Aristote

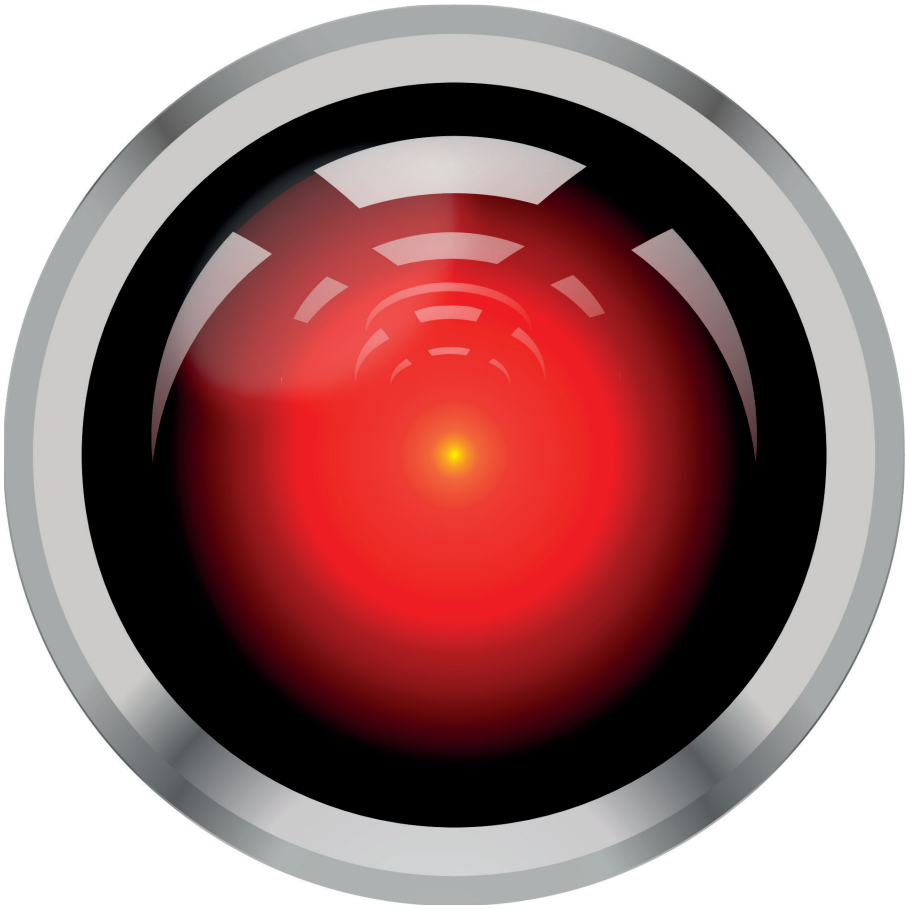
À la croisée des révolutions numériques

École Polytechnique, Palaiseau
Amphithéâtre ARAGO

L'IA est-elle explicable ?

Un coup d'oeil furtif dans la boite noire des algorithmes de l'IA

Jeudi 17 octobre 2019



Coordination scientifique :

- **Christophe DENIS** (LIP6/Sorbonne Université)
- **Jean LATIERE**
- **Julia PETRELLUZZI** (Doctorante en Droit et intelligence artificielle)

LIP6 — SORBONNE UNIVERSITÉ — scai SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Renseignements, programme…
https://www.association-aristote.fr/lia-est-elle-explicable/

ARISTOTE

Séminaire du 17 novembre

« L'IA est-elle explicable ? »

Le séminaire a été organisé par Christophe DENIS (LIP6/Sorbonne Université), Jean LATIERE, et Julia PETRELLUZZI (Doctorante en Droit et intelligence artificielle).

Christophe Calvin, président d'Aristote, introduit rapidement la journée, car le programme est dense. Mais aux nombres de personnes élevé dans la salle, il souligne l'importance de l'explicabilité de l'IA à l'heure actuelle. Et notamment la question juridique que cela pose. *« Ces notions de transparence, d'explicabilité et de reproductibilité vont permettre d'avoir une acceptatibilité - ou non - de l'intelligence artificielle. Cela ne répondra pas à toutes les questions, mais permettront de clarifier notamment les aspects législatifs, un des enjeux majeurs. »*

