

arbitre, mais la plupart ont admis que le potentiel de préparation de l'action était une cause cérébrale de l'action. Libet supposait que le signal électrique qu'il avait enregistré dans le cerveau des participants reflétait leur véritable décision d'agir, prise, à leur insu, par leur cerveau, bien avant qu'ils aient conscience de leur intention d'agir. Il tenait ce signal électrique pour l'indice de «la» cause authentique du mouvement de la main des participants. Or la cause de l'action doit être une motivation d'agir.

Libet supposait donc que la cause cérébrale ou la motivation authentique de leur action précède leur intention consciente d'agir et préempte l'efficacité causale de celle-ci dans le processus de leur action. Mais en 2012, Schurger, Sitt et Dehaene ont proposé une tout autre interprétation du potentiel de préparation de l'action enregistrée par Libet, fondée sur un sens différent du mot «décision».

Loin de refléter la contribution motivationnelle précoce du cerveau des participants à s'engager dans l'action, à leur insu, ce signal électrique représenterait le seuil atteint par l'activité intrinsèque des neurones du système moteur dont la tâche est de prédire le prochain événement moteur, en l'absence de toute stimulation externe. L'activité spontanée des neurones peut être conceptualisée comme une «prise de décision» ou un vote stochastique dont le but est de prédire un événement moteur, et non pas de motiver une action. Quand un nombre suffisant de neurones ont déchargé (comme dans une chute de dominos), l'assemblée neuronale se prononce: l'événement moteur est imminent. La prédiction neuronale ne cause pas l'action.

La majorité des êtres humains, dans la plupart des cultures, se croient libres de leur choix. On ne peut, comme Jean-Paul Sartre, se croire paradoxalement condamné à la liberté que si on croit au dogme du libre arbitre. Aucun fait scientifique attesté ne vient ni confirmer ni infirmer ce dogme. Les hommes sont-ils libres de se libérer de ce dogme? Il revient à la philosophie et aux sciences cognitives d'explorer cette question spéculative. »

Mais suffit-il d'avoir le choix pour satisfaire à l'idée du libre arbitre, ou faut-il que la décision elle-même ne soit pas déterminée?

Pierre Jacob: C'est une bonne question. Comme le linguiste Noam Chomsky l'a fait souvent remarquer, supposez qu'un nazi braque un revolver sur votre tempe

et vous ordonne de crier: «Heil Hitler!» Même si vous obtempérez pour survivre, vous avez le sentiment que vous auriez pu faire autrement. En un sens, vous n'étiez pas libre puisque vous étiez menacé de recevoir une balle de revolver dans le crâne. En un autre sens, vous étiez libre de prendre le risque de mourir.

Pour la majorité des gens, y compris des philosophes, le problème métaphysique du libre arbitre est le problème de savoir si la liberté de nos décisions est compatible ou incompatible avec le déterminisme. On dit souvent qu'un monde déterministe est un monde dans lequel nous ne sommes pas libres parce que tous les événements passés et à venir sont déterminés par des causes.

Pourquoi pensons-nous intuitivement que le libre arbitre est incompatible avec le déterminisme?

Pierre Jacob: Le philosophe Dan Dennett a fait une suggestion intéressante: le libre arbitre serait incompatible avec le déterminisme si dans un monde déterministe tous les événements passés et futurs étaient inévitables. Mais un monde déterministe n'est pas *ipso facto* un monde où tout ce qui advient était inévitable. À la différence du passé, nous ne connaissons pas le futur. En quel sens, se demande Dan Dennett, le futur peut-il être inévitable si nous ne le connaissons pas?

Prenons un événement passé, donc connu: en quel sens était-il inévitable? Pendant un match de football, l'intention du joueur qui tire un penalty est de marquer un but et celle du gardien de l'en empêcher. Chacun dispose d'une liberté de manœuvre qui dépend de son habileté. Il n'y a pas de raison de supposer que le succès avéré de l'un, et l'échec avéré de l'autre, mettent le déterminisme en échec. Mais en quel sens le succès avéré de l'un devrait-il être qualifié d'inévitable? Quelle que soit l'issue de l'événement, les choses auraient pu se passer différemment.

