# Vision croisée de l'IA Explicable entre philosophie et informatique

Cédric Brun, Ikram Chraibi Kaadoud

# Vision croisée de l'IA Explicable entre philosophie et informatique

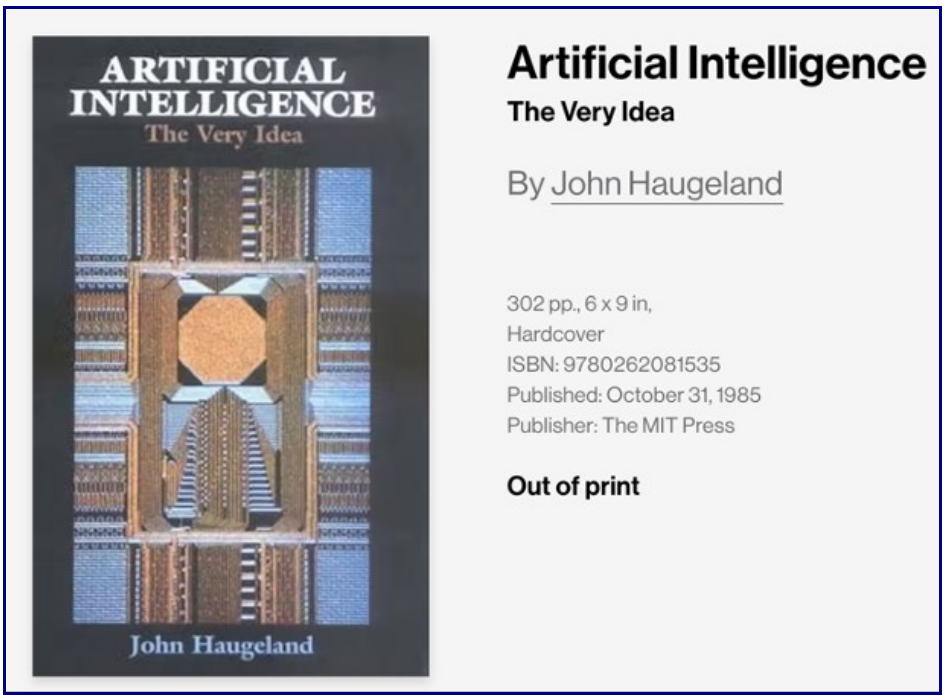
Pourquoi ne pas proposer au grand public **une vision croisée de l'IA explicable entre philosophie et sciences informatiques** des enjeux, des pistes de recherche et des conditions auxquelles l**es outils d'explications pourront être déterminants pour la confiance des utilisateurs et des autorités de régulation** (s'il y en a un jour) de ces outils. Autrement dit, un **fil de philosophie de l'Intelligence Artificielle (IA**) sur l'**opacité des systèmes computationnels complexes d'un point de vue de la philosophie des sciences**. **Cédric Brun,** chercheur en philosophie des sciences en **Neuroscience, Humanities & Society (NeHuS)** et **Ikram Chraibi Kaadoud**, chercheuse en IA explicable et IA digne de confiance, nous partage cette vision. **Thierry Viéville**.

## *Philosophie de l'IA*

La philosophie s'est très tôt intéressée à l'IA comme discipline et comme projet théorique.

D'abord parce qu'une partie de l'IA et de la philosophie de l'esprit posent des **questions proches, avec des outils différents**. Par exemple, **John Haugeland,** feu professeur émérite de philosophie à l'université de Chicago, a discuté en 1980 dans son livre "*Artificial Intelligence, the Very Idea*" (image ci dessous) de l'idée que la **pensée humaine et le traitement formel de l''information dans une machine sont "radicalement les mêmes"**. Le contexte de l'époque opposait alors les humanistes qui soulignaient que "Les machines qui pensent - c'est tout à fait absurde" et les techno-visionnaires qui soutenaient que "L'intelligence artificielle est là et sur le point de surpasser la nôtre".

43 ans après, force est de constater que ces questions, posées probablement différemment, sont toujours d'actualité.




"*Artificial Intelligence, the Very Idea" de* ***John Haugeland*** (Trad fr. J. Henry, *L'esprit dans la machine, fondement de l'intelligence artificielle*, Odile Jacob 1989) Src: https://mitpress.mit.edu/9780262081535

L'une des questions centrales mais souvent évitées est celle de la **nature de l'intelligence :** qu'est-ce que l'intelligence ? L'intelligence humaine peut-elle être reproduite, voire dépassée, par des outils computationnels ? La conscience (ou au moins la connaissance réflexive) d'un agent intelligent humain peut-elle être simulée, reproduite, voire réalisée par des machines ?