Editorial Board

Table des matières

- L'IA VA-T-ELLE MODIFIER LA NOTION D'EXPLICATION ? 4
- « ÉPISTÉMOLOGIE DES MODÈLES ET DEMANDES D'EXPLICABILITÉ POUR L'APPRENTISSAGE MACHINE » 6
- EXPLICABILITÉ ET ÉTHIQUE D'APPLICATIONS IA DANS LE DOMAINE MÉDICAL 10
- APPRENTISSAGE ARTIFICIEL ET INTERPRÉTABILITÉ POUR LA MÉDECINE DE PRÉCISION 13
- INTERPRÉTATION D'IMAGES MÉDICALES : QUELQUES TRAVAUX VERS L'EXPLICABILITÉ 16
- L'EXPLICABILITÉ, LA CLÉ POUR LE DÉPLOIEMENT D'IAS DANS L'ENTREPRISE 18
- LES LIMITES DES ALGORITHMES DE DEEP LEARNING DANS LE DOMAINE DE L'IMAGERIE MÉDICALE 21

EMPATHIE, CONSCIENCE ET CRÉATIVITÉ DE L'IA 24

- EXPLICABILITÉ, OBLIGATION OU OPTION ? ETAT DE L'ART JURIDIQUE 24
- POURQUOI L'IA NE PEUT AVOIR NI EMPATHIE, NI CRÉATIVITÉ NI CONSCIENCE 25
- ENJEUX ET DEFIS DU DÉPLOIEMENT DE L'IA DANS LA SANTÉ 28
- L'IA SERA-T-ELLE CAPABLE DE FAIRE DES « EXPÉRIENCES DE PENSÉE » ? 32

- L'IA va-t-elle modifier la notion d'explication ?

Alexei Grinbaum, du CEA IRFU

En partant de plusieurs photos (un homme qui refuse de serrer la main à un robot, une vieille dame qui joue avec un robot) Alexei Grinbaum pose la question de la prédiction : « *L'homme refuse-t-il car il ne sait pas prédire ce que va faire le robot ?* », mais aussi de l'imitation, « *qui imite qui sur cette image ?* », il y a un problème d'imitation mutuelle, la machine apprend sur les données qui va reproduire ce que fait le robot. Donc il y a beaucoup de notions abstraites derrière l'IA, qui définissent à un certain point la question de l'éthique.

Ceci est d'autant plus important que différents agents tournent autour de la notion d'IA. Entre les designers, les programmeurs, les entraîneurs, les opérateurs, les utilisateurs… Qui est responsable ? On peut ainsi prendre la question de l'éthique de l'IA sous plusieurs angles, « *et un des angles est de cartographier les choses en parlant des valeurs* », explique le chercheur. Sécurité, transparence… La loyauté, par exemple, est une valeur qui a un sens différent pour l'IA que pour les autres sytèmes.

Le chercheur donne ainsi plusieurs exemples : comme Lee Sedol, qui a perdu face à Alphago. Ou avec la reconnaissance faciale, dont les valeurs peuvent dépendre de la culture, des pays, et cela a une incidence in fine sur l'explicabilité. Par exemple, la manière dont on caractérise les visages dépend de la langue par exemple, on n'a pas les mêmes mots, ni le même nombre de mots pour définir les caractéristiques des visages. De même, si l'on parle à la police, l'explicabilité ici, n'est pas un souci. La police veut juste savoir si la personne a été reconnue. Elle s'arrête là. Mais dans le cadre de la médecine, par exemple, il faut aller plus loin. Un médecin ne peut pas se limiter à dire : « là vous avez du rouge ». « *Le même algorithme pour la police ou le médecin, n'a pas besoin d'un même niveau d'explicabilité. Car dans son interaction, le médecin a besoin d'aller plus profondément dans le langage, pour comprendre les défaillances possibles.* » L'explicabilité n'est pas la même selon l'utilisateurs, ou alors où l'on se situe, en Californie, en Russie ou en Chine.

Par l'exemple du film Ex Machina ou avec des chatbots, Alexei Grinbaum, expose des différences de sauts de langages d'éléments inhumains par apprentissage par renforcement. Où à un moment, les robots ne comprennent pas. Car apprendre sans comprendre introduit de l'inhumain dans le comportement des machines. Donc l'explication est au centre des débats. Ce n'est pas le seul problème, mais il est au centre.

Pourquoi est-ce important d'expliquer ?

Ce n'est pas une question de prédiction du comportement, ni pour comprendre une panne que l'explicabilité est nécessaire, mais c'est bien pour la confiance ou la communication sociale. « *Le médecin a besoin d'interagir avec le patient, donc besoin d'expliquer pourquoi. Cela crée le besoin d'ouvrir la boîte noire. Et pourtant ce n'est pas actuellement possible. Aujourd'hui pour un agent conversationnel, les espaces de phases atteignent 1,5 milliards de paramètres, ce n'est pas possible de tout expliquer* », ajoute-t-il.

Ouvrir les boîtes noires veut dire raconter une histoire. On a besoin d'expliquer pourquoi. On ne peut pas être tenu responsable si on n'a pas de récit sur un système.