Par conséquent, c'est une erreur de croire que si un événement est le résultat d'une chaîne causale conforme au déterminisme, cet événement était inévitable. Comme se plaît à le souligner Dan Dennett, dans une perspective évolutionniste, tous les agents vivants humains ou non sont les descendants d'ancêtres qui ont survécu aux épreuves de la sélection naturelle en évitant des obstacles qui n'étaient donc pas inévitables. ■

PROPOS RECUEILLIS
PAR LOÏC MANGIN

BIBLIOGRAPHIE

G. HERBET ET AL., *Disrupting the right pars opercularis with electrical stimulation frees the song*, *J. of Neurosurgery*, vol. 123(6), pp. 1401-1404, 2015.

C. SOON ET AL., *Predicting free choices for abstract intentions*, *PNAS*, vol. 110(15), pp. 6217-6222, 2013.

C. FRITH, *Comment le cerveau crée notre univers mental*, Odile Jacob, 2010.

L'ESSENTIEL

● Selon l'hypothèse bayésienne, le cerveau réalise en permanence des prédictions qu'il confronte ensuite à la réalité des perceptions. Il met ainsi à jour les croyances qu'il a sur le monde.

● Pour faire face à la complexité de l'environnement, ces croyances sont organisées hiérarchiquement selon, par exemple, différents niveaux d'abstraction.

● Cet ensemble de croyances, souvent implicites, constituerait une forme d'inconscient qualifié de « bayésien ».

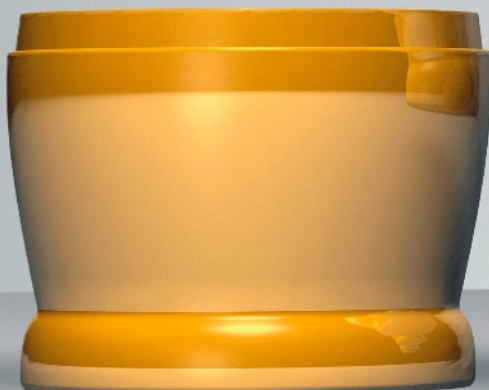
● Cet inconscient bayésien alimente à chaque instant la conscience d'une multitude de modèles hiérarchiques codant la structure causale, logique et temporelle du monde.

LES AUTEURS



HUGO BOTTEMANNE ET STÉPHANE MOUCHABAC sont psychiatres et travaillent sur la modélisation bayésienne en sciences cognitives à l'Institut du cerveau et à l'hôpital Saint-Antoine, à Paris.

Un inconscient bayésien ?



Une « révolution bayésienne » infuserait progressivement les sciences cognitives. Cette théorie fondée sur un célèbre théorème en statistique – le théorème de Bayes – bouleverse notre vision de l'inconscient. Celui-ci reposerait sur une architecture hiérarchique du cerveau faisant de cet organe une « machine à prédictions ».



D

eux urnes contiennent, l'une dix boules noires et trente blanches, l'autre vingt de chaque couleur. On tire une boule au hasard dans l'une des urnes, choisie aléatoirement. La boule est blanche. Quelle est la probabilité qu'elle vienne de la première urne? Pour répondre à cette question, un célèbre théorème attribué à Thomas Bayes au XVIII^e siècle (*voir l'encadré page 47*) est utile. Schématiquement, il permet d'actualiser une probabilité à partir des observations. Loin d'être anecdotique, ce théorème a des applications importantes, en médecine par exemple, pour analyser le résultat de tests en tenant compte des faux positifs, ou encore en neurosciences et en intelligence artificielle pour comprendre comment un humain ou un robot peuvent apprendre à partir d'informations incertaines.