**Hilary Putnam,** philosophe américain co-fondateur du **computationnalisme*** et figure centrale de la philosophie contemporaine américaine, a tenté d'apporter des éléments de réponses à ces questions dans son article "*Minds and Machines"* (Les esprits et les machines) de 1960. Selon lui, les différentes questions et énigmes qui constituent le problème traditionnel du corps et de l'esprit peuvent entièrement être approchées par leur nature linguistique et logique. Cette approche l'a ainsi conduit à conclure **la cognition humaine n'est pas fondamentalement de nature différente d'un traitement formel de symboles par un ordinateur.**


Putnam, H. (1960). Minds and machines.URL: https://philpapers.org/rec/PUTMAM

***Moment Glossaire:***

*****Un **système computationnel** est un modèle qui fait des calculs à partir d'informations données en entrée, et qui donne en sortie un résultat numérique. Source: Collins, A., & Khamassi, M. (2021). Initiation à la modélisation computationnelle

***** Le **computationnalisme** est une théorie fonctionnaliste en philosophie de l'esprit qui conçoit l'esprit comme un système de traitement de l'information et compare la pensée à un calcul (en anglais, *computation*) et, plus précisément, à l'application d'un système de règles. Cette théorie est différente du cognitivisme. Source: https://fr.wikipedia.org/wiki/Computationnalisme

***** Le **cognitivisme** est le courant de recherche scientifique endossant l'hypothèse selon laquelle la pensée est analogue à un processus de traitement de l'information, cadre théorique qui s'est opposé, dans les années 1950, au béhaviorisme. La notion de cognition y est centrale. Elle est définie en lien avec l'intelligence artificielle comme une manipulation de symboles ou de représentations symboliques effectuée selon un ensemble de règles. Elle peut être réalisée par n'importe quel dispositif capable d'opérer ces manipulations.
Source: https://fr.wikipedia.org/wiki/Cognitivisme

Plus récemment, la philosophie de l'IA s'est davantage tournée vers des questions techniques relatives aux différentes architectures et méthodes computationnelles et leurs enjeux épistémologiques, éthiques et politiques **du fait de la pénétration et du développement de l'IA dans la pratique scientifique et dans la société**.

**Daniel Andler** dans son livre *Intelligence artificielle, intelligence humaine : la double énigme* en 2023 (image ci dessous), a introduit **l'idée qu'il existait un écart entre la représentation de la philosophie de l'IA chez les non spécialistes et l'actualité de la recherche en philosophie de l'IA**. Cela est d'autant plus vrai pour lui lorsqu'il s'agit des sujets éthiques et des problèmes fondationnels sur la possibilité d'une Intelligence Générale Artificielle (IGA) ou d'une IA forte, sujets très présents dans la **philosophie du transhumanisme*.**


Src: https://www.gallimard.fr/Catalogue/GALLIMARD/NRF-Essais/Intelligence-artificielle-intelligence-humaine-la-double-enigme

**Une bonne partie de la philosophie de l'IA concerne des enjeux éthiques et politiques de l'IA** tels que par exemple les biais, l'équité, la confiance, la transparence, mais reste toujours liée à des enjeux **épistémologiques*** selon les outils techniques mobilisés en IA, à savoir apprentissage machine supervisé ou non, Réseaux de neurones profonds convolutifs ou encore systèmes symboliques classiques.

***Moment Glossaire:***

***** La **philosophie du transhumanisme ou transhumanisme** est une doctrine philosophique prétendant qu'il est possible **d'améliorer l'humanité par la science et la technologie en libérant l'humanité de ses limites biologiques** notamment en surmontant l'évolution naturelle. Le changement apporté à l'humain serait positif, car cela pourrait signifier la libération des contraintes de la nature, comme la maladie ou la mort. L'idée centrale est celle d'un dépassement de l'humain (et non de son élimination) par l'intermédiaire des techniques qui évoluent de manière très rapide. Source: Le transhumanisme selon https://philosciences.com/ Pour en savoir plus: https://encyclo-philo.fr/transhumanisme-a

*** L'épistémologie** désigne de manière générale l**'étude de la connaissance et de ses conditions de possibilité.** En un sens plus spécifique, c'est un domaine de la philosophie qui **étudie les disciplines scientifiques et les conditions logiques, méthodologiques et conceptuelles de production des connaissances scientifiques**. Pour un domaine scientifique particulier (l'IA par exemple), l'épistémologie désignera l'**étude critique des savoirs** qu'il produit à partir de l'analyse de ses méthodes, pratiques et concepts.