Or, selon lui, nous n'avons pas besoin d'aller trop loin dans l'explication scientifique. Cela dépend de l'utilisateur. On peut facilement mettre en place des outils heuristiques dépendant de l'auditeurs qui permettent d'expliquer « un peu », la boîte noire, de façon satisfaisante. Arrivent alors deux autres questions, celle de la causalité et de la contrefactualité. *« L'utilisateur va spontanément chercher des phénomènes de causalité. L'utilisateur va très souvent chercher l'explicabilité des résultats, mais il ne peut expliquer les résultats qu'à travers des chaînes de causalité.* » Elle ne sera pas forcément du niveau de complexité du fonctionnement des systèmes, mais se fait souvent à travers la contrefactualité. En effet, un ensemble d'énoncés contrefactuels suffisent à expliquer, dans le langage et dans le contexte de l'utilisateur. Alexei Grinbaum revient alors sur le dilemme du tramway. Voir la vidéo ici.
La réponse à ce dilemme dépend des cultures sociales : aux Etats-Unis, en Asie, les réponses ne sont pas les mêmes. La sociologie des victimes du tramway est donc à prendre en compte.

Par l'histoire d'Achan dans un récit religieux sur la Terre promise, le chercheur établit ensuite l'importance de la rétropropagation de l'explication dans le récit. Quand on explique a posteriori. Les enjeux, forcément parfois un peu faux, car simplifiés, peuvent venir s'interposer dès le début du récit, pour l'utilisateur, même si dans la temporalité, le scientifique peut comprendre les choses a posteriori. C'est ici un des enjeux de l'explicabilité. On n'a pas besoin d'une explication complètement vraie. La rétropropagation évacue la question de la vérité, pour favoriser la question de la confiance. *« Cela a été très bien compris par l'Union Européenne, qui parle de Trustworthy AI », explicite le chercheur.*

Ainsi, la vérité n'est pas l'enjeu de l'explicabilité, L'enjeu, c'est la confiance.

- « Épistémologie des modèles et demandes d'explicabilité pour l'apprentissage machine »

Franck Varenne, Université de Rouen

Pourquoi l'explicabilité ? Il y a en ce moment un, demande d'explicabilité notamment en apprentissage machine, pour comprendre et savoir si c'est un effet de mode ou non. Et l'épistémologie se penche en ce moment sur le sujet. Mais ici, la demande d'explicabilité vient de nombreux acteurs. D'une part car à toute nouveauté technologique s'adjoint un certain nombre de questions éthiques, donc le régulateur s'interroge, mais aussi sur ce sujet, de la part des ingénieurs et des développeurs, car le manque d'explicabilité pose des problèmes d'acceptabilité technique. Et enfin le sujet pose des questions pour la recherche académique, car cela pourra bloquer si on n'est plus capable d'expliquer. La science nait de la science. Moins d'explicabilité, c'est moins de questions posées. Serait-ce la fin de la recherche académique ?

