation et la position de contours, et que l'organisation des circuits du cortex cérébral participe à l'analyse progressive de

l'information visuelle. [Hubel & Wiesel (1962)] Il ne s'agit plus seulement d'expliquer comment un signal nerveux se propage, mais comment l'activité de populations neuronales peut réaliser des opérations perceptives spécifiques.

Les recherches de Roger Sperry sur les patients dits « au cerveau divisé » renforcent encore cette dépendance fonctionnelle du mental à l'organisation cérébrale. Chez certains patients atteints d'épilepsie sévère, les chirurgiens avaient sectionné le corps calleux, le grand faisceau de fibres nerveuses qui relie les deux hémisphères cérébraux, parfois avec d'autres connexions entre ceux-ci. Les deux moitiés du cerveau pouvaient dès lors beaucoup moins facilement échanger certaines informations. Des expériences montrent alors qu'une information reçue principalement par un hémisphère peut devenir inaccessible à l'autre, et que des capacités cognitives jusque-là unifiées peuvent se dissocier de manière systématique. [Sperry (1968)] Une modification matérielle des connexions cérébrales ne produit donc pas seulement un déficit global : elle transforme de manière structurée l'organisation des fonctions mentales accessibles au sujet.

Au milieu du XX^e siècle, deux programmes initialement distincts convergent ainsi. La biologie moléculaire rend progressivement intelligibles l'hérédité, la régulation et une part croissante du fonctionnement cellulaire à partir de mécanismes physico-chimiques ; les neurosciences décomposent parallèlement la transmission, le codage et le traitement de l'information nerveuse en mécanismes électriques, chimiques et anatomiques. Rien de cela ne démontre à lui seul que la conscience soit réductible au physique : Sherrington, puis Eccles, montrent précisément qu'on peut accepter une grande partie de cette naturalisation scientifique tout en restant dualiste. Mais l'espace laissé à une intervention mentale indépendante se resserre à mesure que des fonctions autrefois attribuées globalement à l'« esprit » deviennent accessibles à une explication physiologique détaillée.

Cette continuité de programme apparaîtra encore de façon frappante chez Francis Crick : l'un des artisans de la révolution moléculaire de l'ADN consacrera, quarante ans plus tard, *The Astonishing Hypothesis* (1994) à l'idée que la perception, la pensée et le sentiment d'identité doivent eux aussi être abordés comme des produits de l'activité cérébrale. [Crick (1994)] Le succès de cette extension reste précisément ce que les débats contemporains sur la conscience devront déterminer.

VI. L'interlude quantique : une porte entrouverte, puis refermée

La section précédente a suivi jusqu'à la seconde moitié du XX^e siècle la progression propre de la biologie et des neurosciences. Il faut maintenant revenir quelques décennies en arrière, dans les années 1920, car un bouleversement de la physique allait momentanément sembler rouvrir une porte que cette naturalisation progressive paraissait refermer.

L'avènement de la **mécanique quantique** dans les années 1920 remet en cause le cadre déterministe de la physique classique, du moins dans les interprétations qui deviennent rapidement dominantes. La règle de Born, qui permet d'extraire les prédictions expérimentales de la théorie à partir de son formalisme mathématique, n'assigne aux

différents résultats possibles que des probabilités et, dans la théorie orthodoxe de la mesure quantique, aucun élément de l'état quantique ne détermine lequel se réalisera lors d'une observation particulière. Même une connaissance complète de l'état physique ne permettrait donc plus de prévoir avec certitude chaque événement individuel : le déterminisme laplacien paraît abandonné. [Born (1926)]

Cette conclusion dépend toutefois de l'interprétation retenue. L'équation de Schrödinger, qui est l'équation centrale décrivant l'évolution de l'état quantique, est elle-même continue et déterministe. Dans la formulation orthodoxe, on suppose cependant qu'une mesure introduit une rupture dans cette évolution : parmi les différents résultats possibles, un seul est alors obtenu, avec des probabilités données par la règle de Born. Des interprétations formulées après la seconde guerre mondiale, comme la mécanique bohémienne ou l'interprétation d'Everett, éviteront chacun à leur manière cette rupture dans l'évolution déterministe de l'état quantique et offriront ainsi deux manières alternatives de concevoir les processus quantiques sans indéterminisme fondamental. Mais cette pluralité d'interprétations ne se dégagera clairement que plus tard. Pour plusieurs physiciens de l'entre-deux-guerres, la mécanique quantique semble alors rouvrir une porte que la mécanique classique, ses lois de conservation et son déterminisme avaient progressivement refermée sur le dualisme interactionniste.

Arthur Eddington est l'un des premiers à tirer cette conséquence, dans *The Nature of the Physical World* (1928). Sa position reste prudente : il ne propose pas de mécanisme précis par lequel l'esprit agirait sur le cerveau, mais soutient que **l'indétermination quantique** lève un obstacle métaphysique majeur à la **liberté humaine**. Le monde n'étant plus un bloc rigidement déterminé de part en part, rien n'empêche plus, selon lui, de penser que la décision volontaire trouve une place légitime dans les interstices laissés ouverts par la physique elle-même. [Eddington (1928)]

Arthur Compton va plus loin dans ses conférences Terry, publiées sous le titre *The Freedom of Man* (1935). Il défend l'idée que l'esprit pourrait directement influencer l'issue d'événements quantiques indéterminés au sein du cerveau, et que cette issue, initialement microscopique et indéterminée, se trouverait ensuite amplifiée en un effet macroscopique déterminé --- un peu comme un signal électrique infime peut déclencher, via un mécanisme d'amplification, un effet de grande ampleur. Pour Compton, une telle sélection permettrait à la volonté d'agir sur le corps sans injecter nécessairement d'énergie nouvelle dans le système physique. Cette affirmation reste toutefois programmatique : elle ne fournit pas encore de mécanisme montrant comment cette sélection pourrait s'opérer tout en respectant l'ensemble des contraintes de la dynamique physique. [Compton (1935)]

Cette stratégie sera reprise plusieurs décennies plus tard par John Eccles, qui cherchera à localiser dans la dynamique microscopique du cerveau un point d'intervention possible pour une cause mentale non physique. Nous reviendrons sur cette tentative à propos de la clôture causale.

Cette tentative ne convainc cependant pas la majorité des physiciens ni des philosophes de l'esprit. L'objection la plus immédiate est que l'indétermination ne constitue pas encore une

action de l'esprit : remplacer une évolution strictement déterminée par plusieurs issues probabilistes ne donne à une volonté non physique aucun contrôle sur celle qui se réalise. Il reste donc à expliquer comment l'esprit pourrait sélectionner une issue plutôt qu'une autre, et comment cette sélection s'insérerait dans la dynamique quantique sans devenir elle-même une cause supplémentaire laissée hors de l'explication physique.

L'indétermination quantique affaiblit ainsi l'une des contraintes héritées de la mécanique classique — le déterminisme laplacien — mais elle ne réfute ni les lois de conservation ni le principe plus général de clôture causale. L'interlude quantique ne résout donc pas le problème de l'interaction : il déplace le point auquel un mécanisme causal devrait être fourni, du niveau macroscopique vers le niveau quantique.

VII. Le physicalisme s'institue : de la chimie quantique à la clôture causale

Cette section introduit le terme de **physicalisme**, qu'il convient de distinguer de celui de matérialisme employé jusqu'ici. Le matérialisme classique, tel que le défendaient La Mettrie, Diderot ou d'Holbach, affirme qu'il n'existe qu'une seule sorte de substance : la matière, généralement conçue à partir du modèle de la physique de leur époque, comme étendue, localisée et soumise à des lois mécaniques.

Le terme de physicalisme apparaît au XX^e siècle, notamment dans le contexte du Cercle de Vienne. Il s'agit d'un groupe de philosophes et de scientifiques des années 1920-1930 qui cherchent à construire une philosophie aussi rigoureusement articulée que possible aux sciences de leur époque. Ils sont notamment marqués par les révolutions de la relativité et de la mécanique quantique, ainsi que par les progrès de la logique et des fondements des mathématiques. Leur ambition est de rapprocher philosophie et science, de clarifier le langage scientifique et de limiter le recours aux spéculations métaphysiques qui ne peuvent être soumises à une analyse logique ou empirique.

Dans ce contexte, le mot « physicalisme » ne désigne d'ailleurs pas encore exactement ce qu'il désignera dans une grande partie de la philosophie contemporaine. Chez Otto Neurath notamment, il s'agit d'abord d'une thèse sur le langage et l'unité des sciences plutôt que d'une doctrine métaphysique sur la constitution ultime du monde. Les différentes sciences — physique, chimie, biologie, psychologie ou sociologie — doivent pouvoir communiquer au sein d'un langage commun, intersubjectivement contrôlable, décrivant des objets, événements et processus situés dans l'espace et dans le temps. La physique fournit le modèle et le vocabulaire de base de cette entreprise d'unification, sans que cela oblige encore à affirmer métaphysiquement que seules des entités physiques existent. [Neurath (1931) ; Carnap (1932)]

Le sens du terme se déplacera ensuite. Dans une grande partie de la philosophie contemporaine, le physicalisme désigne désormais principalement une thèse ontologique : tout ce qui existe est physique ou dépend entièrement du physique. Ce terme permet alors d'éviter certaines connotations historiques du mot « matérialisme ». La physique moderne reconnaît en effet comme réels des champs, des structures spatio-temporelles et d'autres

entités qui ne correspondent plus à l'image classique de petits corps solides en mouvement. Le physicalisme contemporain ne consiste donc pas à soutenir que tout est matériel au sens du XVIII^e siècle, mais que tout ce qui existe appartient au domaine décrit par une physique complète, ou dépend entièrement de ce domaine.