UNE MACHINE PRÉDICTIVE

En effet, la théorie bayésienne pourrait changer en profondeur notre compréhension du fonctionnement cérébral. Au cœur de cette transformation se trouve l'hypothèse du cerveau bayésien, une nouvelle théorie en neurosciences. Les fondations de cette hypothèse remontent aux travaux d'Hermann von Helmholtz au milieu du XIX^e siècle. Étudiant la perception visuelle, le physiologiste prussien découvre que notre cerveau n'est pas un système passif de traitement des informations, mais plutôt un générateur d'inférences prédisant à chaque instant les entrées sensorielles. Pour Helmholtz, le cerveau serait ainsi une « machine prédictive » réalisant des prédictions qu'il qualifie d'« inférences inconscientes », et qui interféreraient avec nos processus conscients. Cette hypothèse s'est ensuite progressivement enrichie de nombreuses formulations mathématiques et de preuves expérimentales, aboutissant au principe du codage prédictif et à celui de l'énergie libre. De quoi s'agit-il?

La théorie du codage prédictif suggère que le cerveau élabore à chaque instant des modèles de son environnement, c'est-à-dire des « croyances », puis utilise ces modèles pour prédire les futures entrées sensorielles. ➤

➤ Lorsqu'un décalage entre les prédictions et les stimuli apparaît, le cerveau génère une erreur de prédiction, et utilise cette nouvelle information pour modifier ses croyances, et ainsi améliorer ses prédictions futures. En un sens, cette théorie suppose que notre cerveau «internalise» la structure causale du monde, c'est-à-dire son fonctionnement, pour prédire comment ces sensations sont engendrées.

Le principe de l'énergie libre, introduite au début des années 2000 par le neuroscientifique britannique Karl Friston, propose une formulation mathématique du codage prédictif. Il suppose que l'erreur de prédiction produit de l'incertitude pour le cerveau, et associe ce degré d'incertitude à l'énergie libre, un concept issu de la thermodynamique. Selon Karl Friston, la fonction principale du cerveau est alors de réduire cette énergie libre en minimisant les erreurs de prédiction. Pour ce faire, il peut modifier soit les modèles du monde pour les accorder avec les stimuli sensoriels, c'est-à-dire «mettre à jour» sa croyance, soit les entrées sensorielles en agissant différemment dans le monde, notamment en réalisant une action.

UN JAGUAR DANS LA JUNGLE, OU PAS?

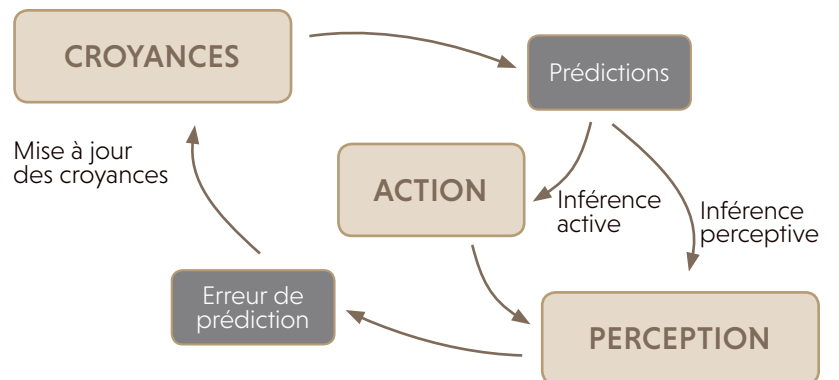
Un exemple. Dans la jungle amazonienne, au crépuscule, votre système visuel détecte une tache orangée fugace dans le feuillage: vous pouvez réduire l'incertitude associée à cette perception en générant une croyance à propos de votre environnement («Je crois qu'il y a un jaguar derrière les buissons») ou bien en agissant afin d'affiner vos connaissances sur l'environnement («Je me rapproche de la végétation pour voir ce qui s'y cache»). Dans les deux cas, l'incertitude associée à l'entrée sensorielle diminue. Dans cette relation bidirectionnelle, le monde fournit les données sensorielles qui constituent la base de l'inférence, et le cerveau agit sur le monde pour modifier le flux sensoriel. Le principe de minimisation de l'énergie libre constitue alors un processus unificateur, en mesure d'expliquer à la fois la perception, l'apprentissage et la prise de décision dans un monde incertain.

