Lignes directrices d'un cadre de spécification pour contrôler les systèmes basés sur de l'apprentissage machine profond utilisant la théorie de le modélisation et de la simulation et un héritage nécessairement revisité de la philosophie de la technique.

Christophe Denis

Lignes directrices d'un cadre de spécification pour contrôler les systèmes basés sur de l'apprentissage machine profond utilisant la théorie de le modélisation et de la simulation et un héritage nécessairement revisité de la philosophie de la technique.

Christophe DENIS
Sorbonne Université - LIP6 - UMMISCO
Panthéon Sorbonne - IHPST
Université de Rouen-Normandie - ERIAC
Saclay Industry Lab for Artificial Intelligence Research
Christophe.Denis@{lip6.fr,edf.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. Pour cela, nous mettons en oeuvre une adaptation de la théorie de la modélisation et de la simulation [1] pour ces types de systèmes, guidée par notre travail de clarification épistémologique des concepts d'explication et de compréhension appliqués à l'apprentissage machine [2]. Le résumé est organisé comme suit. Nous commençons par établir dans le prochain paragraphe une mise en perspective historique de l'évolution des techniques d'apprentissage machine permettant une confrontation avec certaines théories philosophiques portant sur la technique. Nous indiquons ensuite la stratégie retenue par la Commission Européenne pour réglementer les systèmes basés sur de l'apprentissage machine profond [3]. Le dernier paragraphe présente les lignes directrices du cadre de spécification que nous proposons pour mener un contrôle systématique des systèmes conçues avec des techniques d'apprentissage machine quelque soit son supposé impact [4].

Les progrès technologiques et scientifiques réalisés depuis les années 2010 dans le domaine des réseaux de neurones artificiels ont considérablement amélioré les performances des systèmes basés sur de l'apprentissage machine. Ceci représente un contraste marqué avec les premiers réseaux de neurones développés dans les années 1950, qui étaient exclusivement conçus pour explorer les possibilités de mécanisation du raisonnement humain sur ordinateur au sein de la discipline émergente de l'Intelligence Artificielle. Un exemple emblématique de premières méthodes d'apprentissage est le perceptron développé en 1957 par le psychologue et informaticien Franck Rosenblatt. L'objectif du perceptron était de créer une unité de calcul électronique en s'inspirant du fonctionnement des neurones biologiques, dans le but de « *répondre aux questions de savoir comment les informations sur le monde physique sont perçues, sous quelle forme elles sont mémorisées et comment les informations conservées en mémoire influencent la reconnaissance et le comportement* » [5]. L'exemple du perceptron montre une continuité épistémique facilitant l'interprétation et le contrôle des résultats. Toutefois, l'analyse menée par Minsky et Papert en 1969, démontrant que le perceptron et ses améliorations ne permettaient pas de résoudre avec une précision suffisante des problèmes d'intérêt en raison de la linéarité de la frontière de décision, a considérablement limité les financements alloués à cette approche au profit d'une représentation symbolique. Ces deux chercheurs ont proposé de concevoir des réseaux de neurones multicouches, une possibilité rendue seulement réalisable depuis vingt ans, d'abord sur le plan technologique, en raison de la disponibilité d'une puissance de calcul élevée, et au niveau scientifique par la conception d'algorithmes d'optimisation efficaces. Les réseaux de neurones actuels sont qualifiés de « boite noire » car constitués d'un grand enchevêtrement de fonctions de transfert qui nécessite de caler grâce aux algorithmes d'optimisation récemment développés des millions de paramètres. On peut donc considérer un réseau de neurones profond comme un automate numérique dont ses spécifications sont implicitement déterminées lors du calage agnostique de ses paramètres. De plus, cette rétrospective historique met en évidence la nécessité de réexaminer, à la lumière des caractéristiques des nouvelles technologies numériques, les concepts philosophiques sur la technique formulés après la seconde révolution industrielle. Cela englobe des notions telles que l'évolution continue des techniques et la concrétude technique théorisées par Gilbert Simondon [7], nécessitant d'être au moins actualisées voire complétées en raison par exemple d'une méconnaissance des mécanismes internes d'un réseau de neurones profonds.

La performance des réseaux de neurones profonds les rendent disponibles au sens de Heidegger en révélant leur utilité pratique [8]. Toutefois, est-il raisonnable par mesure de précaution d'interdire l'exploitation de tels systèmes dont on peine encore à comprendre ses mécanismes internes bien que leurs performances bouleversent 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. Nous écartons dans la suite du résumé les applications qui sont interdites par l'Union Européenne, comme la notation sociale basée sur le comportement individuel. Il s'agit d'un choix politique indépendante de la technologie, car on peut par exemple mettre en place une notation sociale basé sur un comportement individuel plus simple à prédire ne nécessitant pas la performance prédictive des techniques d'apprentissage machine profond. Pour les systèmes à risque, la réglementation européenne, comme ce fut le cas notamment par le législateur français autour du principe de garantie humaine, impose une analyse d'impact avant la mise sur le marché, et de mener de fournir une explication du fonctionnement des systèmes.

Les réglementations actuelles autour de l'Intelligence Artificielle se basent essentiellement sur une conception instrumentale de la technique en négligeant sa capacité de dévoilement heideggerien qui modifie le rapport de l'être humain au monde. Pour assurer un équilibre entre ces deux facettes de la technique, nous considérons qu'il est important de considérer une relation interpénétrante entre l'explication et la compréhension dont nous avons jusque présent mené séparément une clarification épistémologique [2]. L'établissement d'une explication nécessite au préalable de formaliser le système, 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. Le protocole expérimental revêt d'une importance particulière car permettant d'englober le système dans un cas d'usage, dans un contexte permettant un jugement à partir des décisions prédites [9]. 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 [10]. Ce processus nécessite de lever les ruptures épistémiques induits 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] Bernard P. Zeigler, Alexandre Muzy, Ernesto Kofman, «*Theory of Modeling and Simulation: Discrete Event & Iterative System Computational Foundations* », Academic Press, Third Edition, 2023.

[2] Christophe Denis, Franck Varenne, « *Interprétabilité et explicabilité de phénomènes prédits par de l'apprentissage machine* », Revue Ouverte d'Intelligence Artificielle, 2022.

[3] « *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.

[4] 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.

[5] Franck Rosenblatt, «*The perceptron: A probabilistic model for information storage and organization in the brain* », Psychological Review, 65(6), 386–408, 1957.

[6] Marvin Minsky, Seymour Papert, « Perceptrons », MIT Press, Cambridge, 1969.

[7] Gilbert Simondon, «Du mode d'existence des objets techniques », Éditions Aubier-Montaigne, 1958.

[8] Martin Heidegger, « *Essais et conférences. La question de la technique* », Édition Gallimard, traduction de André Préau, 1958.

[9] Bruno Bachimont, « « Une décision calculée peut-elle tenir lieu de jugement ? Considérations sur la faculté de juger et son instrumentation », Questions de communication, 2022.

[10] Bruno Bachimont, « *Le Sens de la technique : le numérique et le calcul* », Collection À présent, Edition Encre Marine, 2010.