# *Opacité et transparence des Systèmes Computationnels Complexes (SCC)*

L'opacité d'un ***Système Computationnel Complexe (SCC)*** est dérivée du concept d'opacité épistémique. Au sens le plus fort, l'opacité épistémique désigne la complexité (voir l'impossibilité) de suivre et comprendre les processus computationnels impliqués dans un système: les étapes, les justifications et les implications de chaque étape du processus deviennent hors de porté pour des agents cognitifs humains.

Autrement dit, nous ne pouvons expliquer ni pourquoi, ni comment le système produit, en sortie, les résultats (classifications, décisions) qu'il produit selon les données fournies (ou collectées) en entrées du système. On parle alors de **boîte-noire, puisque les processus internes en sont inscrutables.**

Cette opacité s'étend aussi à la **nature exacte des données pertinentes au fonctionnement du système dans le cas de l'apprentissage profond.**

Rappelez-vous les réactions aux premiers résultats de **Parcoursup** en mai 2023, faites une recherche "#PARCOURSUP + opacité" sur X (anciennement Twitter), pour voir. Nous retrouvons alors des opinions comme celle-ci :


En résumé, **l'opacité survient lorsque nous ne savons pas exactement comment le comportement du système est produit ni sur quelles données (ou propriétés de ces données) il s'appuie pour produire ce comportement.**

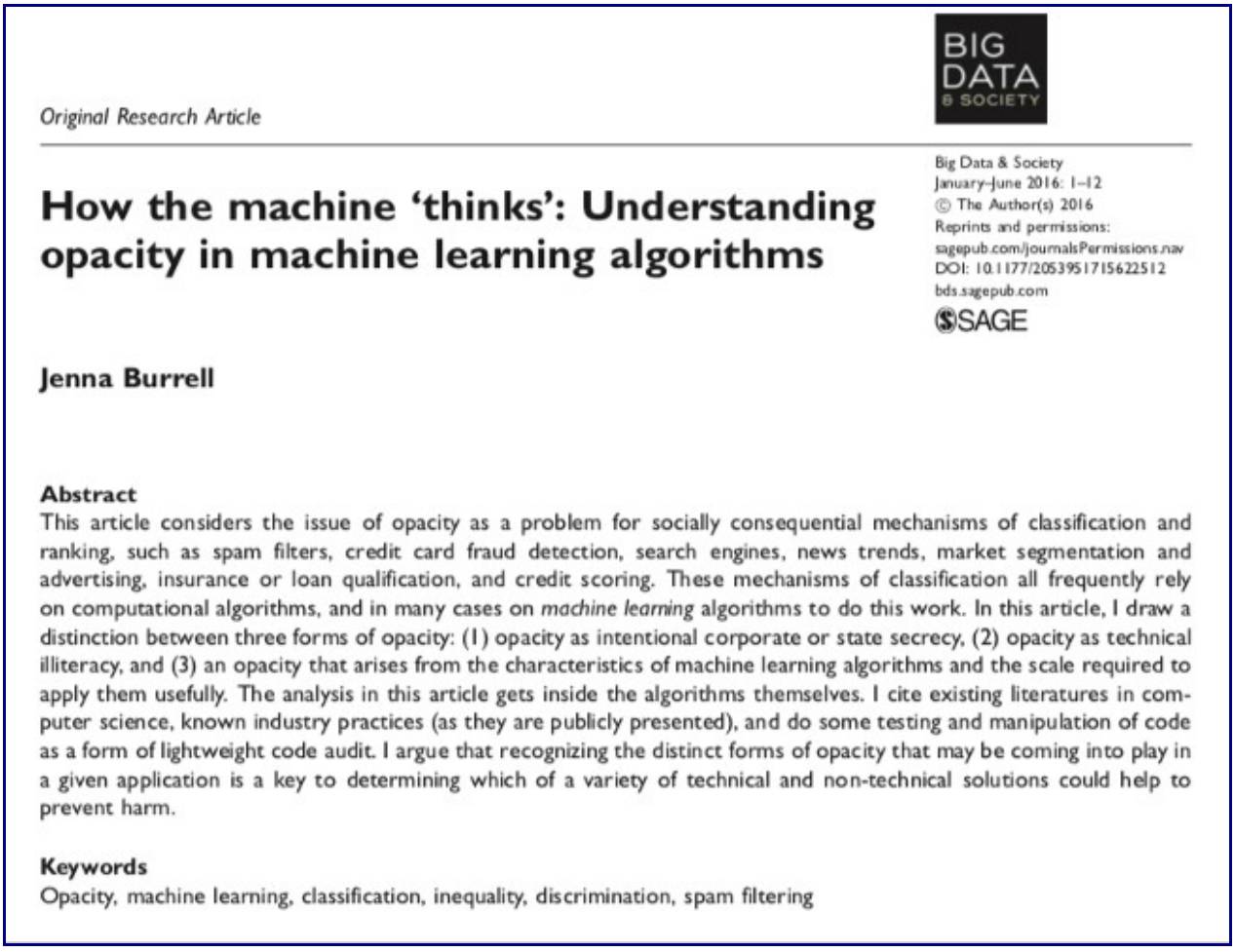
Dans un article de 2016 intitulé "*How the machine 'thinks': Understanding opacity in machine learning algorithms*" (Comment les machines pensent: comprendre l'opacité dans les algorithmes d'apprentissage autonome), **Jenna Burrell** @jennaburrell , alors professeure à l'UC Berkeley, a examiné la question de l'**opacité en tant que problème pour les mécanismes de classification et de classement ayant des conséquences sociales,** tels que les filtres anti-spam, la détection des fraudes à la carte de crédit, les moteurs de recherche, les tendances de l'actualité, la segmentation du marché et la publicité, l'assurance ou la qualification des prêts, et l'évaluation de la solvabilité. Ces mécanismes de classification s'appuyent tous fréquemment sur des algorithmes et, dans de nombreux cas, sur des algorithmes d'apprentissage automatique.

