Les relations difficiles entre l'Intelligence Artificielle et les Neurosciences

Frédéric Alexandre

Les relations difficiles entre l'Intelligence Artificielle et les Neurosciences

Frédéric Alexandre, INRIA Bordeaux – Institut des Maladies Neurodégénératives - Labri

L'Intelligence Artificielle (IA) s'est construite sur une opposition forte entre connaissances et données. Les neurosciences ont tout d'abord fourni des éléments confortant cette vision avec la description de deux formes de mémoire, traitant respectivement de connaissances et de données. Les neurosciences ont ensuite décrit des interactions fortes entre ces deux formes de mémoire et ont suggéré que ces interactions permettent une cognition plus robuste et plus performante. De son coté, l'IA a pâti des limitations résultant de la dualité stricte entre connaissances et données. Pour autant, les chercheurs en IA restent trop souvent bloqués sur ces conceptions initiales et peinent à intégrer les mécanismes suggérés par les neurosciences. Ils se privent ainsi de pistes d'évolution prometteuses et d'un dialogue fertile avec ce domaine.

IA symbolique et numérique

L'Intelligence Artificielle (IA) cherche à définir des méthodes de traitement de l'information capables de rendre compte de caractéristiques de l'intelligence humaine. La recherche en IA s'est construite sur la base d'une polarité entre deux approches exclusives, proposant d'une part une IA dite symbolique car centrée sur la manipulation de connaissances et d'autre part une IA dite numérique car centrée sur la manipulation de données.

Cette polarité fut déclarée dès les origines de l'IA. D'un coté, certains de ses pères fondateurs comme H. Newell, A. Simon ou J. McCarthy soulignaient que, tout comme notre esprit, les ordinateurs manipulent des symboles et peuvent donc construire et manipuler des représentations du monde caractérisques de l'intelligence que l'on appelle connaissances. D'autres pères fondateurs comme J. von Neumann ou N. Wiener proposaient d'aborder l'intelligence en modélisant le cerveau et le calcul numérique de ses neurones pour montrer que des fonctions intelligentes peuvent résulter du traitement numérique de données. Cette dualité est très bien décrite par l'expression des frères Dreyfus « Making a Mind versus Modelling the Brain » (pour obtenir un système intelligent à l'image de l'humain, certains cherchent à créer un esprit logique, d'autres à modéliser le cerveau calculant), dans un article (Dreyfus & Dreyfus, 1991) où ils expliquent que, par leur construction même, ces deux paradigmes de l'intelligence sont faits pour s'opposer. L'approche symbolique met l'accent sur la résolution de problèmes et utilise la logique pour manipuler séquentiellement des symboles. L'approche numérique se focalise sur l'apprentissage et utilise les statistiques pour traiter massivement des données. C'est bien ce qui s'est produit : ces deux approches se sont développées en antagonisme.

En effet, tout au long de l'histoire de l'IA, chaque approche a essayé de s'imposer à l'occasion du succès d'une technique particulière (par exemple les systèmes experts pour l'approche symbolique ou les perceptrons pour l'approche numérique), qui a toujours été suivi de désillusions lorsque, pour chaque technique, des limitations sont apparues, entraînant ce que l'on a appelé des hivers de l'IA. La phrase d'un des frères Dreyfus : « Le domaine de l'intelligence artificielle présente un schéma récurrent : succès précoce et spectaculaire suivi de difficultés soudaines et inattendues. » date de 1965 mais s'est vérifiée ensuite à plusieurs reprises. Aujourd'hui, en dépit de ces revers, l'IA a fait des progrès indéniables, mais nous subissons toujours la dualité et l'alternance entre les approches symboliques et numériques,

même si le vocabulaire a un peu évolué et que l'on parle maintenant d'IA basée sur les connaissances (utilisée pour le web sémantique) ou basée sur les données (à la base des *data sciences*).

