Lignes directrices pour assurer le contrôle humain d'un simulateur par apprentissage machine profond nécessitant une hybridation entre l'explication et la compréhension de ses spécifications formalisées par la théorie de la modélisation et de la simulation

Christophe DENIS
Sorbonne Université - LIP6 - UMMISCO
Panthéon Sorbonne - IHPST
Université de Rouen-Normandie - ERIAC
Christophe.Denis@lip6.fr

Notre contribution a pour objectif de partager les lignes directrices de nos travaux visant à définir un cadre de spécification formelle de systèmes basés sur de l'apprentissage machine profond afin d'assurer un contrôle rigoureux et systématique de ces systèmes. Nos travaux s'inscrivent dans la thématique d'hybridation entre l'apprentissage machine profond et des méthodes de représentation formelle. Notre approche est basée sur une hybridation nécessairement interdisciplinaire : en philosophie, plus précisément concernant une mise en perspective à l'aune de cette technologie des concepts philosophiques en épistémologie, en herméneutique et sur la technique, et en informatique, concernant la modélisation des phénomènes complexes. La performance des réseaux de neurones profonds les rendent en effet disponibles au sens de Heidegger en révélant leur utilité pratique [1]. Ces performances marquent un contraste abyssal avec les premières méthodes d'apprentissage, dont l'exemple emblématique est le perceptron développé en 1957 par le psychologue et informaticien Franck Rosenblatt [2]. Le perceptron, malgré des performances pénalisant son utilisation pratique, s'est construit sur des spécifications possédant une continuité épistémique facilitant l'interprétation et le contrôle des résultats. Est-il donc raisonnable par mesure de précaution d'interdire l'exploitation de systèmes basées sur de l'apprentissage machine profond dont on peine à comprendre ses mécanismes internes bien que leurs performances agissent sur des pans entiers de la société humaine ? La réponse des comités d'éthique est non, du moins pas dans tous les cas, pour ne pas priver l'humanité d'innovation qui pourrait lui être bénéfique comme dans le cadre de la santé ou pour trouver des nouveaux relais de croissance économique. Dans cette optique, la Commission Européenne a récemment adopté une réglementation sur l'Intelligence Artificielle basée sur les risques, dans une approche conséquentialiste et utilitariste de l'éthique, pour trouver un équilibre entre régulation et innovation. Pour les systèmes à risque, et non interdits, la réglementation européenne, comme ce fut le cas notamment par le législateur français autour du principe de garantie humaine [3], impose une analyse d'impact avant la mise sur le marché, et de fournir une explication du fonctionnement des systèmes [4].

Les réglementations autour de l'Intelligence Artificielle se basent essentiellement sur une conception instrumentale de la technique en négligeant sa capacité de dévoilement heideggérien qui modifie le rapport de l'être humain au monde [8]. Pour assurer un équilibre entre ces deux facettes de la technique, nous considérons qu'il est nécessaire d'établir une relation transductive entre l'explication et la compréhension dont nous avons jusque présent mené séparément une clarification épistémologique [5]. L'établissement d'une explication nécessite au préalable de formaliser le système, de l'hybrider avec une représentation des connaissances, que nous proposons d'effectuer en instanciant les concepts de phénomène cible, de modèle, de simulateur et de protocole expérimental issus de la théorie de la modélisation et de la simulation [6][7]. La prise en compte du protocole expérimental est ici cruciale car permettant d'englober le système dans un cas d'usage, dans un contexte permettant un jugement à partir des décisions prédites [8]. La compréhension relève de l'herméneutique comme proposée dans le cadre de la théorie du support formulant que la connaissance est le fruit d'un mécanisme d'interprétation des propriétés matérielles de l'inscription [9]. Ce processus nécessite de lever les ruptures épistémiques induites par la technique d'apprentissage machine profond pour englober le système dans un contexte épistémique cohérent comme ce fut le cas par exemple pour le perceptron.

Références

[1] Martin Heidegger, « *Essais et conférences. La question de la technique* », Édition Gallimard, traduction de André Préau, 1958.
[2] Franck Rosenblatt, «*The perceptron: A probabilistic model for information storage and organization in the brain* », Psychological Review, 65(6), 386–408, 1957.
[3] « *Loi n° 2021-1017 du 2 août 2021 relative à la bioéthique »,* https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000043884384
[4] « *The European Commission released the proposed Regulation on Artificial Intelligence (EU AI Act)* », en ligne sur le site https://www.euaiact.com/, dernière mise à jour le 25 novembre 2022.
[5] Christophe Denis, Franck Varenne, « *Interprétabilité et explicabilité de phénomènes prédits par de l'apprentissage machine* », Revue Ouverte d'Intelligence Artificielle, 2022.
[6] Bernard P. Zeigler, Alexandre Muzy, Ernesto Kofman, «*Theory of Modeling and Simulation: Discrete Event & Iterative System Computational Foundations* », Academic Press, Third Edition, 2023.
[7] Christophe Denis, « *Cadre méthodologique de spécification formelle d'un simulateur par apprentissage machine profond pour assurer sa validation* », Revue des Nouvelles Technologies de l'Information, Extraction et Gestion des Connaissances, RNTI-E-40, 2024.
[8] Bruno Bachimont, « *Une décision calculée peut-elle tenir lieu de jugement ? Considérations sur la faculté de juger et son instrumentation* », Questions de communication, 2022.
[9] Bruno Bachimont, « *Le Sens de la technique : le numérique et le calcul* », Collection À présent, Edition Encre Marine, 2010.