Le bayésianisme trouve ici toute son importance. En effet, on peut mettre en œuvre de façon «algorithmique» les principes du codage prédictif et de la minimisation de l'énergie libre en appliquant les préceptes de l'inférence bayésienne. Techniquement, celle-ci est un processus de détermination de la meilleure hypothèse expliquant les états observés. Dans le cerveau, elle pourrait permettre de renverser le problème en générant un modèle des causes des entrées sensorielles utilisé pour estimer ce qu'elles devraient être dans la réalité. On passe ainsi d'un simple processus inductif déduisant

la cause à partir de l'entrée sensorielle, peu flexible en situation d'incertitude, à une inférence de la cause de l'entrée sensorielle à partir de sa vraisemblance.

L'hypothèse du cerveau bayésien suppose ainsi que le cerveau peut déchiffrer ce qui provoque une entrée sensorielle en combinant à chaque instant les informations probabilistes issues de ses organes sensoriels avec ses prédictions sur les causes possibles de ces informations. Ce traitement en sens inverse, partant du résultat jusqu'à son origine, réduit l'ambiguïté inhérente aux informations sensorielles. Le cerveau peut alors discriminer plus finement les situations où les mêmes entrées sensorielles sont provoquées par des causes différentes, ou encore lorsque des causes différentes provoquent les mêmes entrées sensorielles.

L'inférence bayésienne rend également possible la modélisation de l'effet de l'incertitude sur la mise à jour des croyances après une erreur de prédiction. Ainsi, lorsque les informations que l'on perçoit sont incertaines, il vaut mieux conserver ses anciennes croyances



Dans un environnement donné, sur la base de croyances, notre cerveau établit des prédictions sur les entrées sensorielles possibles et programme des actions en fonction de celles-ci. Le décalage entre ses prédictions et la réalité génère des erreurs de prédictions qui sont utilisées pour mettre à jour les croyances.

UN PEU DE PROBABILITÉS

L'origine du bayésianisme se niche dans l'article posthume du révérend et mathématicien anglais Thomas Bayes (1702-1761), « An essay towards solving a problem in the doctrine of chances ». Cette œuvre publiée par son ami Richard Price en 1763 expose ce qui deviendra le « théorème de Bayes » :

$$P(A/B) = \frac{P(A) * P(B/A)}{P(A) * P(B/A) + P(\bar{A}) * P(B/\bar{A})}$$

Soit A et B deux événements, le théorème de Bayes permet de déterminer la probabilité de A sachant B c'est-à-dire $P(A/B)$, si l'on connaît les probabilités de A , de B et de B sachant A .

En d'autres termes, ce théorème décrit la méthode optimale pour mettre à jour une croyance dans des conditions d'incertitude. Il est largement utilisé en intelligence artificielle, en ingénierie comportementale, en neuroéconomie, en sciences sociales, en psychologie expérimentale, et en neurosciences cognitives pour décrire ou modéliser les comportements.

plutôt que de modifier tous ses modèles du monde. À l'inverse, lorsqu'on reçoit de nouvelles informations très précises qui contredisent ce que l'on pense, nos croyances peuvent évoluer pour s'adapter efficacement aux changements. On obtient *in fine* un système flexible s'adaptant à la structure statistique du monde au gré de nos expériences. Mais ce modèle simplifié reflète-t-il le fonctionnement réel de notre cerveau ?

UNE HIÉRARCHIE DES CROYANCES

Ce serait trop facile ! Notre monde est trop complexe pour un algorithme aussi simple. Les signaux sensoriels que nous percevons proviennent d'un environnement dynamique en constante évolution, avec une multiplicité de structures causales imbriquées. Pour en tenir compte, la théorie du cerveau bayésien suggère que notre système nerveux traite les différents signaux sensoriels simultanément de façon « hiérarchique ». Plutôt que de comparer une seule probabilité aux preuves sensorielles, notre cerveau manipule une hiérarchie de croyances à différents niveaux d'échelle d'implication spatiotemporelle, logique et d'abstraction.