Cette reformulation rend la thèse plus souple face à l'évolution des théories physiques, mais elle soulève une difficulté : que faut-il exactement entendre par « physique complète » ? Si l'on entend par là la physique actuelle, le physicalisme risque d'être faux parce que nos théories sont probablement incomplètes ; si l'on vise une physique future idéale, la thèse devient plus difficile à définir sans circularité. C'est le dilemme de Hempel, que je développe plus longuement dans un autre article. [Hempel (1969), Vila Valls (2026)]

Dans l'usage contemporain, les termes de matérialisme et de physicalisme restent néanmoins souvent employés comme des quasi-synonymes, selon le contexte historique.

La formulation explicite du physicalisme au XX^e siècle ne naît pas seulement d'une réflexion philosophique sur l'esprit. Elle s'appuie également sur les succès spectaculaires de la physique lorsqu'elle est appliquée à des domaines jusque-là considérés comme relativement autonomes, au premier rang desquels la chimie.

En 1927, Walter Heitler et Fritz London proposent le premier traitement quantique de la **liaison covalente** dans la molécule d'hydrogène ⁶. Leur modèle montre comment la stabilité d'une molécule peut être comprise à partir de la mécanique quantique et de l'interaction entre les électrons et les noyaux. Il constitue ainsi l'un des actes fondateurs de la **chimie quantique**, que Linus Pauling et d'autres développeront et systématiseront au cours des années suivantes. Pour la première fois, certaines propriétés chimiques fondamentales apparaissent comme explicables à partir des lois de la physique nouvelle, et non plus seulement comme des régularités empiriques propres à la chimie. [Heitler & London (1927)]

La portée exacte de cette réussite doit cependant être précisée. Heitler et London n'ont pas déduit toute la chimie de la seule équation de Schrödinger, et la chimie quantique ne fournit pas davantage aujourd'hui une réduction complète de l'ensemble des concepts et des lois chimiques. Ses calculs reposent sur des approximations, sur des modèles et parfois sur des informations structurelles qui ne sont pas elles-mêmes directement dérivées des seules lois fondamentales. Des notions telles que la liaison, la valence ou la forme moléculaire conservent par ailleurs un rôle explicatif propre dans la pratique des chimistes.

Ces limites n'empêchent pas les succès de la chimie quantique de produire, à l'époque, une impression profonde d'**unification**. En 1929, Paul Dirac peut ainsi écrire que les lois

⁶ La molécule d'hydrogène, notée H₂, est constituée de deux atomes d'hydrogène, chacun composé d'un proton et d'un électron. Dans une liaison covalente, les deux atomes partagent une paire d'électrons. Cette idée de partage des électrons était déjà utilisée par les chimistes avant la mécanique quantique, mais on ne disposait pas encore d'une explication physique satisfaisante de la raison pour laquelle cette configuration produit une liaison stable. Le traitement quantique de Heitler et London montre que les deux électrons peuvent former un état commun dans lequel l'énergie totale du système est plus basse lorsque les deux noyaux se trouvent à une certaine distance l'un de l'autre que lorsqu'ils sont très éloignés. La molécule liée correspond donc à un état énergétiquement plus stable que les deux atomes séparés. La mécanique quantique fournit ainsi, pour la première fois, un mécanisme physique permettant de comprendre la stabilité d'une liaison covalente.

physiques fondamentales nécessaires à la théorie mathématique d'une grande partie de la physique et de toute la chimie sont désormais connues, la principale difficulté tenant à la complexité insoluble des équations auxquelles elles conduisent. Cette déclaration doit être comprise comme l'expression particulièrement ambitieuse d'un programme : elle affirme que les phénomènes chimiques possèdent une base physique commune, non que toutes les explications chimiques aient déjà été effectivement remplacées par des dérivations quantiques. [Dirac (1929)]

C'est dans ce climat que le Cercle de Vienne — Carnap, Neurath et Hempel notamment — développe, dans les années 1930, son **programme d'unité de la science** : les différentes sciences empiriques devraient pouvoir être articulées dans un langage physicaliste commun. Les avancées de la chimie quantique fournissent à ce programme un modèle de réussite particulièrement impressionnant. Elles rendent plausible l'idée que des phénomènes décrits à un niveau supérieur peuvent dépendre entièrement de mécanismes physiques, sans pour autant démontrer que les concepts et les explications de ce niveau puissent être intégralement éliminés ou déduits des seules lois fondamentales.⁷

C'est dans ce contexte que la **théorie de l'identité esprit-cerveau** prend sa forme moderne, notamment avec Herbert Feigl dans « The “Mental” and the “Physical” » (1958), puis avec U. T. Place et J. J. C. Smart. Ces auteurs s'appuient sur une idée devenue de plus en plus plausible au fil du développement des sciences physiques et biologiques : **le déroulement des processus physiques ne semble exiger l'intervention d'aucune cause non physique supplémentaire. Autrement dit, le domaine physique paraît causalement complet.** [Place (1956) ; Feigl (1958) ; Smart (1959)]

Cette thèse, que l'on désignera plus tard comme **le principe de clôture causale du physique**, affirme que tout effet physique possède une cause physique suffisante. Feigl compte parmi les auteurs qui l'intègrent explicitement au problème de l'esprit et du corps, mais il ne faut pas lui en attribuer une invention isolée : le principe résulte plutôt de la convergence progressive des lois de conservation, de la physiologie mécaniste, de la neurologie et de l'unification croissante des sciences.

La clôture causale fournit alors un argument puissant en faveur de la théorie de l'identité esprit-cerveau. Si les états mentaux produisent réellement des comportements physiques, alors que chaque événement physique possède déjà une cause physique suffisante, deux possibilités semblent rester ouvertes : soit les effets corporels sont systématiquement produits par deux chaînes causales distinctes, l'une mentale et l'autre physique ; soit les états mentaux sont eux-mêmes identiques aux états physiques qui jouent ce rôle causal. Place, Feigl et Smart considèrent la seconde hypothèse comme la plus économique : la douleur, la perception ou la pensée ne sont pas des événements supplémentaires accompagnant l'activité cérébrale, mais cette activité elle-même, décrite sous un autre vocabulaire.

⁷ Sur cette distinction entre dépendance physique, réduction et autonomie explicative des niveaux supérieurs, voir également : « Émergence, réduction et real patterns », *The Naturalistic Outlook*, <https://thenaturalisticoutlook.com/2026/07/19/emergence-reduction-et-real-patterns/>

La même année, Paul Oppenheim et Hilary Putnam donnent à ce mouvement une formulation systématique dans « Unity of Science as a Working Hypothesis » (1958). Ils proposent une hiérarchie de niveaux — des particules physiques aux atomes, aux molécules, aux cellules, aux organismes vivants, puis aux groupes sociaux — et envisagent que chaque niveau puisse, en principe, être relié par des relations de micro-réduction au niveau inférieur. Pour être tout à fait explicite, l'idée est que les phénomènes propres à chaque niveau pourraient, en principe, être expliqués à partir des entités et des processus du niveau immédiatement inférieur : les propriétés des atomes à partir de celles des particules qui les composent ; celles des molécules à partir de la physique des atomes ; les phénomènes cellulaires à partir de mécanismes moléculaires ; ceux des organismes à partir du fonctionnement de leurs cellules ; les phénomènes psychologiques à partir notamment de l'activité du système nerveux ; et, finalement, les phénomènes sociaux à partir des comportements et interactions des individus. Les succès de la physique atomique et de la chimie quantique leur servent de modèle et renforcent la plausibilité de ce programme. Comme l'indique cependant le titre de leur article, il s'agit explicitement d'une hypothèse de travail, et non du constat que toutes les réductions de la hiérarchie ont déjà été accomplies. [Oppenheim & Putnam (1958)]

VIII. Intermède philosophique : le statut du principe de clôture causale

Avant de poursuivre le récit historique, il vaut la peine de s'arrêter sur la nature exacte du principe de clôture causale du physique, parce que son statut épistémique est singulier et qu'il structure une grande partie des débats contemporains sur l'esprit.⁸

Le premier fait notable est que ce principe est aujourd'hui largement accepté dans la philosophie de l'esprit naturaliste, y compris par plusieurs auteurs qui rejettent pourtant le physicalisme réductionniste. Il ne fait cependant pas l'objet d'un consensus absolu : les dualistes interactionnistes et certains émergentistes continuent de le contester. Son influence vient surtout de la pression qu'il exerce sur toute théorie attribuant au mental une efficacité causale propre. Cette pression peut être rendue transparente par la présentation **du problème de l'exclusion causale**, formulée dans sa version canonique par le philosophe Jaegwon Kim en 1998.

