

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?

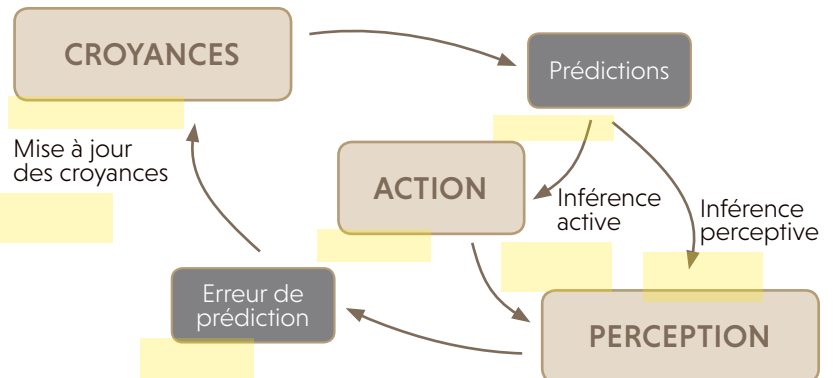
Un exemple. Dans la jungle amazonienne, au crépuscule, votre système visuel détecte une tache orangée fugace dans le feuillage: vous pouvez réduire l'incertitude associée à cette perception en générant une croyance à propos de votre environnement («Je crois qu'il y a un jaguar derrière les buissons») 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 *médié* 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, ➤