La chercheuse distingue ainsi **3 types d'opacité** :

- **(1) Intentionnel :** l'opacité en tant que secret d'entreprise ou d'État intentionnel,
- **(2) Educationnelle** : l'opacité en tant qu'analphabétisme technique (technical illiteracy)
- **(3) Opératoire** : l'opacité qui découle des caractéristiques des algorithmes d'apprentissage automatique et de l'échelle requise pour les appliquer de manière utile.

Les deux premiers types ne sont pas spécifiques à l'apprentissage machine/profond. On les retrouve dans tous les domaines techniques et scientifiques : essayez de démonter et réparer un écran OLED de dernière génération, pour voir ; ou de dépanner vous-mêmes votre voiture hybride. Ne serait-ce que comprendre les processus engagés entre l'action réalisée par votre votre index sur la télécommande et le résultat sur l'écran ( par exemple le changement de chaîne) représente un défi si vous n'avez pas de connaissances poussées en physique et en électronique. Votre téléviseur est une boîte-noire, autrement dit une "lucarne MAGIQUE".

La spécificité des SCC, en tout cas de certains, c'est que **même si vous avez accès au code et que vous avez toutes les connaissances nécessaires pour concevoir ce système, son caractère récursif, l'échelle à laquelle il fonctionne et l'organisation dynamique de ses données produisent une opacité opératoire** qui vous affecte quasiment au même titre que le béotien[1] !



*Original Research Article*

BIG DATA & SOCIETY

# How the machine 'thinks': Understanding opacity in machine learning algorithms

Big Data & Society
January–June 2016: 1–12

DOI: 10.1177/2053951715622512
bds.sagepub.com
SAGE

**Jenna Burrell**

**Abstract**

This article considers the issue of opacity as a problem for socially consequential mechanisms of classification and ranking, such as spam filters, credit card fraud detection, search engines, news trends, market segmentation and advertising, insurance or loan qualification, and credit scoring. These mechanisms of classification all frequently rely on computational algorithms, and in many cases on *machine learning* algorithms to do this work. In this article, I draw a distinction between three forms of opacity: (1) opacity as intentional corporate or state secrecy, (2) opacity as technical illiteracy, and (3) an opacity that arises from the characteristics of machine learning algorithms and the scale required to apply them usefully. The analysis in this article gets inside the algorithms themselves. I cite existing literatures in computer science, known industry practices (as they are publicly presented), and do some testing and manipulation of code as a form of lightweight code audit. I argue that recognizing the distinct forms of opacity that may be coming into play in a given application is a key to determining which of a variety of technical and non-technical solutions could help to prevent harm.