Celui-ci peut être présenté comme une tension entre plusieurs thèses qui, prises séparément, paraissent plausibles. La première thèse est justement **le principe de clôture causale**, qui affirme que tout événement physique possède une cause physique suffisante. La deuxième thèse est **l'efficacité causale du mental** : nous attribuons normalement une efficacité causale à nos états mentaux : décider de lever le bras semble contribuer réellement au mouvement qui s'ensuit. La troisième est **le principe d'exclusion causal** : sauf cas particulier de véritable surdétermination, un même effet ne devrait pas posséder simultanément deux causes suffisantes entièrement distinctes. On

⁸ Épistémique signifie ici « relatif à la connaissance et à sa justification ». Parler du statut épistémique d'une proposition revient donc à demander sur quelles raisons nous pouvons la tenir pour vraie et avec quel degré de certitude : démonstration logique, résultat expérimental, généralisation inductive, hypothèse de travail, etc.

attribue par exemple généralement une cause unique au décès d'un individu : maladie, meurtre, accident de voiture, etc. Il peut arriver, par un hasard extraordinaire, qu'un individu puisse être touché par la foudre en même temps que renversé par une voiture, et que ces deux causes soient seules suffisantes pour le tuer sur le coup. Il s'agit là d'un cas de surdétermination, mais ce genre de cas reste exceptionnel. Il faut distinguer ce cas de celui où un même enchaînement causal peut être décrit à plusieurs niveaux. Une maladie générale peut ainsi provoquer un dysfonctionnement local qui entraîne la mort : dire que la personne est morte de la maladie ou du dysfonctionnement qu'elle a provoqué ne revient pas nécessairement à postuler deux chaînes causales indépendantes. C'est là qu'intervient la quatrième thèse, qui est **la non-identité du physique et du mental** : il s'agit de l'idée que les processus physiques cérébraux et les états mentaux qui les accompagnent constituent deux ensembles de faits distincts. [Kim (1998)]

Ces quatre thèses semblent difficilement pouvoir être maintenues simultanément. Par exemple, si on accepte la clôture causale, l'efficacité causale du mental et le principe d'exclusion, il paraît difficile de maintenir la non-identité du physique et du mental. Si en effet, un événement corporel comme le hochement de mes épaules est à la fois causé par des processus physiques dans mon cerveau et par les états mentaux les accompagnant, et que par ailleurs, il ne peut pas y avoir deux chaînes causales différentes menant au même effet, alors cela semble établir que les états mentaux et les états physiques ne correspondent pas à deux ensembles de faits distincts. On est alors poussé à adopter une forme ou une autre de physicalisme. Toujours si on accepte la clôture causale, les deux autres solutions au problème semblent pour le moins étranges. La première consisterait à rejeter l'efficacité causale du mental : notre pensée ne serait alors pas responsable des mouvements corporels que nous décidons d'effectuer. La deuxième consisterait à rejeter le principe d'exclusion causale, ce qui signifierait qu'il y aurait une double chaîne causale responsable de chacun de nos mouvements volontaires. Dans ce cadre, le problème de l'exclusion causal, accompagné de l'acceptation de la clôture causale du physique, rend la solution physicaliste extrêmement attractive.

Le dualiste interactionniste peut évidemment refuser cette conclusion en rejetant directement la clôture causale. Mais il lui faut alors préciser où et comment une cause mentale non physique intervient dans une chaîne physiologique dont une part croissante semble pouvoir être décrite en termes physiques. John Eccles constitue à cet égard un cas particulièrement intéressant. Neurophysiologiste majeur et pourtant dualiste interactionniste, il ne se contente pas d'affirmer abstraitement l'existence d'une action de l'esprit sur le cerveau : il cherche à identifier les processus cérébraux au niveau desquels celle-ci pourrait s'exercer.

Dans ses travaux tardifs, Eccles propose que l'action mentale puisse modifier les probabilités de certains événements microscopiques associés à la transmission synaptique, effets qui seraient ensuite amplifiés par l'activité des réseaux neuronaux. Avec Friedrich Beck, il développe notamment un modèle quantique de l'exocytose synaptique – la libération de molécules au niveau des synapses qui permettent à deux neurones de communiquer – dans lequel une intention consciente pourrait augmenter momentanément

la probabilité de libération de neurotransmetteurs. Eccles propose également de décrire certaines unités fonctionnelles du cortex comme des « dendrons », auxquelles correspondraient des unités mentales appelées « psychons ». L'appel à la mécanique quantique vise précisément à permettre une influence mentale sans supposer qu'une volonté immatérielle injecte simplement de l'énergie dans le cerveau. [Eccles (1992) ; Beck & Eccles (1992)]

Cette proposition présente l'intérêt philosophique de rendre explicite ce qu'exige le rejet de la clôture causale : il ne suffit plus d'affirmer qu'une cause mentale pourrait intervenir ; il faut identifier un point d'intervention et proposer un mécanisme susceptible, au moins en principe, d'être confronté à l'étude du cerveau. Les hypothèses précises d'Eccles n'ont cependant pas reçu de confirmation empirique reconnue. Son cas illustre ainsi à la fois la possibilité logique de rejeter la clôture et la charge empirique qui accompagne ce rejet : à mesure que les mécanismes neuronaux sont décrits sans faire apparaître de rupture nécessitant une cause non physique, l'interactionniste doit indiquer de plus en plus précisément où rechercher cette rupture.

Le second fait notable est que la clôture causale n'est ni un théorème logique ni une loi établie par une expérience unique. David Papineau la présente plutôt comme une **généralisation inductive**⁹ construite progressivement par l'histoire des sciences. Les explications physiques ont réussi de manière croissante à rendre compte des effets physiques sans qu'il soit nécessaire d'introduire des forces vitales, spirituelles ou mentales supplémentaires. Le trajet historique parcouru jusqu'ici — développement des lois de conservation, mécanisation de la physiologie, progrès de la neurologie et intégration physique croissante de la chimie — ne constitue donc pas seulement le contexte intellectuel du principe : il fournit l'essentiel de l'argument empirique en sa faveur. [Papineau (2001)]

Pour de nombreux naturalistes, cette généralisation est d'autant plus persuasive qu'elle s'inscrit dans une histoire plus vaste : celle du remplacement progressif d'explications faisant intervenir des intentions, des finalités, des forces vitales ou des principes spirituels par des explications physiques et physico-chimiques. Le mouvement commence avec les phénomènes naturels — tonnerre, éclairs, tremblements de terre ou éruptions volcaniques — puis s'étend au vivant, avec la physiologie, la biochimie, l'embryologie ou l'hérédité, avant de gagner une part croissante des processus mentaux. Cette histoire est aussi jalonnée de phénomènes dont on a pu penser qu'ils resteraient, par principe, hors de portée d'une explication naturelle. Plusieurs de ces frontières ont pourtant reculé. Cela ne démontre évidemment pas que toute difficulté restante recevra un jour une explication physicaliste ; mais cela rend moins persuasive, pour un naturaliste, la simple proclamation selon laquelle tel phénomène serait nécessairement inexplicable en termes physiques.¹⁰

⁹ Une généralisation inductive est une inférence faite à partir de nombreuses observations concordantes. Si par exemple après des siècles à avoir étudié des tigres, il n'existe aucun témoignage de tigre ailé, on peut inférer que le tigre est un animal sans aile.

¹⁰ Une explication est dite téléologique lorsqu'elle explique un phénomène par la fin ou le but auquel il tend. La biologie contemporaine utilise encore couramment un vocabulaire fonctionnel de ce type, mais cherche généralement à l'expliquer sans faire intervenir une finalité consciente ou un principe vital indépendant, notamment par la sélection naturelle et les mécanismes de régulation des organismes.

L'histoire de la mécanique céleste offre un exemple célèbre. Newton ne pensait pas que l'ordre du système solaire puisse être entièrement compris comme le simple produit de causes mécaniques : il faisait intervenir Dieu dans son explication générale de l'ordonnement du système, et les perturbations mutuelles entre les planètes semblaient pouvoir menacer sa stabilité à long terme. Au siècle suivant, Laplace montra qu'une partie importante de ces irrégularités pouvait au contraire être comprise comme la conséquence des interactions gravitationnelles elles-mêmes : certaines perturbations apparemment cumulatives sont périodiques et se compensent à très long terme. C'est dans ce contexte que se situe la célèbre anecdote selon laquelle, Napoléon lui demandant pourquoi Dieu n'apparaissait pas dans sa Mécanique céleste, Laplace aurait répondu : « Sire, je n'avais pas besoin de cette hypothèse. » La formulation exacte est peut-être légendaire, mais elle résume bien le déplacement conceptuel : ce qui avait semblé exiger une intervention extérieure devenait explicable par les lois mêmes de la mécanique. [Newton (1713) ; Laplace (1799-1825)]

Même les plus grands physiciens du 19^{ème} siècle ont parfois cru rencontrer de telles limites. En 1873, James Clerk Maxwell s'émerveille de l'extraordinaire uniformité des molécules : les molécules d'un même type semblent posséder exactement les mêmes propriétés, où qu'elles se trouvent dans l'Univers. Reprenant une expression de John Herschel, il estime que cette identité leur donne le caractère d'« articles manufacturés ». Maxwell considère alors que l'on atteint ici une limite du pouvoir explicatif des sciences naturelles : selon lui, ni l'existence des molécules ni l'identité exacte de leurs propriétés ne peuvent être attribuées à aucune des causes que nous appelons naturelles. Il affirme même que, sur ce point, « la science doit s'arrêter » et conclut que les constituants fondamentaux de la matière ont été créés. Le problème qu'il identifiait était réel, mais la limite qu'il en déduisait ne l'était pas : la mécanique quantique montrera au siècle suivant que l'uniformité des atomes peut elle-même recevoir une explication physique, leurs propriétés découlant de la structure quantifiée des états permis par les lois de la théorie. [Maxwell (1873)]

On retrouve aujourd'hui encore le même type de proclamations à propos de problèmes encore ouverts, comme l'abiogenèse. Nous ne disposons pas encore d'un récit complet et établi du passage de la chimie prébiotique aux premiers systèmes vivants. Mais l'absence actuelle d'une telle explication n'implique pas, à elle seule, qu'un principe non physique soit nécessaire. L'histoire précédente incite plutôt le naturaliste à distinguer une difficulté scientifique non résolue d'une impossibilité de principe.