Nous sommes actuellement dans une période où c'est l'approche numérique qui domine et la technique phare est l'apprentissage de réseaux à couches profonds avec le Deep Learning. Elle rapporte des performances impressionnantes car l'accès large aux données et à la puissance de calcul permet aujourd'hui de nourrir ces architectures profondes. Mais des voix commencent à s'élever pour prédire un nouveau déclin proche si l'on n'est pas capable d'associer ces modèles numériques à des techniques d'interprétabilité (Lipton, 2017), permettant transparence (quelles informations extraient les couches cachées ?) et explications (sur quels critères fondent-elles leurs décisions ?), deux notions du monde des connaissances.

Sommes-nous encore partis pour un cycle d'alternance, à toujours nous demander laquelle de ces deux approches, à force de s'améliorer, finira par démontrer qu'elle était la bonne solution, ou saurons-nous sortir du cadre et trouver une solution plus élaborée qu'un choix exclusif entre l'une de ces deux approches ? C'est dans cette dernière perspective que je propose de revenir aux fondamentaux. Puisque les deux approches s'accordent au moins sur le fait qu'elles cherchent à reproduire nos fonctions cognitives dites intelligentes, ne devrait-on pas commencer par se demander si notre cognition est numérique ou symbolique ?

Mémoires implicites et explicites dans le cerveau

En fait, à cette question, les Sciences Cognitives nous répondent tout d'abord « les deux ». Bien sûr, nous sommes basés sur des connaissances car notre cognition a cette capacité unique de transformer un monde continu et incertain en des symboles discrets et de la logique, tel qu'on le constate en faisant un peu d'introspection. Mais nous sommes aussi basés sur le traitement numérique de données, car il a été montré (Saffran, 1999) que, même pour une fonction aussi symbolique que le langage, l'enfant utilise de l'apprentissage statistique pour extraire les mots des phrases qu'il perçoit.

Cette double compétence repose sur une dualité observée en clinique dans les années 1950 et théorisée dans les années 1990 (Squire, 2004) : notre mémoire à long terme est parfois explicite (on dit aussi déclarative) quand elle nous permet de manipuler explicitement des connaissances. Elle est parfois implicite (non déclarative) quand elle nous permet d'acquérir une compétence à partir d'expériences multiples (assimilables à des données). D'une part, comme exemple de mémoire explicite, nous pouvons nous souvenir explicitement de notre repas d'hier soir (mémoire épisodique) ou avoir la connaissance que le ciel est bleu (mémoire sémantique), deux types de mémoire que l'on sait déclarer; d'autre part, comme exemple de mémoire implicite, nous avons appris implicitement (par la pratique) notre langue maternelle et nous pouvons apprendre à faire du vélo (mémoire procédurale, non déclarative ; on sait le faire).

Certains pans de notre mémoire manipulent explicitement des connaissances alors que d'autres s'entrainent et se nourrissent de données. Nous savons que (et nous en sommes conscients, nous savons l'expliquer ou le déclarer) ou nous savons faire (et nous pouvons en faire la démonstration, sans être capable de ramener cette connaissance au niveau conscient ; c'est peut-être d'ailleurs pour cette raison que l'introspection a plutôt tendance à nous faire croire basés sur les connaissances plutôt que sur les données).

Il est également intéressant de mentionner que les neurosciences ont pu identifier des circuits cérébraux spécifiques et distincts pour ces deux types de mémoire, soulignant bien leur différence fondamentale, avec en particulier les circuits neuronaux reliant les ganglions de la base et le cortex plutôt impliqués dans la mémoire implicite, alors que l'hippocampe et ses relations avec l'ensemble du lobe temporal médial, sont essentiels pour la mémoire explicite (le cas clinique fondateur des années 1950 concernait un patient qui, après une ablation de l'hippocampe, avait perdu la fonction de mémoire explicite).