Retournons dans la jungle. La génération de la croyance qu'un jaguar est caché sous le feuillage est issue de la conjonction d'une multitude de modèles statistiques imbriqués (la probabilité qu'un animal au pelage fauve dans

la jungle soit un jaguar, la probabilité qu'un animal se cachant avant de bondir soit un jaguar, la dangerosité de l'incertitude sur la présence d'un possible jaguar...). La génération d'inférences, c'est-à-dire d'hypothèses sur le monde, à partir d'une hiérarchie de prédiction avec un large grain temporel, spatial et causal aide à réduire l'ambiguïté massive des multiples entrées sensorielles, et à obtenir une représentation plus ou moins optimale de l'environnement. De simple processus neuro-cognitif, l'inférence bayésienne devient alors un puissant mécanisme adaptatif.

LA CLÉ DE VOÛTE DU TRAITEMENT DE L'INFORMATION

Aussi élégante soit-elle sur le plan théorique, la théorie bayésienne est-elle soutenue par l'organisation neuronale du cerveau ? Plusieurs travaux ont suggéré que les inférences seraient réalisées par un assemblage hiérarchique de connexions neuronales descendantes qui implémenteraient les prédictions, et de connexions ascendantes qui transmettraient les erreurs de prédiction. Ces prédictions et ces erreurs pourraient notamment être codées par des neurones particuliers associés à un vaste réseau de dendrites et de synapses, les cellules pyramidales.

Dans ce modèle prédictif hiérarchique, les prédictions descendantes depuis les niveaux élevés de la hiérarchie sont utilisées pour réduire les erreurs de prédiction des niveaux inférieurs. À chaque niveau, seules les erreurs de prédiction qui n'ont pas pu être expliquées par les prédictions sont transmises à l'étage supérieur, et peuvent alors être utilisées pour optimiser les croyances sur le monde.

Ces nouvelles prédictions sont ensuite transmises aux niveaux inférieurs par des connexions neuronales descendantes, pour générer de nouvelles inférences. De cette façon, toute nouvelle expérience inattendue est comparée aux connaissances antérieures à travers un vaste réseau hiérarchique associant différentes échelles temporelles, causales et spatiales. Ce processus de codage prédictif hiérarchique *medié* par la signalisation neuronale constituerait ainsi la clé de voûte de l'ensemble du traitement de l'information corticale.

Plusieurs travaux suggèrent d'ailleurs que la hiérarchie des croyances est corrélée à la complexité des représentations. Plus on monte, plus les représentations codées sont abstraites et globales, plus le niveau d'implication logique est important. Plus on descend, plus le niveau de détail se renforce, mais plus l'espace de prédiction diminue, jusqu'à arriver à des modèles prédictifs extrêmement précis. Les aires cérébrales de haut niveau dévolues aux représentations complexes envoient ainsi des signaux prédictifs aux aires sensorielles primaires, ➤

- déterminant en partie les motifs causaux et perceptifs attendus.

Les modèles causaux de haut niveau permettraient au cerveau de faire des prédictions sur une temporalité longue ou sur des associations causales complexes. À l'inverse, les modèles de bas niveau dans la hiérarchie prédisent, certes avec précision, la nature d'une entrée sensorielle simple, mais au mieux seulement quelques secondes à l'avance. Le cerveau ne peut déduire les conséquences de ses actions que s'il peut modéliser l'avenir. Cette capacité de faire des prédictions sur des échelles temporelles longues est ainsi primordiale pour définir les actions à réaliser. La structure hiérarchique lui permet de faire des prédictions sur des choses qui ne se sont jamais produites et pourraient ne jamais se produire : une forme de simulation d'un monde contrefactuel.

Pour le philosophe Jakob Hohwy, le cerveau serait ainsi un «miroir de la nature», simulant à chaque instant la structure causale du monde en fonction des prédictions de ses caractéristiques statistiques. Le codage prédictif et la théorie du cerveau bayésien pourraient ainsi offrir une conception globale de notre fonctionnement cérébral, avec une portée explicative aussi importante que celle de la psychanalyse au siècle précédent.