Cette tentation de transformer une difficulté actuelle en limite de principe est un thème récurrent. Le théorème d'incomplétude de Gödel a ainsi été invoqué par certains auteurs, notamment J. R. Lucas puis Roger Penrose, pour soutenir que l'esprit humain ne pouvait pas être entièrement computationnel ¹¹ : pour tout système formel suffisamment puissant, il existerait des vérités qu'il ne pourrait démontrer, alors qu'un esprit humain pourrait, du moins en apparence, reconnaître leur vérité. Cet argument demeure très controversé et

¹¹ Dire qu'un processus mental est computationnel revient à dire qu'il fonctionne, au moins en partie, comme un traitement d'informations : certains états servent d'entrées, des règles les transforment, et d'autres états en résultent.

l'incomplétude de Gödel ne constitue pas, à elle seule, une réfutation généralement admise du computationnalisme. [Lucas (1961) ; Penrose (1994)]

Les progrès récents de l'intelligence artificielle offrent un autre exemple de cette naturalisation progressive de fonctions longtemps associées à l'esprit. Des systèmes artificiels peuvent aujourd'hui produire et interpréter du langage, programmer, résoudre certains problèmes mathématiques ou manipuler des concepts abstraits avec une efficacité qui aurait paru spectaculaire quelques décennies plus tôt. Pourtant, personne n'a besoin de postuler, pour expliquer ces performances, une cause qui échapperait au domaine physique : elles sont réalisées par des architectures matérielles et des processus computationnels entièrement intégrés à celui-ci. [Stanford HAI (2026)]

Cela ne montre évidemment ni que l'esprit humain est entièrement computationnel, ni que ces systèmes sont conscients. Mais cela illustre une distinction importante : le fait qu'une capacité paraisse intellectuellement sophistiquée, voire presque « magique » lorsqu'on l'observe, ne constitue pas en soi une raison de penser qu'elle exige un principe mental non physique. L'intelligence artificielle étend ainsi, sur un terrain nouveau, le mouvement historique par lequel des fonctions autrefois attribuées globalement à l'esprit deviennent progressivement accessibles à une explication mécaniste et physique.

Cette justification du principe de clôture causale demeure révisable. Aucune série finie d'observations ne peut démontrer qu'aucune cause non physique n'interviendra jamais dans aucun processus. La force du principe vient plutôt de la convergence de deux constats : l'absence persistante de phénomènes physiques reconnus qui exigeraient une cause non physique et la capacité croissante des sciences à compléter leurs explications sans sortir du domaine physique. La clôture causale doit donc être comprise comme une thèse empirique générale, fortement étayée mais exposée en principe à la découverte d'un contre-exemple.

Plusieurs programmes de recherche controversés ont néanmoins cherché à mettre en évidence des phénomènes qui paraîtraient incompatibles avec la complétude causale du domaine physique. Parmi les plus souvent invoqués figurent les expériences Ganzfeld, destinées à tester une éventuelle transmission extrasensorielle d'informations dans des conditions de réduction sensorielle. Les premières séries ont fait l'objet de critiques portant sur la randomisation, les risques de fuites sensorielles, la sélection des données et les méthodes statistiques. L'amélioration et l'automatisation ultérieures des protocoles n'ont pas entièrement fait disparaître les effets rapportés : plusieurs méta-analyses concluent encore à un résultat moyen légèrement supérieur au hasard. Le dossier reste donc réellement controversé, plutôt que simplement réfuté. Ces résultats n'ont cependant pas convaincu la majorité des scientifiques, parce que l'effet demeure faible, variable selon les études et les méthodes d'analyse, et qu'il n'a pas donné lieu à une réplique indépendante, simple et suffisamment robuste pour établir l'existence d'un transfert d'information inexplicable physiquement. Pour ma part, je considère que cette absence de clarification progressive constitue une raison importante de scepticisme. L'amélioration des protocoles n'aurait pas nécessairement dû accroître la taille de l'effet, puisque les premiers résultats pouvaient être gonflés par des biais ou des défauts de contrôle ; elle aurait toutefois dû

améliorer le rapport entre le signal et le bruit, stabiliser l'estimation de l'effet et permettre de préciser les conditions dans lesquelles il apparaît. Or, après plusieurs décennies de recherches, le Ganzfeld demeure un faible excès statistique dont la taille, la robustesse et les conditions de production restent discutées. Ce profil me paraît davantage compatible avec un effet nul accompagné de biais résiduels, d'hétérogénéité méthodologique et de flexibilité analytique qu'avec l'existence d'un canal réel de transmission de l'information, sans que cette conclusion constitue pour autant une réfutation logique définitive. [Storm et al. (2010) ; Hyman (2010) ; Tressoldi & Storm (2021)]

Les expériences de mort imminente constituent un autre argument fréquemment invoqué. Certaines personnes réanimées rapportent une impression de sortie du corps, le passage dans un tunnel, une lumière intense, la rencontre d'êtres ou la perception de la scène médicale depuis un point de vue extérieur. La force subjective et la cohérence de ces expériences ne sont pas contestées. Mais leurs principales composantes phénoménologiques possèdent des analogues produits par l'hypoxie (manque d'oxygène), certaines substances psychoactives, l'anesthésie, les crises épileptiques, la stimulation cérébrale ou d'autres états altérés de conscience. Surtout, les récits de perceptions prétendument véridiques obtenues hors du corps reposent généralement sur des témoignages rétrospectifs dont le moment exact et les conditions d'acquisition ne peuvent être établis avec certitude. Les recherches prospectives utilisant des informations ou des cibles placées hors de la vue des patients n'ont jusqu'ici fourni aucun cas largement reconnu comme démontrant l'acquisition d'informations en l'absence de fonctionnement cérébral capable de la permettre. [Parnia et al. (2023)]

Ni les expériences Ganzfeld ni les expériences de mort imminente ne peuvent donc être simplement ignorées : elles constituent des objets de recherche légitimes et certaines données continuent de faire débat. Mais elles n'ont pas produit de résultat assez robuste, reproductible et indépendant pour être reconnues comme des contre-exemples à la clôture causale. Leur intérêt actuel réside moins dans une réfutation de la clôture causale que dans la possibilité, toujours ouverte en principe, de soumettre cette thèse à des tests empiriques.

Le principe de clôture causale fonctionne également, le plus souvent de manière implicite, dans des pratiques de raisonnement extérieures aux sciences naturelles. La médecine légale et le droit recherchent ainsi des chaînes causales établissables à partir de mécanismes naturels, de traces matérielles et de régularités publiquement contrôlables. Un tribunal ne retient pas l'envoûtement, la possession ou l'intervention d'une volonté immatérielle comme explication concurrente d'un événement, simplement parce que leur inexistence absolue n'a pas été démontrée.

Cela ne signifie pas que les institutions juridiques adoptent officiellement une doctrine métaphysique physicaliste. Leur engagement est d'abord méthodologique et lié aux exigences de preuve : une possibilité purement imaginable, qui n'est soutenue par aucun indice positif ni intégrée à un cadre explicatif fiable, ne constitue pas une hypothèse recevable. La présomption d'innocence n'oblige donc pas à traiter toute cause surnaturelle concevable comme un doute raisonnable. Sous cette forme pratique et limitée, l'idée que les

événements physiques doivent recevoir des explications physiquement établissables borne silencieusement l'espace ordinaire de nos raisonnements causaux.

Une dernière précision technique est nécessaire. La formulation selon laquelle « tout effet physique possède une cause physique suffisante » semble naturelle dans le cadre d'une physique déterministe : un état physique antérieur et les lois déterminent alors entièrement l'état ultérieur. Elle devient moins claire lorsque la théorie fondamentale n'attribue aux différents résultats que des probabilités.

Cela ne signifie pourtant pas que l'indéterminisme ouvrirait automatiquement une brèche dans la clôture causale. Une théorie physique probabiliste peut être complète au sens où elle fixe entièrement les probabilités des issues possibles et ne laisse aucun rôle explicatif à une variable non physique supplémentaire. Une intervention mentale qui sélectionnerait l'une de ces issues constituerait toujours une addition à la dynamique physique, même si elle ne violait pas une évolution strictement déterministe.

Il est donc préférable de formuler le principe en termes de complétude causale ou explicative : pour tout événement physique, les lois et les conditions physiques fournissent toute l'explication causale requise par la meilleure théorie physique, qu'elles déterminent un résultat unique ou seulement une distribution de probabilités. Cette formulation reste neutre entre les théories quantiques fondamentalement déterministes, comme les approches bohmiennes et everettiennes, et les théories introduisant de véritables événements stochastiques, comme les modèles de réduction dynamique. Elle préserve le cœur du principe : aucun élément non physique supplémentaire n'est nécessaire pour compléter l'explication de l'évolution du domaine physique.

IX. Du behaviorisme au fonctionnalisme : de la disposition au rôle causal

Nous revenons maintenant à notre histoire de la montée en puissance du physicalisme au cours du XX^e siècle, en nous penchant sur l'avènement de l'une de ses formulations les plus influentes : **le fonctionnalisme**. Cette position peut être comprise comme le point de convergence de plusieurs programmes de recherche. D'un côté, le behaviorisme avait cherché à définir scientifiquement les états mentaux à partir de leurs relations avec les stimuli et les comportements observables. De l'autre, la cybernétique, la théorie du calcul et les premières sciences informatiques avaient montré qu'il était possible de décrire des systèmes très différents par les fonctions qu'ils accomplissent et par les relations entre leurs entrées, leurs états internes et leurs sorties, indépendamment de la matière précise dont ils sont constitués. Le fonctionnalisme conservera quelque chose de ces deux intuitions, tout en réintroduisant explicitement des états internes que le behaviorisme avait largement évacués.

Le **behaviorisme** de John Watson, formulé en 1913, peut être lu comme l'aboutissement, côté science du comportement, de ce que Sherrington établit côté nerf et Loeb côté matière organique : Watson refuse d'invoquer toute cause mentale dans l'explication du comportement, et ne veut plus traiter que du couple stimulus-réponse, directement observable et mesurable. [Watson (1913)]

Le béhaviorisme ne reste cependant pas cantonné à une règle méthodologique pour la psychologie expérimentale. Il reçoit également, au milieu du XX^e siècle, une formulation philosophique. Alors que Watson propose de ne prendre pour objets scientifiques que les relations observables entre stimuli et réponses, le béhaviorisme logique cherche à analyser la signification même des termes mentaux à partir de comportements ou de dispositions à se comporter.

