Calcul algorithmique, responsabilité humaine

Raja Chatila

Calcul algorithmique, responsabilité humaine

Raja Chatila
Institut des Systèmes Intelligents et de Robotique
Sorbonne Université, Paris

Proposition pour l'ouvrage « Penser, calculer, délibérer »

Table des matières

Question traitée dans cet article

Les méthodes d'Intelligence Artificielles telles que définies scientifiquement peuvent-elle être considérées responsables et comptables de leurs résultats ? Où se situe la responsabilité humaine ?

L'Intelligence Artificielle

Les systèmes basés sur les techniques d'Intelligence Artificielle (IA) ont atteint des performances impressionnantes ces dernières années, et ont été adoptés dans de nombreux secteurs industriels ou de service.

Clarifions tout d'abord l'objet de notre discours. Qu'est-ce que l'Intelligence Artificielle ? Il s'agit d'abord, et surtout, d'un programme de recherche, énoncé par ses fondateurs en 1955 comme étant « *basé sur la conjecture que tout aspect de l'apprentissage, ainsi que tout autre trait de l'intelligence, peut en principe être décrit de manière tellement précise qu'une machine peut être construite pour le simuler. Une tentative sera faite pour trouver comment les machines peuvent utiliser le langage, former des abstractions et des concepts, résoudre toute sorte de problèmes actuellement du ressort des humains, et s'améliorer.* »

Ce programme de recherche est toujours d'actualité. Il a donné naissance à de nombreux travaux qui ont contribué – au-delà (ou en-deçà…) de l'objectif recherché de l'intelligence de la machine lui-même – à rendre l'ordinateur de plus en plus performant et facile d'usage. Des recherches portant sur la représentation des connaissances dans la machine ont par exemple produit des langages de programmation pouvant exprimer des relations entre objets et classes d'objets, et à manipuler ces relations par des calculs. Ceci a abouti à la possibilité de créer des bases de données qui peuvent organiser des

informations diverses, et qui peuvent être consultées pour rechercher des informations spécifiques. Ces outils sont utilisés quotidiennement.

Il a également été possible de formaliser le raisonnement déductif logique pour inférer la valeur de vérité de propositions à partir de prémisses, et ainsi démonter des théorèmes à partir d'axiomes.

Des méthodes de calcul ont été élaborées qui permettent de parcourir efficacement des structures de données exprimant des relations de connectivité entre « états ». Un état décrit une situation donnée, par exemple les positions des pièces sur un échiquier, ou les coordonnées d'une voiture sur le réseau routier. Ces méthodes dites de « recherche » fournissent des solutions, des chemins optimaux, pour passer d'un état donné à un autre. Le GPS d'une voiture ou d'un smartphone utilise de tels algorithmes pour proposer un chemin que le conducteur ou le piéton décidera d'emprunter pour rejoindre sa destination. Un programme de jeu d'échec utilise un algorithme tout à fait similaire pour jouer – et très souvent gagner la partie.

Nous venons d'utiliser le terme « algorithme ». Admettons pour l'instant qu'il s'agit d'une méthode de calcul. Nous en verrons la signification précise dans la prochaine section.

Des méthodes dites d'apprentissage machine ont été élaborées dès la naissance du domaine de l'IA. Elles consistent à améliorer le comportement, le calcul, effectué par la machine sur la base de l'expérience, c'est-à-dire en évaluant les résultats obtenus par l'application de ce calcul aux étapes précédentes. Ainsi le programme de jeu de dames élaboré par Arthur Samuel dès 1957 pouvait améliorer ses performances en mieux choisissant ses mouvements à la lumière des gains obtenus ou non au cours des parties précédentes.

Mais ce qui a fait de l'IA un tel succès depuis une dizaine d'années, c'est une technique particulière de l'apprentissage machine inventée dans les années 1980-1990 (Le Cun 2019), *l'apprentissage profond*. Ces techniques d'apprentissage sont essentiellement fondées sur des méthodes statistiques utilisant des réseaux de neurones formels organisés en couches.

La formalisation mathématique des fonctions de la cellule biologique neuronale a été faite en 1943 par McCulloch et Pitts. Ce modèle formel, qui ne rend pas compte de toute la complexité de ce qui se déroule dans cette cellule, explique son fonctionnement comme une transformation des signaux d'entrée en provenance d'autres neurones ou de capteurs. Un neurone calcule une somme pondérée de ses signaux d'entrée, suivie d'une opération mathématique particulière, par exemple une fonction sigmoïde (le lecteur intéressé par le détail de ce fonctionnement peut consulter par exemple (LeCun 2019)). Le résultat est un signal transmis à d'autres neurones.