**Keywords**

Opacity, machine learning, classification, inequality, discrimination, spam filtering

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. Big data & society, 3(1), 2053951715622512. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2660674

# Quid de la communauté scientifique IA ?

Les chercheur.e.s en IA, les scientifiques qui utilisent des outils complexes de traitement computationnel des données massives et les philosophes de l'IA ont fait valoir que **réduire l'opacité, et amener de la transparence** exigeait : **(i)** un effort en direction de **plus de transparence,** d'interprétabilité ou d'explicabilité (tous ces concepts doivent être soigneusement distingués, mais cela nous engagerait dans un long développement technique). Cela s'est parfois traduit dans des règlements internationaux (RGPD par exemple en Europe), **(ii)** des **programmes de recherche (privés et publics)** et **(iii)** un grand nombre de **publications en IA, en SHS, en droit, en sciences politiques**, autrement dit, de la pluridisciplinarité!

Cette exigence de transparence repose sur l'**espoir d'une plus grande confiance des utilisateurs** proximaux ou finaux. L'argument étant le suivant: **la confiance dans une personne provient de la capacité à exhiber les raisons de ses décisions**. Si ces raisons sont impénétrables, inaccessibles ou opaques, alors il n'y aura pas de pleine confiance. Certains auteurs ont cependant déjà prévenu (il y a

déjà un certain temps) que cette opacité des SCC était **inhérente, insurmontable et inéliminable** et qu'il fallait faire avec.

## *- Deal with it! -*

**Paul Humphreys**, Professeur britannique de philosophie à l'université de Virginie, spécialisé dans la philosophie des sciences, la métaphysique et l'épistémologie, s'est intéressé à la métaphysique et à l'épistémologie de l'émergence, à la science informatique, à l'empirisme et au réalisme. En 2009, il explique dans son article "*The philosophical novelty of computer simulation methods*" (La nouveauté philosophique des méthodes de simulation informatique) que **les simulations informatiques** et **la science computationnelle** sont un ensemble de **méthodes scientifiques distinctement nouvelles** qui introduisent de nouvelles questions à la fois épistémologiques et méthodologiques dans la philosophie des sciences.

Ces outils numériques, utilisés à grande échelle, **modifient profondément la pratique scientifique, mais surtout les buts de la recherche scientifique:** La modélisation et la simulation computationnelles nous conduiraient à envisager la recherche scientifique comme visant **la prédiction** de phénomènes ou processus modélisés, plutôt que leur **compréhension ou explication.**

Synthese (2009) 169:615–626
DOI 10.1007/s11229-008-9435-2

# The philosophical novelty of computer simulation methods

**Paul Humphreys**

Received: 5 December 2007 / Accepted: 17 October 2008 / Published online: 28 November 2008


**Abstract** Reasons are given to justify the claim that computer simulations and computational science constitute a distinctively new set of scientific methods and that these methods introduce new issues in the philosophy of science. These issues are both epistemological and methodological in kind.

**Keywords** Computer simulations · Computational science · Epistemic opacity · Semantics · Temporal dynamics · Approximations

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. Synthese, 169, 615-626. URL : https://philpapers.org/rec/HUMTPN

Certains chercheurs considèrent que si l'on ne parvient pas à réduire cette opacité fondamentale en choisissant des architectures et méthodes plus transparentes que l'apprentissage machine non supervisé ou l'apprentissage profond, il faudrait alors **exclure l'utilisation des SCC de certains cas.**

Par exemple, en médecine, dans le domaine judiciaire ou l'éducation, les SCC devraient être suffisamment transparents du point de vue opératoire car sans cela, leur utilisation ne devrait pas être autorisée.