Dire qu'une personne croit qu'il va pleuvoir, qu'elle souffre ou qu'elle désire boire ne renverrait donc pas à un événement privé se déroulant dans une substance mentale inaccessible. Cela reviendrait à dire qu'elle est disposée, dans certaines circonstances, à prendre un parapluie, à se plaindre, à éviter une blessure ou à chercher de l'eau. Cette orientation se retrouve sous des formes différentes chez Carnap et Hempel, puis dans *The Concept of Mind* de Gilbert Ryle, publié en 1949. Ryle y dénonce comme une erreur de catégorie l'idée cartésienne selon laquelle l'esprit serait une chose supplémentaire, comparable au corps mais cachée derrière lui — ce qu'il appelle ironiquement le « fantôme dans la machine ». Il illustre ce type d'erreur par l'exemple d'un visiteur à qui l'on aurait montré les collèges, les bibliothèques, les amphithéâtres et les bureaux d'une université, et qui demanderait ensuite : « Mais où est l'Université elle-même ? » La question suppose à tort que l'Université doit être une chose supplémentaire, placée à côté de toutes les institutions qui la composent. De la même manière, selon Ryle, chercher « l'esprit » comme une entité distincte située derrière les comportements, les dispositions et les capacités d'une personne revient à se tromper sur le type de chose que désigne le concept d'esprit. [Ryle (1949)]

Le béhaviorisme philosophique permet ainsi de dissoudre une partie du problème corps-esprit : si les prédicats mentaux désignent des dispositions comportementales plutôt que des états d'une substance intérieure, il n'existe plus deux domaines qu'il faudrait faire interagir. Mais cette stratégie paraît laisser de côté l'existence d'**états internes** qui peuvent varier sans produire immédiatement de comportement observable. Une personne peut éprouver une douleur tout en la dissimulant ; inversement, elle peut simuler tous les comportements associés à la douleur sans la ressentir. Ces difficultés contribueront à faire abandonner le béhaviorisme strict au profit de théories qui réintroduisent des états internes, d'abord en les identifiant à des états cérébraux, puis en les définissant par leur rôle causal dans une organisation plus large.

Le **fonctionnalisme** naîtra précisément de cette double insatisfaction : conserver l'idée béhavioriste que les états mentaux doivent être caractérisés par ce qu'ils font, tout en réintroduisant des états internes capables de jouer un rôle causal entre les entrées sensorielles et les comportements. Avant d'y parvenir, la **théorie de l'identité esprit-cerveau** fournit une première manière de réinternaliser le mental en l'identifiant directement à des états cérébraux.

La théorie de l'identité esprit-cerveau, présentée en section 7, ne survit pas longtemps sous sa forme initiale – et c'est la révision qu'elle subit qui installe le cadre théorique dominant pour les décennies suivantes.

Il faut d'abord préciser de quelle identité il s'agit. La théorie défendue par Place et Smart est généralement comprise comme **une identité type à type** : chaque type d'état mental serait identique à un type déterminé d'état cérébral. L'exemple devenu canonique est celui de la douleur identifiée à l'activation des fibres C. Ces fibres existent bien et participent à la transmission de certains signaux associés à la douleur. Mais l'équation « douleur = activation des fibres C » doit être comprise ici comme une illustration philosophique volontairement simplifiée, et non comme une identification validée par les neurosciences : la douleur consciente dépend de processus beaucoup plus complexes, et l'activité de ces fibres ne suffit pas, à elle seule, à l'expliquer.

Une théorie d'identité type à type affirme néanmoins quelque chose de cette forme : partout où se produit un même type d'état mental, on devrait retrouver le même type physique. Si la douleur était réellement identique à l'activation des fibres C, aucun organisme dépourvu de telles fibres ne pourrait éprouver de douleur.

Hilary Putnam porte à cette thèse une objection devenue classique dans « Psychological Predicates », publié en 1967 et repris en 1975 sous le titre « The Nature of Mental States ». C'est l'argument de la **réalisation multiple**. Des organismes très différents — êtres humains, autres mammifères, oiseaux ou pieuvres — semblent capables d'éprouver de la douleur alors même que leurs systèmes nerveux ne possèdent pas nécessairement la même organisation physique. On peut également concevoir qu'une machine ou un extraterrestre constitué d'un autre substrat réalise le même état mental. Si la douleur était strictement identique à un type cérébral particulier, ces systèmes ne pourraient donc pas ressentir de douleur à moins de partager exactement ce même type physique. [Putnam (1967)]

L'idée peut être comprise avec une analogie simple. Un même type d'outil peut être réalisé de manières physiquement très différentes. Un piège à souris peut fonctionner grâce à un ressort, à une trappe, à une cage ou à un capteur électronique. Ce qui permet de regrouper ces dispositifs sous une même catégorie n'est pas la matière dont ils sont faits ni la forme précise de leurs composants, mais **le rôle qu'ils remplissent** : réagir à la présence d'une souris de manière à la capturer.

Pour le **fonctionnaliste**, il en va de même des états mentaux. La douleur ne serait pas définie par un tissu nerveux particulier, mais par **la place qu'elle occupe dans le fonctionnement de l'organisme** : elle est généralement causée par une lésion ou une stimulation nocive, modifie les croyances et les désirs, attire l'attention et tend à provoquer l'évitement, le retrait ou la plainte. Un cerveau humain, le système nerveux d'une pieuvre ou un dispositif artificiel pourraient donc réaliser le même état fonctionnel par des mécanismes physiques différents.

Cette thèse ne doit cependant pas être confondue avec l'identité occurrence à occurrence. Un fonctionnaliste physicaliste peut soutenir que chaque épisode mental particulier est réalisé par un événement physique particulier. Mais le fonctionnalisme, considéré isolément, affirme d'abord que les types mentaux sont définis par leurs rôles causaux ; il ne détermine pas à lui seul le statut métaphysique ultime de chacune de leurs occurrences.

Le recours de Putnam au vocabulaire fonctionnel s'inscrit dans le contexte plus général du développement de l'**informatique** et des **sciences cognitives**. La notion de **machine de Turing**¹² avait montré qu'un calcul pouvait être défini abstraitement par un ensemble d'états et de règles de transition, indépendamment de la matière dont la machine chargée de l'exécuter est constituée. Une même organisation computationnelle peut ainsi être réalisée par des dispositifs physiques différents. [Turing (1936)]

Putnam transpose ce modèle à l'esprit dans ce qu'on appellera le fonctionnalisme-machine. Un état mental est défini par la place qu'il occupe dans l'organisation globale du système : les entrées susceptibles de le produire, ses relations avec les autres états internes et les sorties comportementales auxquelles il contribue. David Lewis prolonge cette approche en 1972, dans « Psychophysical and Theoretical Identifications », avec une version dite du rôle causal : les termes mentaux du langage ordinaire — douleur, croyance, désir — définissent implicitement un rôle causal, et l'état mental correspondant est simplement ce qui occupe ce rôle dans un système donné, qu'il s'agisse d'un cerveau humain, d'un cerveau de pieuvre ou, en principe, d'un circuit de silicium. [Lewis (1972)]

Le fonctionnalisme s'impose rapidement comme la position dominante en philosophie de l'esprit, à partir de la fin des années 1960 et pendant toute la décennie 1970, pour une raison qui dépasse la seule philosophie : il offre un cadre théorique naturel pour les sciences cognitives naissantes et pour l'intelligence artificielle. Si l'esprit se définit par son organisation fonctionnelle plutôt que par son substrat, alors il devient en principe possible qu'un ordinateur, correctement organisé, pense — une conséquence que Putnam assumait explicitement, et qui vaut au fonctionnalisme un lien étroit avec le programme de l'IA classique de cette période.

Cette victoire ne va cependant pas sans opposition. Ned Block, dans "Troubles with Functionalism" (1978), imagine un scénario resté célèbre : si l'ensemble de la population chinoise, suivant un protocole de communication précis, se mettait à simuler collectivement l'organisation fonctionnelle d'un cerveau humain — chaque personne jouant le rôle d'un neurone, transmettant des signaux selon les mêmes relations causales —, ce système géant réaliserait-il un véritable esprit, avec une expérience subjective ? L'intuition de Block est que non : le système reproduirait fidèlement l'organisation fonctionnelle requise, sans que rien ne garantisse qu'une expérience consciente en résulte. C'est l'argument dit des **qualia**¹³ **absents** : un système peut, semble-t-il, satisfaire entièrement la description fonctionnelle d'un état mental sans que cet état s'accompagne d'aucune expérience vécue. Face à cette objection, la majorité des fonctionnalistes choisissent d'avalier la pastèque plutôt que d'abandonner la thèse : ils acceptent que, aussi étrange que cela paraisse intuitivement, un tel système satisferait bien les conditions requises pour avoir une expérience subjective — la

¹² Une machine de Turing est un modèle abstrait de calcul proposé par Alan Turing en 1936. Elle consiste, schématiquement, en un dispositif capable de lire et d'écrire des symboles sur un ruban, puis de changer d'état selon un ensemble de règles. Son intérêt est de montrer qu'un calcul peut être défini par l'organisation des opérations réalisées, indépendamment du support matériel qui les exécute.

¹³ Un qual (au pluriel, qualia) désigne caractère qualitatif d'une expérience vécue : ce que cela fait, par exemple, de voir du rouge, d'éprouver une douleur ou de sentir une odeur.

bizarrerie de la conclusion ne suffisant pas, à leurs yeux, à invalider l'argument fonctionnaliste lui-même. [Block (1978)]

X. Vers le problème difficile : de Nagel à Chalmers

Les neuf sections précédentes ont raconté l'histoire d'une victoire progressive du physicalisme, consolidée par la clôture causale du physique. Mais c'est précisément à l'intérieur de ce camp victorieux, et non contre lui de l'extérieur, qu'un problème résiduel va resurgir dans le dernier tiers du XXe siècle — porté par des auteurs qui, pour la plupart, ne se pensent pas eux-mêmes comme des adversaires du physicalisme, mais comme des témoins de ses limites.