Dans un premier temps, la distinction entre ces différentes formes de mémoires a pu conforter le positionnement bi-polaire de l'IA qui a en particulier utilisé l'opposition implicite/explicite ou encore procédural/déclaratif pour justifier le développement de modèles capables d'apprendre une procédure à partir de données (comme des réseaux de neurones) ou de modèles capables de produire des connaissances à partir de règles (comme les systèmes experts). On peut aussi noter que ces deux écoles distinctes de l'IA se sont développées en antagonisme pour explorer chacune de ces pistes alors que les neurosciences indiquent que le cerveau abrite ces deux formes de mémoires. Certains travaux ont été menés pour associer des approches numériques et symboliques dans des structures hybrides (Sun & Alexandre, 2013), mais ils sont restés l'exception et n'abordaient le sujet que de façon superficielle.

Par ailleurs, et de façon plus notable, des travaux ultérieurs en neurosciences indiquent que, au-delà d'une simple cohabitation entre ces formes de mémoire, le cerveau organise en fait des interactions et des échanges complexes entre ces mémoires. Ces travaux montrent également que ces interactions sont à la base de certaines propriétés importantes de notre cognition, qui ne seraient pas forcément présentes si on considérait ces mémoires comme étanches. C'est le cas en particulier pour deux phénomènes que nous décrivons maintenant : la consolidation et la formation des habitudes.

Les mécanismes de la consolidation

Concernant la consolidation, le constat initial est celui d'une mise en œuvre différente par le cerveau de ces mémoires complémentaires (McClelland et al., 1995), avec un apprentissage lent de la mémoire procédurale dans le cortex et la formation rapide de la mémoire épisodique dans l'hippocampe. Prenons un exemple : comme je vais toujours faire mes achats dans le même supermarché, je vais former, après de nombreuses visites, la procédure pour accéder à son parking, mais à chaque visite, je dois aussi me souvenir de l'endroit précis où j'ai laissé ma voiture. Les modèles computationnels d'apprentissage permettent de mieux comprendre ce qui est à l'œuvre ici. Les modèles d'apprentissage procédural implicite doivent extraire statistiquement des régularités. Pour ce faire, on utilise généralement des réseaux de neurones organisés en couches successives qui ont cette capacité (et on remarque que ce sont aussi des couches successives de neurones qui assurent cette fonction dans le cerveau). Il leur faut pour cela de nombreux exemples dont les représentations doivent se recouvrir pour permettre des généralisations. Ce grand nombre d'exemples à traiter explique que le temps d'apprentissage va être long. Si on cherche à apprendre plus vite ou si on choisit tout à coup des exemples très différents, on va observer ce qu'on appelle un oubli catastrophique, c'est à dire qu'on oubliera les premiers exemples et on ne retiendra que les derniers, ce qui correspond à un apprentissage implicite de mauvaise qualité.

Inversement, dans un modèle d'apprentissage explicite de cas particuliers, on va chercher ce qui est spécifique plutôt que ce qui est régulier dans l'information (ce qui me sera utile pour retrouver ma voiture : je ne dois surtout pas généraliser sur plusieurs exemples mais me

souvenir de ce qui distingue ce cas précis). Ces modèles vont être généralement réalisés avec des réseaux de neurones récurrents (réseaux incluant des connexions entre tous les neurones et pas seulement de couches en couches) qui ont cette capacité de mémorisation de configurations (et on remarquera que les circuits correspondants du cerveau, dans l'hippocampe, sont fortement récurrents). On va aussi privilégier ici un modèle de codage dit clairsemé (aussi présent dans l'hippocampe), qui essaie de représenter (et de mémoriser) ce qui est distinctif. Cet apprentissage peut alors être beaucoup plus rapide, puisqu'on ne cherchera pas à se confronter à d'autres exemples mais qu'on apprendra plutôt par cœur un cas particulier. Par contre, l'expérimentation avec ce type de modèles montre des risques d'interférence si on apprend trop d'exemples proches (on risque de les mélanger) ainsi qu'un coût élevé pour le stockage des informations (d'une part, il y a une densité de connexion plus importante dans un réseau récurrent que dans un réseau à couche ; d'autre part, le codage doit être précis dans le réseau récurrent car on veut retrouver l'exemple précis qui a été mémorisé alors que le codage du réseau à couches peut supporter la généralisation). Il est donc impératif de limiter le nombre d'exemples stockés en mémoire épisodique, comme on le constate en neurosciences : la mémoire n'est pas permanente dans l'hippocampe, pouvant durer jusqu'à quelques années chez l'humain et quelques semaines chez le rat.