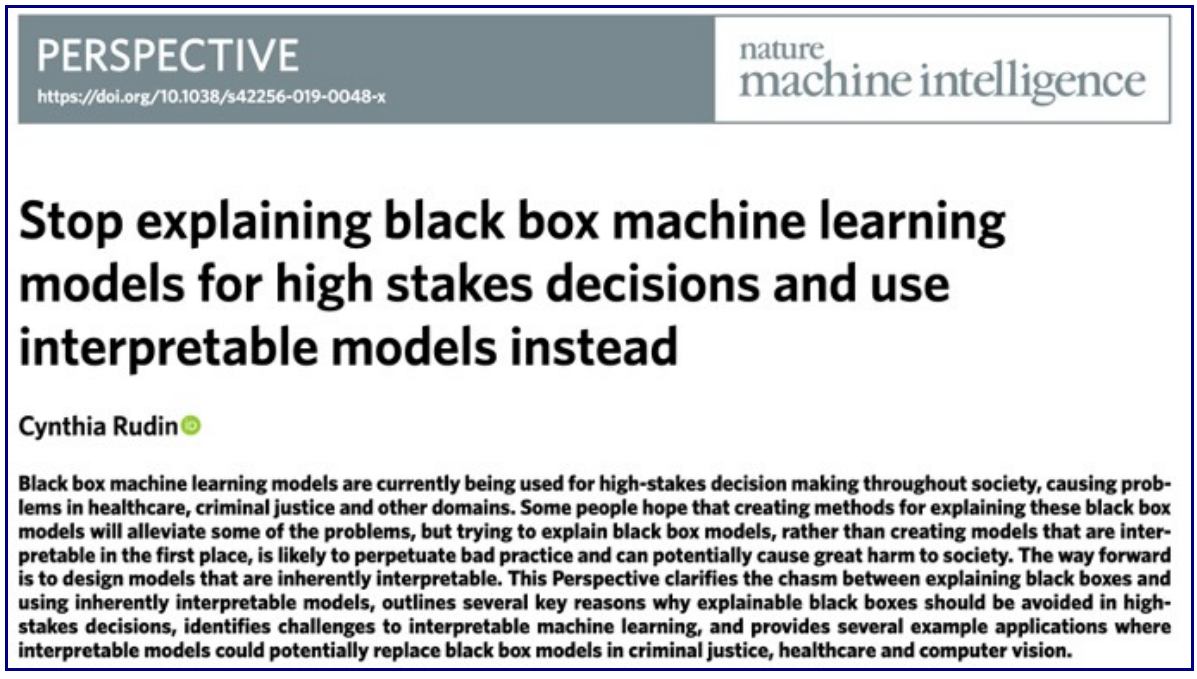
C'est autour de cette idée que se sont construits les travaux de **Cynthia Rudin** @CynthiaRudin qui ont annoncé un changement de cap dans le domaine de **l'explicabilité en IA** notamment centrée-humain.

Informaticienne et statisticienne américaine spécialisée dans l'apprentissage automatique, elle est notamment connue pour ses travaux sur **l'interprétabilité des algorithmes d'apprentissage**

**automatique**. Directrice de l'*Interpretable Machine Learning Lab* à l'université Duke, où elle est professeur d'informatique, d'ingénierie électrique et informatique, de science statistique, de biostatistique et de bio-informatique, elle a remporté en 2022 le Squirrel AI Award for Artificial Intelligence for the Benefit of Humanity de l'Association for the Advancement of Artificial Intelligence (AAAI) pour ses travaux sur l'importance de la transparence des systèmes d'IA dans les domaines à haut risque.

Dans son article de 2019 intitulé "*Stop explaining black-box machine learning models for high-stakes decisions and use interpretable models instead.*" (Cessez d'expliquer des modèles d'apprentissage automatique à boîte noire pour des décisions à fort enjeu et utilisez plutôt des modèles interprétables.), Cynthia Rudin se penche sur l'idée que la création de méthodes permettant d'expliquer ces modèles boîte noire atténuera certains des problèmes éthiques recensés dans la littérature. Elle y discute notamment l'**idée que s'échiner à expliquer les modèles boîte noire, plutôt que de créer des modèles interprétables en premier lieu, risque de perpétuer les mauvaises pratiques et peut potentiellement causer un grand préjudice à la société**.

La voie à suivre, selon elle, consiste à **concevoir des modèles intrinsèquement interprétables** comme pour les décisions à fort enjeu notamment dans la justice pénale, les soins de santé et la vision par ordinateur.



PERSPECTIVE

https://doi.org/10.1038/s42256-019-0048-x

nature machine intelligence

# Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

**Black box machine learning models are currently being used for high-stakes decision making throughout society, causing problems in healthcare, criminal justice and other domains. Some people hope that creating methods for explaining these black box models will alleviate some of the problems, but trying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society. The way forward is to design models that are inherently interpretable. This Perspective clarifies the chasm between explaining black boxes and using inherently interpretable models, outlines several key reasons why explainable black boxes should be avoided in high-stakes decisions, identifies challenges to interpretable machine learning, and provides several example applications where interpretable models could potentially replace black box models in criminal justice, healthcare and computer vision.**

Src: Rudin, C. (2019). Stop explaining black-box machine learning models for high-stakes decisions and use interpretable models instead. Nature machine intelligence, 1(5), 206-215. URL: https://tinyurl.com/4vtac7zh

En parallèle à ce mouvement lancé par Cyntia Rudin, d'autres chercheurs et industriels, pensent qu'en distinguant des types de transparence on pourra limiter l'opacité opératoire et gagner en confiance, ainsi qu'en maîtrise (recherche de bugs par les modélisateurs) et en capacité explicative (pour les

scientifiques utilisant ces outils). C'est notamment ce que propose **Kathleen Creel** @KathleenACreel dans un article de 2020 extrêmement éclairant, "*Transparency in Complex Computational Systems*" ("Transparence des systèmes computationnels complexes").