Thomas Nagel ouvre cette séquence en 1974 avec un article resté célèbre, *What Is It Like to Be a Bat?* (en Français : « Qu'est-ce que cela fait d'être une chauve-souris ? »). Son argument part d'une observation simple : pour tout organisme conscient, il y a quelque chose que cela fait d'être cet organisme — une manière subjective dont le monde se présente à lui. Nagel choisit la chauve-souris précisément parce qu'elle perçoit le monde par écholocation, un sens dont les humains sont dépourvus : même en connaissant absolument tout de la physique, de la neurologie et du comportement de la chauve-souris, rien dans cette description ne semble nous donner accès à ce que cela fait, du point de vue de la chauve-souris elle-même, de percevoir ainsi. C'est cette dimension subjective de l'expérience que l'on appelle la **phénoménalité** : le fait qu'un état conscient possède un certain caractère vécu, une certaine manière d'être éprouvé de l'intérieur. Le point de Nagel est méthodologique autant que métaphysique : la science physique procède par objectivation, elle cherche à décrire le monde indépendamment de tout point de vue particulier — ce qu'on appelle parfois une « vue de nulle part ». Or la phénoménalité semble, par construction, être liée à l'existence d'un point de vue subjectif. [Nagel (1974)]

C'est Saul Kripke qui, en 1980 dans *Naming and Necessity*, fournit sans le vouloir l'outil conceptuel qui permettra à ses successeurs de durcir l'intuition de Nagel en argument véritable. Kripke part d'un principe : une identité scientifique vraie, découverte a posteriori — comme "la chaleur est le mouvement moléculaire" — est, si elle est vraie, nécessairement vraie ; il n'existe aucun monde possible où la chaleur ne serait pas du mouvement moléculaire. Pour le dire autrement : nous ignorions avant le XIXème siècle que la chaleur correspondait au mouvement moléculaire. Mais dès que nous avons compris que les conséquences causales de la chaleur pouvaient entièrement être déduites par le mouvement moléculaire, le lien entre chaleur et mouvement moléculaire apparaissait nécessaire. La chaleur ne peut pas être autre chose que le mouvement moléculaire, et le mouvement moléculaire réalise nécessairement la chaleur. Mais alors, comment expliquer qu'on ait malgré tout l'impression qu'elle aurait pu ne pas l'être ? Kripke répond que cette impression de contingence est une illusion dissoluble : ce qui semble vraiment possible, c'est que le mouvement moléculaire ait existé sans produire en nous la sensation de chaleur — puisque "chaleur" désigne le phénomène physique lui-même, non la sensation qu'il cause habituellement, un écart peut s'ouvrir entre les deux, et c'est cet écart qui explique l'illusion sans menacer l'identité. [Kripke (1980)]

Kripke applique ensuite ce même raisonnement au cas de la douleur et de l'activation des fibres C. Si cette identité est vraie, elle doit elle aussi être **nécessaire**, et l'impression de contingence qu'on peut avoir à son sujet devrait, de la même manière, pouvoir s'expliquer comme une **illusion**. Mais ici le mécanisme se bloque : contrairement à la chaleur, il n'existe aucun écart entre la douleur et la sensation de douleur — être en douleur, c'est ressentir cette sensation, rien de plus. On ne peut donc pas dire que l'activation des fibres C aurait pu exister sans être ressentie comme douleur tout en restant, par ailleurs, de la douleur : si elle n'est pas ressentie comme douleur, elle n'est tout simplement pas de la douleur. L'écart qui permettait de dissoudre l'illusion dans le cas de la chaleur est donc structurellement absent ici — non pas provisoirement introuvable, mais impossible en principe, puisque la douleur est elle-même la sensation qu'elle est censée expliquer. L'apparente contingence de "la douleur est l'activation des fibres C" ne peut donc pas être expliquée comme une illusion, ce qui suggère, à la différence du cas de la chaleur, qu'elle n'en est pas une : douleur et activation des fibres C seraient alors deux choses réellement distinctes.

Frank Jackson donne à ce type d'intuition sa forme la plus mémorable avec l'argument de la connaissance, dans "Epiphenomenal Qualia" (1982). Il imagine Mary, une brillante neuroscientifique qui a passé toute sa vie enfermée dans une chambre en noir et blanc, mais qui a eu accès à la totalité du savoir physique concernant la vision des couleurs — chaque longueur d'onde, chaque processus rétinien, chaque voie neuronale impliquée. Le jour où Mary sort de sa chambre et voit une tomate rouge pour la première fois, apprend-elle quelque chose de nouveau ? L'intuition de Jackson, largement partagée, est que oui : elle apprend ce que ça fait de voir du rouge, une information qu'aucune description physique, aussi complète soit-elle, ne lui avait donnée. Si Mary apprend un fait nouveau alors qu'elle connaissait déjà tous les faits physiques, c'est qu'il existe des faits — les faits phénoménaux — qui ne se réduisent pas aux faits physiques. Jackson défendra initialement, à partir de cet argument, un **dualisme des propriétés** : les états cérébraux ont des propriétés physiques (par exemple, des neurones qui s'activent et communiquent) et des propriétés phénoménales (les fameux qualia, par exemple, ce que ça fait de voir du rouge et de ressentir de la douleur). Il est notable qu'il reviendra plus tard sur cette conclusion : dans "Mind and Illusion" (2003), il finira par accepter une forme de physicalisme, jugeant que l'intuition de nouveauté que Mary semble éprouver peut s'expliquer sans postuler de faits non physiques — un revirement rare venant de l'auteur de l'argument anti-physicaliste le plus cité du XX^e siècle. [Jackson (1982, 2003)]

Entre Kripke et Jackson d'un côté, et Chalmers de l'autre, Joseph Levine introduit en 1983 une nuance décisive avec la notion de fossé explicatif ("Materialism and Qualia: The Explanatory Gap"). Levine part d'une comparaison : même si l'on accepte que la douleur est bel et bien l'activation des fibres C — c'est-à-dire même si l'on accepte l'identité —, il manque encore quelque chose que l'on possède dans d'autres cas d'identité scientifique. Savoir que l'eau est H₂O permet, via la théorie chimique, d'expliquer pourquoi l'eau bout à telle température, pourquoi elle est transparente, pourquoi elle dissout le sel : la structure microphysique explique les propriétés macroscopiques. Rien de comparable n'existe pour la

douleur : même en acceptant que la douleur soit l'activation des fibres C, aucune théorie ne montre pourquoi cette activation particulière devrait produire cette sensation-là plutôt qu'une autre, ou plutôt que rien du tout. Le geste de Levine est cependant plus prudent que celui de Jackson : il insiste sur le fait que ce fossé est de nature épistémique — un manque dans notre compréhension actuelle — et qu'il n'implique pas nécessairement de conclusion ontologique. On peut, selon lui, rester matérialiste tout en admettant que notre appareil explicatif actuel ne comble pas ce fossé. [Levine (1983)]

C'est David Chalmers qui, en 1995 puis dans *The Conscious Mind* (1996), rassemble et systématise ces différentes difficultés. Il distingue d'abord les **problèmes « faciles »** de la conscience — expliquer comment le cerveau discrimine les stimuli, intègre l'information, contrôle le comportement ou permet à un sujet de rapporter ses états internes — **du problème difficile : pourquoi l'accomplissement de toutes ces fonctions s'accompagne-t-il d'une expérience subjective plutôt que de se dérouler sans que rien ne soit vécu de l'intérieur ?** [Chalmers (1995, 1996)]

Chalmers reprend alors et généralise la stratégie modale introduite par Kripke. Kripke demandait si un état cérébral déterminé pouvait exister sans l'expérience à laquelle on prétendait l'identifier. Chalmers étend cette dissociation à la totalité du domaine physique avec l'argument des **zombies philosophiques**. Un zombie serait un être physiquement et fonctionnellement identique à un être humain — même cerveau, même comportement, mêmes paroles et mêmes réactions — mais entièrement dépourvu d'expérience subjective.

Si un tel être est métaphysiquement possible, alors l'ensemble complet des faits physiques ne suffit pas à déterminer les faits phénoménaux. Le physicalisme, compris comme la thèse selon laquelle une duplication parfaite du physique entraîne nécessairement une duplication de tout ce qui existe, serait donc faux. Cette conclusion ne contredit cependant pas directement la clôture causale du physique. Chalmers peut admettre que tous les effets physiques possèdent une explication physique complète, tout en soutenant que des propriétés phénoménales supplémentaires accompagnent certains processus cérébraux en vertu de lois psychophysiques. Son argument vise ainsi la réduction ontologique de la conscience au physique, non la complétude causale des explications physiques.

Comme chez Kripke, toute la difficulté tient cependant au passage entre ce que nous pouvons concevoir et ce qui est réellement possible. Le simple fait qu'un zombie nous paraisse imaginable ne suffit pas à montrer qu'un monde physiquement identique au nôtre mais dépourvu de conscience est métaphysiquement possible. [Chalmers (1996, 2002)]

Chalmers cherche donc à préciser ce que signifie ici « concevoir ». Il distingue notamment une concevabilité superficielle, qui peut simplement refléter notre ignorance, d'une concevabilité idéale : une situation resterait concevable même après une réflexion rationnelle suffisamment approfondie et avec toutes les informations pertinentes disponibles.

Cette distinction est importante parce que certaines identités scientifiques paraissent d'abord contingentes alors qu'elles sont en réalité nécessaires. Avant de connaître la chimie, on pouvait par exemple imaginer que l'eau ne soit pas H₂O. Mais, une fois découvert que la

substance que nous appelons « eau » est H_2O , il n'existe pas réellement de monde possible dans lequel cette même substance serait autre chose que H_2O . Ce que nous imaginions était en fait un monde contenant un liquide ressemblant à l'eau, mais possédant une autre composition chimique.