Les signaux traversent les neurones organisés en réseaux et se combinent atteignant des neurones particuliers en sortie. Chez les êtres vivants, certains neurones ainsi activés peuvent transmettre un signal électrique excitant un muscle, déclenchant une action.

Mais ces réseaux peuvent aussi reconnaître la présence d'un objet particulier dans une image.

Ainsi des neurones combinant des signaux en provenance d'autres neurones émettront un signal corrélant ou reconnaissant la présence simultanée d'éléments dont le rôle sera de signaler la reconnaissance d'un objet particulier.

La propriété essentielle des réseaux de neurones est que les pondérations des signaux se modifient dans un processus dit d'entraînement, ou d'apprentissage, qui consiste à modifier leurs valeurs (appelés poids synaptiques en référence aux connexions entrantes des neurones biologiques) pour améliorer la reconnaissance des objets d'intérêt.

Dans les réseaux de neurones artificiels, cet apprentissage est effectué sur un très grand nombre de données contenant les objets en question. Pour résumer très sommairement, ces réseaux effectuent des opérations mathématiques sur les données afin d'en extraire des caractéristiques élémentaires communes qui sont ensuite intégrées pour reconnaître des formes. Ils peuvent ainsi détecter des régularités qui permettent de regrouper des données similaires en catégories différentes. L'ajustement des poids synaptiques est guidé par des algorithmes d'optimisation. Notons que ce processus est celui d'une corrélation statistique entre les caractéristiques contenues dans les données, et non celui de la propagation d'un raisonnement causal.

Il existe deux grandes catégories de méthodes : les méthodes supervisées, dans lesquelles les classes d'intérêt sont connues et l'optimisation orientée pour classer les données dans l'une ou l'autre des catégories attendues, et les méthodes non supervisées recherchant des régularités et regroupant les données en diverses classes qui les contiennent. Le résultat de ces processus est un « modèle » exprimant les informations que le système a extraites des données d'entrée. L'intérêt de ces systèmes est qu'ils évitent de rechercher des caractéristiques prédéfinies dans les données pour effectuer des classements ou une reconnaissance du contenu. La clé est le traitement statistique des composantes élémentaires présentes dans les données. Et compte-tenu de leur démarche statistique, pour être capables de fournir des résultats pertinents, toutes ces méthodes ont besoin de grandes quantités de données.

D'autres méthodes, dites d'apprentissage par renforcement sont fondées sur des processus séquentiels qui sélectionnent les choix les plus pertinents à une étape donnée en fonction d'un critère mathématique, et réitèrent ces choix jusqu'à obtenir les valeurs les plus élevées pour ce critère. Il s'agit ainsi d'une sorte d'optimisation par « essai-erreur ». De telles méthodes ont permis par exemple d'améliorer les performances du logiciel AlphaGo puis AlphaZero dans le jeu de Go.

Les méthodes d'apprentissage susmentionnées sont celles qui sont communément utilisées aujourd'hui. Elles s'appliquent dès lors que l'on dispose de grandes quantités de données sur un domaine d'intérêt. La numérisation et la diffusion de données massives sous différentes formes, images, signaux, textes, sons, dans quasiment tous les domaines a été l'élément qui a favorisé leur succès, conjugué à l'accroissement permanent de la vitesse de calcul.

Mais la force de ces méthodes est aussi leur faiblesse. La démarche usuelle de leur développement dans un domaine d'application, par exemple l'imagerie médicale, est d'utiliser un ensemble de données pour la phase d'entraînement et d'autres ensembles de données pour une phase de validation et de test, et de réitérer l'ajustement des paramètres déterminant le fonctionnement du réseau jusqu'à atteindre une précision satisfaisante du résultat. En d'autres termes, le système a été développé sur un échantillon et testé sur un autre, mais rien n'assure qu'il fonctionnera correctement sur des données tout à fait différentes de l'un et de l'autre. Ces échantillons devraient en effet être statistiquement corrects, représentatifs de la réalité, sans biais et la précision atteinte par le système être acceptable et compatible avec les enjeux de la mise en œuvre. Il y a là une première question centrale liée à la robustesse technique de ces systèmes, ainsi qu'une problématique spécifique liée au biais qui dépasse les questions techniques. Nous reviendrons sur ces questions.