D'UNE CROYANCE À L'AUTRE

L'un des aspects les plus importants du réseau prédictif est la notion d'implication logique. Elle désigne le phénomène par lequel une croyance donnée peut en impliquer un ensemble d'autres. On peut l'illustrer avec la notion de transitivité du philosophe néerlandais Bas Van Fraassen : si je crois qu'un vase se casse lorsqu'il tombe sur le sol, alors je crois aussi qu'il se casse quand on le projette contre le mur, lorsqu'on le frappe avec un marteau... Dans ce modèle prédictif, la hiérarchie de représentations définit des relations complexes d'implication logique qui assurent une cohérence globale au sein de notre réseau de croyances. Ainsi, si je crois qu'un jaguar a quatre pattes, je crois aussi qu'il n'a pas cinq pattes, pas six pattes, et une infinité d'autres croyances associées. Même si je ne suis pas conscient de ces croyances implicites, elles existent du fait de la structure logique de mon réseau de croyances, constituant une forme d'inconscient «bayésien».

Ce processus est toutefois complexe à modéliser, car à propos d'une seule expérience ou d'une seule croyance, on peut souvent générer une infinité de relations d'implication logique. Notre cerveau doit, à chaque instant, faire le tri entre toutes les hypothèses envisageables à partir des expériences passées. Par exemple, si je crois que tous les objets en cuivre conduisent

l'électricité, je peux rationnellement croire que tous les futurs objets en cuivre que je rencontrerai conduiront l'électricité. En revanche, si je crois que tous les objets posés sur mon bureau ce matin peuvent conduire l'électricité, je ne dois pas pour autant me mettre à croire que tous les objets qui seront sur mon bureau au cours des prochains mois seront conducteurs ou que tout ce qui est conducteur se trouve sur mon bureau.

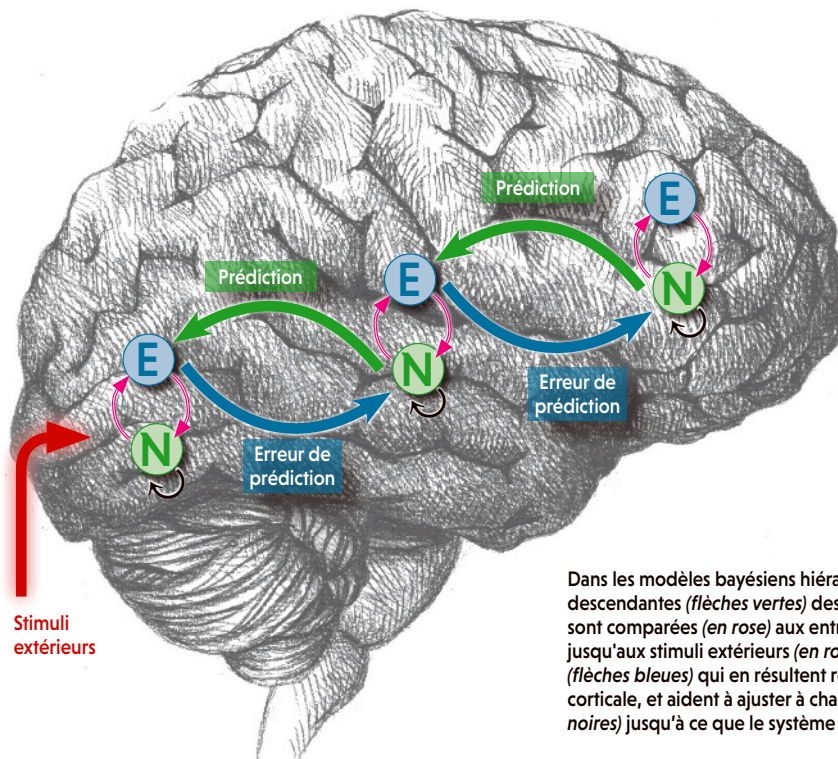
Notre cerveau doit ainsi représenter précisément le type d'implication logique entretenu par chacun des modèles internes qu'il entretient au sein de sa hiérarchie de croyances. Cette distinction est d'autant plus importante que la plupart de ces croyances ne sont pas conscientes et pourtant interagissent activement avec les perceptions, les pensées et les actions. La hiérarchie de croyances que nous avons détaillée représente ainsi une «matrice» déterminant la cognition, un modèle qui rappelle étrangement le concept d'inconscient mobilisé par la psychanalyse freudienne.