Le constat initial est donc celui d'apprentissages complémentaires, avec la mémoire implicite apprenant la structure cachée des données et la mémoire explicite apprenant des cas individuels. Des observations essentielles en neurosciences concernent les relations et les échanges entre ces mémoires, avec en particulier des transferts de l'hippocampe vers le cortex, que l'on appelle consolidation. Lors de périodes de repos et en particulier pendant le sommeil, l'hippocampe va reformer des souvenirs de cas particuliers et va les renvoyer vers le cortex. Ceci va permettre au cortex de s'entrainer de façon progressive, en alternant cas anciens et cas nouveaux et va donc lui éviter le phénomène d'oubli catastrophique mentionné plus haut. Une autre propriété intéressante de la consolidation est discutée en neurosciences. Lorsque les cas particuliers sont trop nombreux à stocker dans la mémoire dite épisodique de l'hippocampe, ce mécanisme de transfert vers le cortex va permettre de chercher leurs points communs pour former une nouvelle mémoire, sémantique, également explicite mais se focalisant sur les dimensions importantes. Il semble que cela permettra à l'hippocampe d'oublier les cas particuliers correspondants. En résumé, la consolidation se décrit comme un échange dynamique entre les mémoires implicites et explicites pour avoir une cognition efficace, tirant le meilleur profit des deux mécanismes et les soulageant de leurs faiblesses. On verra plus loin que ces résultats n'ont pas été transposés en IA.

Les mécanismes de la formation des habitudes

Concernant les mécanismes de formation des habitudes, le constat initial est celui de deux modes de prise de décision, nommés réflexif et réflectif (Dolan & Dayan, 2013) parce que reposant respectivement sur des réflexes ou sur de la réflexion. Historiquement, ils ont été proposés respectivement par les behavioristes, pour qui le comportement émergeait implicitement d'un ensemble d'associations réflexes entre stimulus et réponses, et par les cognitivistes, qui imaginaient plutôt la construction de représentations internes (des cartes cognitives) pour manipuler explicitement des connaissances lors de phases de réflexion. Là aussi, les études ultérieures en neurosciences ont montré que nous développions ces deux modes de décision grâce aux deux types d'apprentissage décrits ici, implicite et explicite. En première approximation, on explique généralement que pour prendre une décision, on commence par se construire une représentation du monde qui nous permettra de façon prospective, d'anticiper les conséquences que pourraient avoir nos actions et de choisir l'action dont les conséquences

seront les plus intéressantes ou les plus proches des buts que nous poursuivons. Les neurosciences rapportent que, avec sa capacité à former rapidement des associations arbitraires, l'hippocampe semble impliqué dans la construction de ces cartes cognitives explicites.

Ensuite, après avoir longuement utilisé cette approche dite dirigée par les buts (où la représentation explicite des buts est nécessaire), on peut finir par dégager des régularités et se rendre compte, uniquement par une analyse rétrospective portant sur de nombreux cas, que dans une certaine situation, c'est toujours la sélection d'une certaine action qui se révèle la plus intéressante, et on peut donc former une association situation-action (ou stimulus-réponse) dans le cortex, par cet apprentissage lent, sans avoir à se représenter explicitement le but qui motive ce choix. On appelle cela la formation des habitudes.

On retrouve ici les principes évoqués précédemment, avec la mémoire explicite qui se forme en premier dans l'hippocampe, avec un coût cognitif important, pour stocker et manipuler explicitement des connaissances et la mémoire implicite qui se forme ensuite dans le cortex, avec l'aide de la précédente, en construisant sur la base de statistiques, des habitudes plus longues à former, qui sont des automatismes, mais plus rapides à déclencher et moins sujets aux interférences.

Mais que fait l'IA ?