L'Intelligence et le calcul

Un point, essentiel, est que ces méthodes effectuent en réalité des traitements statistiques en corrélant des éléments des données qui leurs sont fournies, sans aucune recherche de causalité. Un algorithme est une séquence de traitements, de calculs élémentaires sur des données fournies, explicitement définis et organisés par un concepteur pour obtenir un résultat qui peut être une classification de ces données. Ils se traduisent par des programmes exécutés par les ordinateurs. Les méthodes basées sur l'apprentissage évitent cette programmation explicite et la remplacent par l'utilisation de modèles statistiques obtenus par apprentissage, modèles auxquels on va ensuite confronter de nouvelles données pour les classifier ou y reconnaître des similarités avec ce qui été appris et « prédire » ce résultat.

Dans tous les cas, il faut bien retenir que ces données n'ont aucun sens pour la machine. Il s'agit toujours de valeurs fournies ou bien mesurées par des capteurs (par exemple des images). Leur sémantique, leur signification dans la réalité du monde, n'est connue que par le concepteur humain. Le résultat du traitement, dont le sens est toujours interprété par l'utilisateur du système, est issu dans un cas d'une séquence définie par le programmeur (algorithme explicite), et dans l'autre par la comparaison des composantes des nouvelles données avec le modèle implicite issu de l'échantillon d'apprentissage.

En d'autres termes, la raison d'être du système utilisé et la compréhension de ce qu'il produit est exclusivement du ressort de l'être humain. En revanche, les calculs qui s'effectuent dans le système peuvent n'avoir aucune intelligibilité pour l'utilisateur, voire pour le concepteur dans le cas des systèmes à apprentissage, compte tenu du grand nombre de paramètres et des calculs les combinant. A cela s'ajoute la non connaissance par le système d'intelligence artificielle du contexte exact de son utilisation.

C'est dans ce sens que nous affirmons qu'un système d'intelligence artificielle n'a en réalité aucune intelligence.

Délibération éthique et calcul

Le développement des véhicules à conduite automatisée – souvent et à tort appelés véhicules autonomes – soulève de nombreuses questions éthiques et sociétales que nous ne traiterons pas ici (voir CNPEN 2021)). Nous nous focaliserons sur la question du dilemme éthique qui a motivé les travaux de nombreux auteurs et chercheurs. Ce dilemme – dit dilemme du trolley - est à l'origine une expérience de pensée proposée par Philippa Foot en 1967 à propos des choix éthiques humains (Foot 2002). Vous êtes témoin de la scène suivante : un tram (trolley) a perdu ses freins et avance à toute vitesse. Plus loin sur la voie se trouvent cinq personnes inconscientes du danger qui les menace. Entre le tram et ces personnes se trouve un aiguillage qui, si actionné, pourrait envoyer le tram sur une voie secondaire sur laquelle se trouve une personne. Vous pouvez actionner l'aiguillage, provoquant la mort probable de cette personne ou ne rien faire, laissant les cinq autres personnes à leur sort tout aussi probablement néfaste. Que faites-vous ?

Transposition au cas de la voiture à conduite automatisée : dans une situation hypothétique où un choix doit être fait entre plusieurs actions, toutes pouvant provoquer une collision faisant des victimes, que doit décider le système de conduite automatique (Bonnefon 2016) ? De nombreuses variantes ont été étudiées où les victimes possibles sont des piétons, des cyclistes, les occupant du véhicule lui-même, des enfants, etc. La question qui nous intéresse ici est cependant la suivante : le système d'intelligence artificielle du véhicule est-il capable de délibération éthique ?

Derrière cette question se trouve la problématique de la responsabilité en cas d'accident dans lequel est impliqué un tel véhicule. S'agira-t-il du constructeur, du concepteur du système d'IA, du propriétaire du véhicule, etc. ? Ou du véhicule lui-même ? Cette question est exacerbée, dans le cas du système d'IA apprenant, par les problématiques de robustesse technique ou de biais que nous avons évoquées.

Si tant est que cette question de dilemme a un sens dans la vie réelle, il devrait pourtant être clair que le concept d'agent moral ou éthique artificiel est un non-sens (nous utiliserons ici « moral » et « éthique » dans un sens équivalent). Le calcul effectué par la machine basé sur les données de ses capteurs est toujours le résultat d'un algorithme explicite programmé par des humains, ou celui d'un modèle obtenu par une statistique, probablement dans ce cas à partir de simulations. Dans l'un ou l'autre cas, il ne peut s'agir de délibération morale effectuée par la machine, à partir de valeurs et de concepts qui ne peuvent être compris que par les humains, comme la dignité humaine ou la responsabilité.