La « sémantique bidimensionnelle » développée par Chalmers cherche précisément à formaliser cette différence entre deux questions : ce que nous aurions appelé « eau » si le monde s'était révélé différent de ce que nous pensions, et ce que la substance effectivement désignée par le mot « eau » aurait pu être une fois son identité réelle connue. Chalmers soutient que cette stratégie permet d'expliquer l'apparente concevabilité de la négation de nombreuses identités scientifiques (pourquoi nous trouvons concevable que l'eau ne soit pas H_2O) sans conclure à une différence réelle, mais qu'elle ne dissout pas de la même manière le cas de la conscience. Selon lui, même en fixant tous les faits physiques et en disposant de toutes les informations pertinentes, un monde physiquement identique au nôtre mais dépourvu d'expérience subjective resterait concevable.

Cette étape demeure l'un des points les plus contestés de son raisonnement. Ses adversaires peuvent reconnaître l'existence d'un fossé conceptuel entre les descriptions physiques et phénoménales tout en niant que ce fossé révèle une possibilité métaphysique réelle. Ils soutiennent notamment que nos concepts phénoménaux pourraient masquer une identité nécessaire avec des propriétés physiques, de sorte que notre incapacité à déduire cette identité, ou à concevoir autrement le rapport entre cerveau et expérience, signalerait une limite cognitive plutôt qu'une distinction ontologique. Chalmers va ainsi plus loin que Nagel, qui restait agnostique, et que Levine, qui limitait son fossé à l'épistémologie : il transforme un problème de description en un problème d'ontologie, ce qui vaut à sa formulation de devenir la référence standard sous le nom de « problème difficile de la conscience ». [Loar (1990/1997)]

Il vaut la peine de noter, en clôture de cette section, la parenté structurelle entre cette inférence et celle qui ouvrirait ce texte. L'argument de Descartes en Méditation VI — concevoir clairement et distinctement l'esprit séparé du corps, en tirer une possibilité réelle, puis une distinction de substance — et l'argument de Kripke sur la douleur et les fibres C, repris et durci par Chalmers avec les zombies philosophiques, partagent le même schéma : de l'échec à concevoir autrement, ou à dissoudre une intuition modale, à une conclusion ontologique sur la distinction des choses. La différence tient à la garantie invoquée : Descartes s'appuyait sur un Dieu non trompeur pour assurer que toute conception claire et distincte correspond à une possibilité réelle — une garantie externe, dont ce texte a souligné plus haut la fragilité face à l'objection d'Arnauld. Kripke, plus prudent, ne postule aucune garantie de ce genre : il cherche systématiquement un mécanisme susceptible d'expliquer l'apparence de contingence comme une illusion, et ne conclut à une distinction réelle que là où ce mécanisme échoue structurellement. Cette prudence ne met cependant pas l'argument à l'abri d'une objection de même nature : l'échec à expliquer l'illusion pourrait signaler une limite de nos concepts phénoménaux — opaques, insuffisamment structurés pour révéler leurs conditions de réalisation physique, comme le suggèrent des auteurs tels que Brian Loar — plutôt qu'une distinction métaphysique réelle. Trois siècles séparent Descartes de

Chalmers, mais l'un et l'autre franchissent, chacun à sa manière, le même pas entre ce que l'esprit n'arrive pas à concevoir autrement et ce qui est réellement le cas.

Conclusion

L'histoire que nous venons de parcourir permet d'abord de mesurer à quel point les débats contemporains sur la conscience sont étroitement liés aux transformations successives des sciences. Le problème de l'esprit et du corps n'est pas resté identique depuis Descartes : il s'est déplacé à mesure que la physique, la physiologie, la neurologie, la chimie et, plus récemment, les sciences cognitives ont modifié ce qu'il devenait raisonnable d'attendre d'une explication naturelle.

Les progrès scientifiques n'ont pas, à eux seuls, tranché les questions métaphysiques. Mais ils ont profondément transformé l'espace des positions disponibles. Des phénomènes autrefois expliqués par l'action d'agents, de finalités, de forces vitales ou de principes propres au vivant ont progressivement été intégrés à des chaînes explicatives physiques, chimiques et biologiques. L'histoire du problème de la conscience illustre ainsi une idée plus générale du naturalisme philosophique : les problèmes philosophiques ne se développent pas indépendamment des sciences. Les avancées scientifiques ne dictent pas automatiquement la bonne métaphysique, mais elles modifient continuellement les contraintes auxquelles celle-ci doit satisfaire, affaiblissent certaines intuitions et en font émerger de nouvelles.

C'est dans ce contexte qu'a pris une importance particulière le principe de clôture causale du physique : l'idée selon laquelle l'évolution du domaine physique n'exige aucune cause non physique supplémentaire pour être expliquée. Ce principe n'est ni une vérité logique ni une doctrine universellement acceptée. Il demeure contesté notamment par certains dualistes interactionnistes et certains émergentistes. Il structure néanmoins une grande partie de la philosophie de l'esprit contemporaine, parce qu'il rend particulièrement difficile toute théorie qui attribuerait à un esprit non physique une efficacité causale propre sur le cerveau et le comportement.

Un point essentiel doit alors être souligné : le problème difficile de la conscience ne constitue pas, en lui-même, une remise en cause de cette clôture causale. Il apparaît au contraire largement à l'intérieur d'un cadre où celle-ci est tenue pour acquise. On peut admettre que les mouvements du corps, les discriminations perceptives, les décisions, les rapports verbaux et l'ensemble des traitements fonctionnels réalisés par le cerveau reçoivent une explication physique complète, tout en continuant à demander pourquoi ces processus s'accompagnent d'une expérience subjective.

C'est même, en un sens, ce qui rend le problème particulièrement difficile. À mesure que les fonctions autrefois attribuées à une âme ou à un esprit distinct deviennent accessibles à une explication physique et fonctionnelle, la phénoménalité se trouve de plus en plus isolée. Le problème difficile apparaît ainsi non pas malgré les succès de l'explication naturaliste, mais en partie à cause d'eux : une fois expliqués les mécanismes de perception, de comportement, de mémoire, de raisonnement ou de contrôle de l'action, demeure la

question de savoir pourquoi leur accomplissement s'accompagne de quelque chose que cela fait de les vivre.

Le parcours historique fait également apparaître plusieurs des grandes familles de positions que nous retrouverons dans les débats actuels. Le matérialisme puis le physicalisme prolongent la stratégie consistant à faire de l'esprit un aspect du monde physique lui-même. Le dualisme maintient, sous des formes profondément renouvelées, qu'une dimension de la réalité mentale ne peut être identifiée au physique. Le monisme neutre cherche une réalité plus fondamentale dont le mental et le physique constitueraient deux manifestations ou deux aspects. L'émergentisme soutient que certaines propriétés peuvent apparaître à des niveaux d'organisation supérieurs sans être simplement identiques aux propriétés de leurs composants. Le mystérianisme, enfin, envisage que la difficulté ne provienne pas nécessairement de la nature de la conscience elle-même, mais des limites cognitives de l'esprit humain lorsqu'il tente de la comprendre.

Ces catégories ne doivent cependant pas donner l'illusion d'un paysage simple. Chacune recouvre des positions parfois très différentes, et certains des désaccords les plus profonds opposent des auteurs qui appartiennent pourtant à la même grande famille. Il existe plusieurs physicalismes, plusieurs dualismes, plusieurs formes d'émergentisme ou de monisme neutre, et leurs différences concernent aussi bien le statut de la phénoménalité que la causalité mentale, la réduction, l'explication ou encore le rapport entre concevabilité et possibilité métaphysique.

Le problème difficile n'est donc pas une question isolée apparue soudainement avec Chalmers dans les années 1990. Sa formulation contemporaine est l'aboutissement provisoire d'une longue transformation du problème de l'esprit et du corps sous la pression conjointe des arguments philosophiques et des découvertes scientifiques. Le second article partira de ce point pour examiner les principales réponses aujourd'hui proposées : ce qu'elles prétendent expliquer, ce qu'elles doivent abandonner ou modifier, et quels coûts théoriques chacune d'elles accepte de payer pour rendre compte de l'existence d'une expérience subjective.

Bibliographie

- Adrian, E. D. (1932). « The Activity of the Nerve Fibres ». Nobel Lecture, 12 décembre 1932.
- Alexander, S. (1920). *Space, Time, and Deity*. 2 vol. London : Macmillan.
- Arnauld, A. (1641). « Quatrièmes Objections », dans R. Descartes, *Méditations métaphysiques*.
- Avery, O. T., MacLeod, C. M. & McCarty, M. (1944). « Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types ». *Journal of Experimental Medicine*, 79(2), p. 137-158.
- Beck, F. & Eccles, J. C. (1992). « Quantum Aspects of Brain Activity and the Role of Consciousness ». *Proceedings of the National Academy of Sciences*, 89(23), p. 11357-11361.
- Block, N. (1978). « Troubles with Functionalism », dans C. W. Savage (dir.), *Perception and Cognition: Issues in the Foundations of Psychology*, Minnesota Studies in the Philosophy of Science, vol. 9, p. 261-325. Minneapolis : University of Minnesota Press.
- Born, M. (1926). « Zur Quantenmechanik der Stoßvorgänge ». *Zeitschrift für Physik*, 37, p. 863-867.
- Broad, C. D. (1925). *The Mind and Its Place in Nature*. London : Routledge & Kegan Paul.
- Buchner, E. (1897). « Alkoholische Gährung ohne Hefezellen ». *Berichte der Deutschen Chemischen Gesellschaft*, 30, p. 117-124.
- Carnap, R. (1932). « Psychologie in physikalischer Sprache ». *Erkenntnis*, 3, p. 107-142.
- Chalmers, D. J. (1995). « Facing Up to the Problem of Consciousness ». *Journal of Consciousness Studies*, 2(3), p. 200-219.
- Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York : Oxford University Press.
- Chalmers, D. J. (2002). « Does Conceivability Entail Possibility? », dans T. S. Gendler & J. Hawthorne (dir.), *Conceivability and Possibility*, p. 145-200. Oxford : Oxford University Press.
- Compton, A. H. (1935). *The Freedom of Man*. New Haven : Yale University Press.
- Crick, F. (1994). *The Astonishing Hypothesis: The Scientific Search for the Soul*. New York : Scribner.
- Darwin, C. (1871). *The Descent of Man, and Selection in Relation to Sex*. London : John Murray.
- Descartes, R. (1641). *Meditationes de prima philosophia* [Méditations métaphysiques]. Paris.
- Diderot, D. (1769). *Le Rêve de d'Alembert* (rédigé en 1769 ; publié en 1830).
- Driesch, H. (1908). *The Science and Philosophy of the Organism: The Gifford Lectures Delivered before the University of Aberdeen in the Year 1907-1908*. London : A. & C. Black.
- du Bois-Reymond, E. (1872). *Über die Grenzen des Naturerkennens*. Leipzig : Veit & Comp.