DE FREUD À BAYES

Lorsqu'il définit son triptyque conscient-inconscient-préconscient, Sigmund Freud s'inscrit dans le sillage d'illustres prédécesseurs (Thomas Laycock, Wilhelm Wundt...). Il postule que des processus cognitifs inconscients peuvent déterminer causalement des processus conscients et que des processus conscients peuvent devenir inconscients, c'est-à-dire que tout ce qui ne relève pas de la conscience à un moment donné reste silencieusement actif et peut continuer à interférer avec les processus conscients.

L'influence causale de l'inconscient sur le conscient a depuis été illustrée par de nombreuses preuves expérimentales (*voir Empêcher l'inconscient de dicter sa loi*, par J. Bargh, page 52). Cependant, une multitude de facteurs épistémologiques ont provoqué la scission entre les théories neurocognitives de l'inconscient et la théorie freudienne, favorisant l'émergence de concepts alternatifs comme celui d'«inconscient cognitif», dont la formulation bayésienne est l'une des nuances.

Malgré cette dissonance, un certain nombre de points d'achoppement ont été soulignés entre l'inconscient freudien et cognitif. Dans l'hypothèse du cerveau bayésien hiérarchique, l'inconscient est «structurel» au sens où il repose sur l'architecture fonctionnelle du cerveau, c'est-à-dire l'organisation et la connectivité des neurones et des synapses. Les représentations qui y sont codées n'ont pas vocation à devenir directement conscientes, mais elles participent néanmoins à la conscience à différents niveaux de la hiérarchie corticale. Elles alimentent à chaque instant la conscience d'une multitude de



Dans les modèles bayésiens hiérarchiques, les prédictions descendantes (flèches vertes) des niveaux corticaux (N) supérieurs sont comparées (en rose) aux entrées des niveaux inférieurs (E) jusqu'aux stimuli extérieurs (en rouge). Les erreurs de prédiction (flèches bleues) qui en résultent remontent dans la hiérarchie corticale, et aident à ajuster à chaque niveau les prédictions (flèches noires) jusqu'à ce que le système parvienne à les réduire.

prédictions hiérarchiques simulant la structure causale, logique et temporelle du monde. Elles influent alors directement sur la perception et la prise de décision, sans que l'on soit nécessairement conscient de cette influence. En miroir avec l'inconscient freudien, ces représentations inconscientes peuvent devenir conscientes au cours de l'inférence perceptive, et les représentations conscientes peuvent devenir inconscientes au cours de la mise à jour hiérarchique des croyances.

L'INCERTITUDE DU MONDE

Mais alors, quel est l'impact de la conscience sur ces processus inconscients encodés dans l'anatomie fonctionnelle du cortex ? À mesure que nous évoluons dans notre environnement, notre cerveau inscrit la structure statistique du monde dans son anatomie fonctionnelle, afin de simuler ensuite la charpente causale de son environnement. Les modifications de la connectivité synaptique nous permettent d'encoder cette expérience en modifiant nos croyances. Plutôt que de voir la conscience comme une instance de contrôle régulant les phénomènes inconscients, on peut alors l'imaginer comme un catalyseur des attentes prédictives et des entrées sensorielles, définissant la structure phénoménologique de nouvelles associations à encoder. Grâce à elle, en capitalisant sur l'héritage de notre passé et sur une simulation à long terme de notre avenir, nous minimisons l'incertitude du monde. ■

BIBLIOGRAPHIE

- J. HOHWY, Priors in perception: Top-down modulation, Bayesian perceptual learning rate, and prediction error minimization, *Conscious Cogn.*, vol. 47, pp. 75-85, 2017.
- J. HOHWY, *The Predictive Mind*, Oxford University Press, Oxford, 2013.
- S. KIEBEL ET AL., A hierarchy of time-scales and the brain, *PLoS Comput. Biol.*, vol. 4(11), art. e1000209, 2008.
- T. SHULTZ, The Bayesian revolution approaches psychological development, *Developmental Science*, vol. 10, pp. 357-364, 2007.
- D. KNILL ET A. POUGET, The Bayesian brain: The role of uncertainty in neural coding and computation, *Trends in Neurosciences*, vol. 27(12), pp. 712-719, 2004.