Cette confusion entre calcul et possibilité de délibération éthique s'est cependant immiscée dans la philosophie et dans le droit (Cervantes 2020). Ces concepts ne sauraient être exprimés de manière mathématique et être l'objet d'un traitement algorithmique.

Délibération, éthique et intelligence

Le manque de robustesse peut avoir des conséquences catastrophiques dans les domaines de la santé ou du transport par exemple, où des vies sont en jeu. Dans ces domaines, des normes ont été élaborées afin de garantir la sûreté de fonctionnement des systèmes techniques qui y sont déployés. On peut donc envisager des contraintes sur le développement des systèmes d'IA (SIA) pour les rendre compatibles avec ces normes, et élaborer en particulier des procédures de certification pour garantir leur conformité avec les normes dans ces secteurs.

Mais dans d'autres domaines, il est plus difficile de mesurer les conséquences d'un fonctionnement non désiré. Il en est ainsi par exemple dans les secteurs de l'assurance, du recrutement ou des décisions de justice dans lesquels les biais des données ou ceux de conception du système pourront résulter en un traitement non équitable de certaines personnes. Si le biais « purement » statistique pourrait être éventuellement corrigé par un traitement spécifique des données pour les rendre plus représentatives, il n'en est pas de même du biais social. En effet, les données pourraient être statistiquement représentatives, mais refléter des situations sociales discriminatoires réelles. Des décisions prises par les SIA pourront ainsi défavoriser systématiquement certaines catégories de la population, là ou un décideur humain, ayant conscience de cet état de fait, pourrait prendre une décision allant dans le sens opposé, avec l'objectif de compenser un déséquilibre réel. Certes un autre décideur humain pourrait faire l'opposé et une diversité de décision exister. La particularité des SIA sera d'*automatiser* et de généraliser la décision qui ira toujours dans le même sens, celui de renforcer le biais existant.

Le rôle du décideur humain, capable de délibération est ainsi essentiel pour prendre des décisions impactant les êtres humains dans leurs vies quotidiennes, leurs droits ou leurs valeurs.

Conclusion : la responsabilité humaine

Ce qui rend l'être humain responsable de ses actes est son *autonomie* de décision : sa capacité à décider de lui-même de ses objectifs et de la manière de les accomplir. Un agent artificiel ne présente aucune de ces capacités.

Comme pour tous les systèmes socio-techniques qui jouent un rôle important dans nos sociétés, il s'agit donc de garantir la responsabilité humaine dans le développement, le déploiement et l'usage des systèmes d'intelligence artificielle. Ils doivent être alignés sur les principes et les normes éthiques, et techniquement robustes dans leurs contextes d'utilisation. Une législation spécifique comme celle qui est proposée par la Commission européenne est un pas dans ce sens. Dans le souci de ne pas sur-légiférer, celle-ci est basée sur la notion de risque, et vise donc plus particulièrement des secteurs et des usages où ce risque peut être plus facilement formalisé et mesuré. Cependant, les domaines qui peuvent mettre en cause l'autonomie humaine et la démocratie elle-même, comme les systèmes de recommandation au cœur des réseaux sociaux pourraient échapper à cette approche. Il s'agira donc de d'aller vers une plus grande responsabilisation des acteurs humains et de leurs organisations.

Références

CNPZN. Comité National Pilote d'Ethique du Numérique. https://www.ccne-ethique.fr/sites/default/files/cnpen-avis-vehicule-autonome-avril-2021.pdf

Yan Le Cun. Quand la machine apprend : La révolution des neurones artificiels et de l'apprentissage profond. Odile Jacob, Paris (2019).

Philippa Foot. Virtues and Vice. Oxford: Clarendon (2002).

Jean-François Bonnefon, Azim Shariff, Iyad Rahwan. The Social Dilemma of Autonomous Vehicles, SCIENCE, 6293 (352), pp. 1573-1576. (2016). DOI: 10.1126/science.aaf2654

Cervantes, JA., López, S., Rodríguez, LF. et al. Artificial Moral Agents: A Survey of the Current Status. Sci Eng Ethics 26, 501–532 (2020). https://doi.org/10.1007/s11948-019-00151-x