- Eccles, J. C. (1992). « Evolution of Consciousness ». *Proceedings of the National Academy of Sciences*, 89(16), p. 7320-7324.
- Eddington, A. S. (1928). *The Nature of the Physical World*. Cambridge : Cambridge University Press.
- Élisabeth de Bohême & Descartes, R. (1643). *Correspondance sur l'union de l'âme et du corps, notamment les lettres de mai-juillet 1643*.
- Feigl, H. (1958). « The 'Mental' and the 'Physical' », dans H. Feigl, M. Scriven & G. Maxwell (dir.), *Minnesota Studies in the Philosophy of Science*, vol. 2, p. 370-497. Minneapolis : University of Minnesota Press.
- Franklin, R. E. & Gosling, R. G. (1953). « Molecular Configuration in Sodium Thymonucleate ». *Nature*, 171, p. 740-741.
- Heitler, W. & London, F. (1927). « Wechselwirkung neutraler Atome und homöopolare Bindung nach der Quantenmechanik ». *Zeitschrift für Physik*, 44, p. 455-472.
- Hempel, C. G. (1969). « Reduction: Ontological and Linguistic Facets », dans S. Morgenbesser, P. Suppes & M. White (dir.), *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*, p. 179-199. New York : St. Martin's Press.
- Hobbes, T. (1641). « Troisièmes Objections », dans R. Descartes, *Méditations métaphysiques*.
- Hodgkin, A. L. & Huxley, A. F. (1952). « A Quantitative Description of Membrane Current and Its Application to Conduction and Excitation in Nerve ». *The Journal of Physiology*, 117(4), p. 500-544.
- Holbach, P.-H. Thiry d' (1770). *Système de la nature, ou Des lois du monde physique et du monde moral*. Londres [Amsterdam].
- Hubel, D. H. & Wiesel, T. N. (1962). « Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex ». *The Journal of Physiology*, 160(1), p. 106-154.
- Huxley, T. H. (1874). « On the Hypothesis that Animals Are Automata, and Its History ». *Nature*, 10, p. 362-366.
- Hyman, R. (2010). « Meta-Analysis that Conceals More than It Reveals: Comment on Storm et al. (2010) ». *Psychological Bulletin*, 136(4), p. 486-490.
- Jackson, F. (1982). « Epiphenomenal Qualia ». *The Philosophical Quarterly*, 32(127), p. 127-136.
- Jackson, F. (2003). « Mind and Illusion ». *Royal Institute of Philosophy Supplement*, 53, p. 251-271.
- Jacob, F. & Monod, J. (1961). « Genetic Regulatory Mechanisms in the Synthesis of Proteins ». *Journal of Molecular Biology*, 3(3), p. 318-356.
- Kim, J. (1998). *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Cambridge (MA) : MIT Press.
- Kripke, S. A. (1980). *Naming and Necessity*. Cambridge (MA) : Harvard University Press.
- La Mettrie, J. O. de (1747). *L'Homme Machine*. Leyde : Elie Luzac.

- Laplace, P.-S. (1799-1825). *Traité de mécanique céleste*. 5 vol. Paris.
- Leibniz, G. W. (1695). « Système nouveau de la nature et de la communication des substances ». *Journal des savants*.
- Leibniz, G. W. (1714). *Monadologie*.
- Levine, J. (1983). « Materialism and Qualia: The Explanatory Gap ». *Pacific Philosophical Quarterly*, 64, p. 354-361.
- Lewes, G. H. (1875). *Problems of Life and Mind, First Series*, vol. 2. London : Trübner.
- Lewis, D. (1972). « Psychophysical and Theoretical Identifications ». *Australasian Journal of Philosophy*, 50(3), p. 249-258.
- Loar, B. (1990/1997). « Phenomenal States ». *Philosophical Perspectives*, 4 (1990), p. 81-108 ; version révisée dans N. Block, O. Flanagan & G. Güzeldere (dir.), *The Nature of Consciousness* (1997), p. 597-616. Cambridge (MA) : MIT Press.
- Loeb, J. (1899). « On the Nature of the Process of Fertilization and the Artificial Production of Normal Larvae (Plutei) from the Unfertilized Eggs of the Sea Urchin ». *American Journal of Physiology*, 3, p. 135-138.
- Loeb, J. (1912). *The Mechanistic Conception of Life: Biological Essays*. Chicago : University of Chicago Press.
- Lucas, J. R. (1961). « Minds, Machines and Gödel ». *Philosophy*, 36(137), p. 112-127.
- Malebranche, N. (1674-1675). *De la recherche de la vérité*. Paris.
- Maxwell, J. C. (1873). « Molecules ». *Nature*, 8, p. 437-441.
- McGinn, C. (1989). « Can We Solve the Mind-Body Problem? ». *Mind*, 98(391), p. 349-366.
- Mill, J. S. (1843). *A System of Logic, Ratiocinative and Inductive*. London : John W. Parker.
- Monod, J. (1970). *Le Hasard et la Nécessité : essai sur la philosophie naturelle de la biologie moderne*. Paris : Éditions du Seuil.
- Morgan, C. L. (1923). *Emergent Evolution*. London : Williams & Norgate.
- Nagel, T. (1974). « What Is It Like to Be a Bat? ». *The Philosophical Review*, 83(4), p. 435-450.
- Neurath, O. (1931). « Physicalism: The Philosophy of the Viennese Circle ». *The Monist*, 41(4), p. 618-623.
- Newton, I. (1713). « General Scholium », dans *Philosophiae Naturalis Principia Mathematica*, 2e éd. Cambridge.
- Oppenheim, P. & Putnam, H. (1958). « Unity of Science as a Working Hypothesis », dans H. Feigl, M. Scriven & G. Maxwell (dir.), *Minnesota Studies in the Philosophy of Science*, vol. 2, p. 3-36. Minneapolis : University of Minnesota Press.
- Papineau, D. (2001). « The Rise of Physicalism », dans C. Gillett & B. Loewer (dir.), *Physicalism and Its Discontents*. Cambridge : Cambridge University Press.
- Parnia, S. et al. (2023). « AWAreness during RESuscitation – II: A Multi-Center Study of Consciousness and Awareness in Cardiac Arrest ». *Resuscitation*, 191, 109903.

- Pauly, P. J. (1987). *Controlling Life: Jacques Loeb and the Engineering Ideal in Biology*. New York : Oxford University Press.
- Penrose, R. (1994). *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford : Oxford University Press.
- Place, U. T. (1956). « Is Consciousness a Brain Process? ». *British Journal of Psychology*, 47, p. 44-50.
- Putnam, H. (1967). « Psychological Predicates », dans W. H. Capitan & D. D. Merrill (dir.), *Art, Mind, and Religion*, p. 37-48. Pittsburgh : University of Pittsburgh Press.
- Ryle, G. (1949). *The Concept of Mind*. London : Hutchinson.
- Sherrington, C. S. (1906). *The Integrative Action of the Nervous System*. New Haven : Yale University Press.
- Sherrington, C. S. (1940). *Man on His Nature*. Cambridge : Cambridge University Press.
- Smart, J. J. C. (1959). « Sensations and Brain Processes ». *The Philosophical Review*, 68(2), p. 141-156.
- Sperry, R. W. (1968). « Hemisphere Deconnection and Unity in Conscious Awareness ». *American Psychologist*, 23(10), p. 723-733.
- Spinoza, B. (1677). *Ethica ordine geometrico demonstrata* [Éthique].
- Stanford Institute for Human-Centered AI (HAI) (2026). *AI Index Report 2026*. Stanford University.
- Storm, L., Tressoldi, P. E. & Di Risio, L. (2010). « Meta-Analysis of Free-Response Studies, 1992-2008: Assessing the Noise Reduction Model in Parapsychology ». *Psychological Bulletin*, 136(4), p. 471-485.
- Tressoldi, P. E. & Storm, L. (2021). « Stage 2 Registered Report: Anomalous Perception in a Ganzfeld Condition – A Meta-Analysis of More than 40 Years Investigation ». F1000Research.
- Turing, A. M. (1936). « On Computable Numbers, with an Application to the Entscheidungsproblem ». *Proceedings of the London Mathematical Society*, 42, p. 230-265.
- Vila Valls, A. (2026). « Émergence, réduction et real patterns ». *The Naturalistic Outlook*, 19 juillet 2026. <https://thenaturalisticoutlook.com/2026/07/19/emergence-reduction-et-real-patterns/>
- Vila Valls, A. (2026). « Le naturalisme à l'épreuve du dilemme de Hempel ». *The Naturalistic Outlook*, 19 juillet 2026. <https://thenaturalisticoutlook.com/2026/07/19/le-naturalisme-a-lepreuve-du-dilemme-de-hempel/>
- Watson, J. B. (1913). « Psychology as the Behaviorist Views It ». *Psychological Review*, 20, p. 158-177.
- Watson, J. D. & Crick, F. H. C. (1953). « Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid ». *Nature*, 171, p. 737-738.
- Wilkins, M. H. F., Stokes, A. R. & Wilson, H. R. (1953). « Molecular Structure of Deoxypentose Nucleic Acids ». *Nature*, 171, p. 738-740.