Professeure assistante à la Northeastern University, Kathleen Creel mène des travaux sur les implications morales, politiques et épistémiques de l'apprentissage automatique tel qu'il est utilisé dans la prise de décision automatisée non étatique et dans la science. Elle a notamment contribué à intégrer les enseignements d'éthique aux programmes informatiques de Stanford afin de permettre l'acquisition de compétences aux étudiants qui leur permettraient de discuter et de réfléchir aux dilemmes éthiques qu'ils pourraient rencontrer dans leur carrière professionnelle.

Dans cet article de 2020, Kathleen Creel propose **une analyse de la transparence sous trois formes** : **(i)** la transparence de l'algorithme, **(ii)** la réalisation de l'algorithme dans le code et **(iii)**la manière dont le code est exécuté sur un matériel et des données particuliers. En visant la transparence sous ces trois formes, cela permettrait de cibler **la transparence la plus utile pour une tâche donnée** en fournissant **une transparence partielle lorsque la transparence totale est impossible**, tout en évitant un usage instrumentaliste des systèmes opaques.

# Transparency in Complex Computational Systems

forthcoming, *Philosophy of Science*

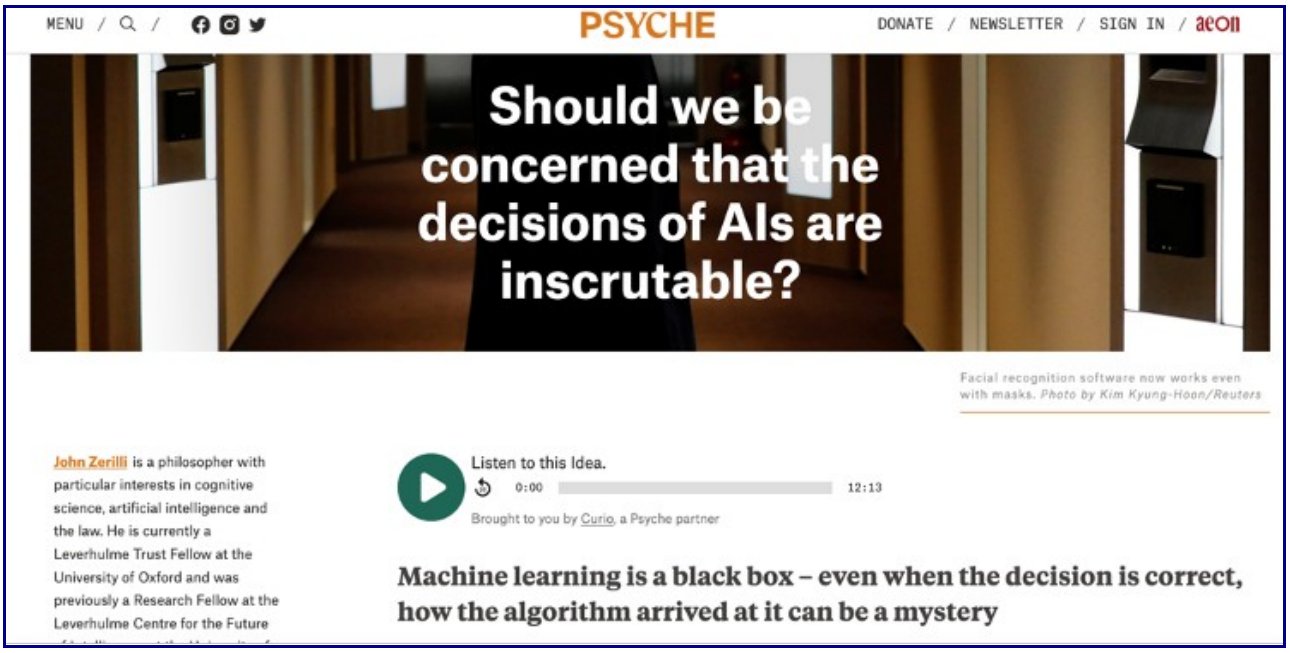
Kathleen A. Creel
kac284@pitt.edu

**Abstract:**

Scientists depend on complex computational systems that are often ineliminably opaque, to the detriment of our ability to give scientific explanations and detect artifacts. Some philosophers have suggested treating opaque systems instrumentally, but computer scientists developing strategies for increasing transparency are correct in finding this unsatisfying. Instead, I propose an analysis of transparency as having three forms: transparency of the algorithm, the realization of the algorithm in code, and the way that code is run on particular hardware and data. This targets the transparency most useful for a task, avoiding instrumentalism by providing partial transparency when full transparency is impossible.