L'IA s'est emparée de ces travaux pour assimiler cette dualité implicite/explicite aux aspects numériques/symboliques ou encore basés sur les données et sur les connaissances. Pour des tâches de classification et d'apprentissage par renforcement, elle a proposé des modèles opérationnels pour la catégorisation et pour la décision ayant ces deux types de caractéristiques. Elle n'a cependant pas intégré l'ensemble des résultats que nous venons d'évoquer, qui montrent que, bien au-delà d'une simple dualité, les mémoires implicites et explicites interagissent très subtilement pour former notre cognition. Essayer de s'approprier ces mécanismes pourrait fournir des modèles d'IA plus performants, mais aussi pourrait fournir une vision normative aux neurosciences, qui n'ont pas complètement élucidé toutes les questions que posent ces observations.

D'une part, concernant la consolidation, il est important de mentionner que l'hippocampe est en fait alimenté presque exclusivement par des représentations provenant du cortex, donc correspondant à l'état courant de la mémoire implicite, ce qui indique que ces deux mémoires sont en fait interdépendantes et co-construites. Comment ces échanges se réalisent entre le cortex et l'hippocampe et comment ils évoluent mutuellement restent des mécanismes très peu décrits et très peu connus en neurosciences.

D'autre part, concernant la formation des habitudes, il est clair aussi que cette automatisation de notre comportement n'est pas à sens unique et que nous savons, en fonction de l'évolution des contingences du monde, figer un comportement puis le réviser par une remise en cause explicite quand il n'est plus efficace puis le reprendre quand c'est pertinent. Là aussi, ces mécanismes sont encore très peu compris en neurosciences. On ne sait pas expliquer précisément comment nous construisons notre modèle du monde, comment nous choisissons d'en automatiser une partie, comment cette automatisation s'effectue et comment nous gardons un niveau de contrôle qui nous permet de remettre en cause cette distribution entre parties du comportement contrôlées ou automatisées.

Comprendre les principes de transfert et d'association entre connaissances et données devrait être au cœur des préoccupations d'une IA soucieuse de résoudre ses points de blocage et d'offrir des modèles plus puissants. Par ailleurs, comme nous l'avons évoqué plus haut, la modélisation a toujours été une source d'inspiration pour aider les neurosciences à formaliser et à décrire les mécanismes de traitement de l'information à l'œuvre dans notre cerveau. Pourtant, concernant ces modalités d'associations flexibles entre nos mémoires implicites et explicites, l'IA ne joue pas son rôle d'aiguillon pour aider les neurosciences à avancer sur ces questions, car elle reste bloquée sur cette dualité rigide et stérile entre les données et les connaissances. Il est donc temps de demander à l'IA de s'emparer de ces résultats des neurosciences pour progresser mais aussi pour renvoyer vers les neurosciences son analyse normative.

R. Sun and F. Alexandre, eds. (2013) *Connectionist-Symbolic Integration : from Unified to Hybrid Approaches.* Taylor and Francis. https://www.taylorfrancis.com/books/9780203763667

Dreyfus H.L., Dreyfus S.E. (1991) Making a Mind Versus Modelling the Brain: Artificial Intelligence Back at the Branchpoint. In: Negrotti M. (eds) *Understanding the Artificial: On the Future Shape of Artificial Intelligence. Artificial Intelligence and Society*. Springer, London.
https://link.springer.com/chapter/10.1007/978-1-4471-1776-6_3

Lipton, Z. C. (2017). *The Mythos of Model Interpretability*. http://arxiv.org/abs/1606.03490

Saffran, J.R. et al. (1999) Statistical learning of tone sequences by human infants and adults. *Cognition* 70, 27–52.

Squire, L. R. (2004). Memory systems of the brain : a brief history and current perspective. *Neurobiology of Learning and Memory*, 82, 171–177.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457.

Dolan, R. J., & Dayan, P. (2013). Goals and Habits in the Brain. *Neuron*, *80*(2), 312–325. https://doi.org/10.1016/j.neuron.2013.09.007