Creel, K. A. (2020). Transparency in complex computational systems. Philosophy of Science, 87(4), 568-589. URL: http://philsci-archive.pitt.edu/16669

Enfin, d'autres considèrent qu'en exigeant une telle transparence des SCC, nous faisons deux poids-deux mesures puisque **cette opacité opératoire n'est qu'une sous-catégorie de l'opacité épistémique** dans laquelle nous nous trouvons face à nos congénères :




https://psyche.co/ideas/should-we-be-concerned-that-the-decisions-of-ais-are-inscrutable

Au fond, nous serions face à l'aide à la décision apportée par un.e docteur.e en médecine avant de donner notre consentement éclairé pour une opération comme nous sommes face à un SCC d'aide à la décision en apprentissage profond. Seul le **contexte d'interaction permettrait de fonder notre confiance**, sans compter que des travaux menés en explicabilité centrée humain, montrent que **dans certains contextes, l'accès à des explications est plus source de stress (et donc de rejets de l'information) que de confiance et d'acceptabilité des SCC.**

Dr Juliette @FerryDanini, enseignante chercheuse en philosophie à l'université de Namur, a fait une communication sur ce débat en 2021 au *Congress of the Quebec Philosophy Society*, nous vous conseillons de la voir si ça vous intéresse, la vidéo étant ci dessous:

Vidéo youtube: https://www.youtube.com/watch?v=xNWe3PsfNng

*TAKE HOME MESSAGE*

Alors que retenir de cette vision croisée de l'IA Explicable, entre philosophie et informatiques ? des réflexions et probablement des questionnements aussi !

**'Take home message #1 :** Comme toujours en philosophie des sciences et techniques, les problèmes éthiques sont liés à des problèmes épistémologiques qui supposent une compréhension des questions pratiques et théoriques centrales soulevées par l'usage de ces méthodes : pas d'indépendance des deux.

**Take-home message #2 :** La philosophie de l'IA suppose une certaine familiarité avec des questions techniques. Idéalement, savoir coder est potentiellement une exigence à viser.

**Take-home message #3 :** Un gros travail interdisciplinaire de définition des concepts centraux (transparence, explicabilité, interprétabilité, opacité) doit être fait pour stabiliser le champ et les stratégies théoriques et pratiques, voire industrielles.

**Take-Home message #4 :** La confiance comme vertu cardinale du rapport aux SCC nous semble à questionner. Il y a du boulot à faire :)

**Cédric Brun,** chercheur en philosophie des sciences en **Neuroscience, Humanities & Society (NeHuS)** et **Ikram Chraibi Kaadoud**, chercheuse en IA explicable et IA digne de confiance.

[1] L'adjectif béotien : de la région de Béotie. Les habitants de la Béotie, province de l'ancienne Grèce, avaient, à Athènes, la réputation d'être un peuple inculte, lourdaud et peu raffiné. De nos jours, l'adjectif béotien, béotienne qualifie un individu peu ouvert aux lettres et aux arts, aux goûts grossiers. Source: https://www.projet-voltaire.fr/culture-generale/beotien-marathon-sybarite-ces-mots-francais_toponymes-grecs-antiques-noms-lieux-grecs/

## *Pour en savoir plus/Références*


- "How a new program at Stanford is embedding ethics into computer science?" Juin 2022, Site : stanford.edu/ , URL article: https://urlz.fr/mf3z
- L'intelligence artificielle explicable (XAI) :
  - Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 58, 82-115. URL: https://www.sciencedirect.com/science/article/abs/pii/S1566253519308103
  - Comment saisir ce que font les réseaux de neurones ? Série de trois articles du blog binaire sur les concepts d'intéprétabilité et d'explicabilité: https://www.lemonde.fr/blog/binaire/2020/09/04/comment-saisir-ce-que-font-les-reseaux-de-neurones/