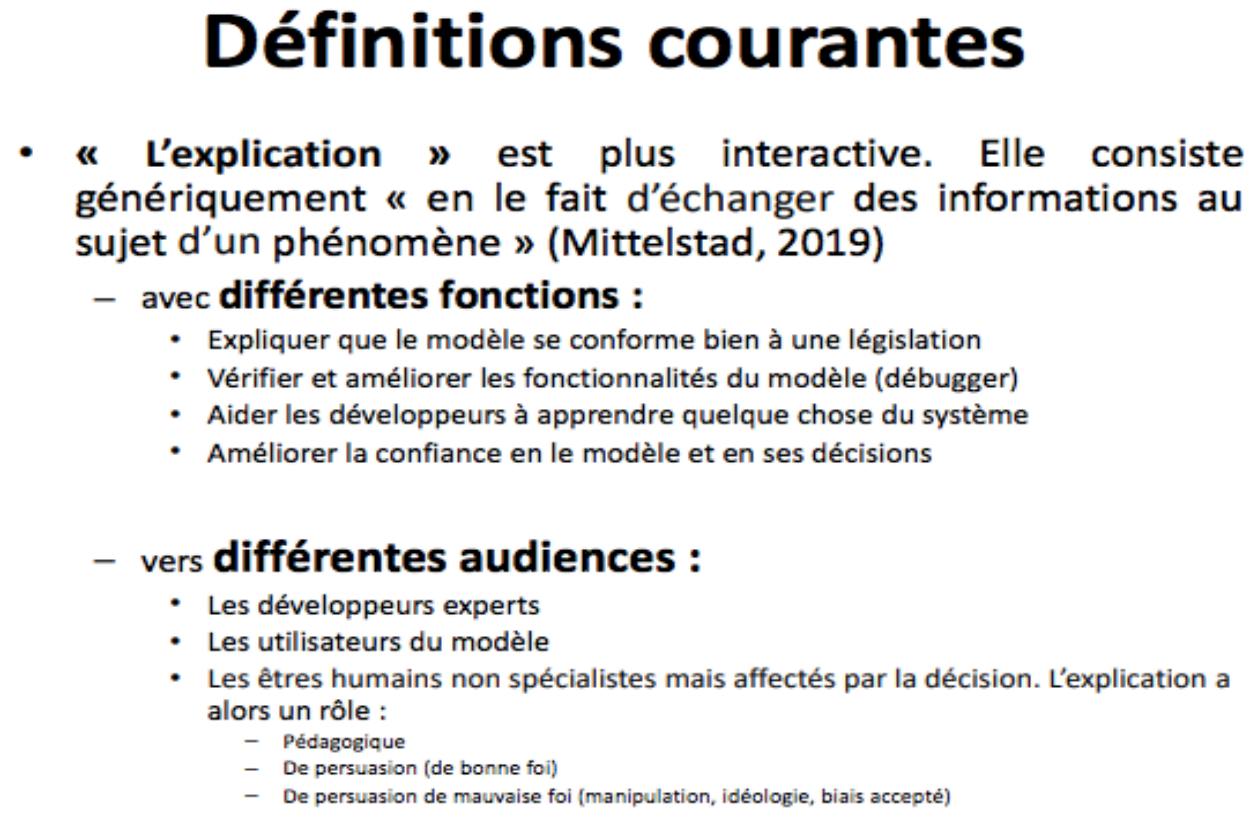
Les définitions

Mittelstadt a défini l'interprétabilité comme « *la chose qui réfère au degré de compréhensibilité humaine d'un modèle de type boîte noire ou d'une décision* ». Franck Varenne bondit ! Car Mittelstadt mélange beaucoup de choses sans les définir. Il fait appel à la notion de compréhension, ou de transparence, par exemple, sans aucune précision.

Revenant sur l'article de Mittelstadt, Franck Varenne expose : cet article se concentre ensuite sur « *l'explication interprétable [compréhensible] post-hoc* » et constate que les explications post-hoc par modélisation compréhensive (simplifiante) de modèle par apprentissage machine ne sont pas des explications fiables.

On se rend compte alors que l'explication a essentiellement une fonction de persuasion, avec ses limites rhétoriques, sa manipulation et ses biais : « *si vous cachez des biais, vous êtes dans la manipulation.* » Donc c'est la question de l'audience et de qui a intérêt à dire les choses.

Mittelstadt conclut en affirmant que si l'on ne peut pas expliquer le détail des modèles, on ne se contente que des grands gestes, des grandes lignes, et donc la fiabilité de ces explications de modèles est discutable.



Beaucoup de confusion dans toutes les définitions ! *« En outre, le terme interprétation lui-même est vague : cela est dû au fait que l'on conditionne dès le début toute interprétation à une compréhension humaine alors que c'est l'inverse qui est le plus vraisemblable »*, ajoute le chercheur. Donc Franck Varenne et Christophe Denis ont retravaillé les définitions :

Interprétabilité d'un modèle : « propriété qu'a un modèle de se voir composé d'élément (signes, symboles, figures, concepts, données, etc.) qui ont chacun un sens [c'est-à-dire un référent possible] pour un sujet humain »

Explicabilité (de l'algorithme à AM ou des sorties de l'AM) : « capacité de déploiement et d'explicitation de cet algorithme ou de ses sorties en séries d'étapes reliées entre elles par ce qu'un être humain peut interpréter sensément comme des causes ou des raisons »

Compréhension d'un phénomène, ou d'un calcul (cum-prehendere) : il y a compréhension d'un phénomène quand notre esprit dispose de la possibilité d'en unifier les manifestations successives ou diverses sous une représentation unique et aisée à concevoir.

Franck Varenne revient ensuite sur les différents types de modèles. Parmi les 21 fonctions de modèles qui existent, il en retient quatre qu'il détaille :

– l'analyse ou la réduction de données
– la description

– la prédiction
– l'explication

La présentation est visible ici.

Modèles explicables

Dans une troisième partie de son exposé, l'épistémologue revient sur les différences entre les modèles expliqués et modèles explicables. *« En physique, certains modèles sont explicables par l'explication du système cible qu'ils modélisent. C'est le cas pour la chute des corps par exemple. Mais ce n'est pas toujours le cas, et plus particulièrement, ce n'est pas le cas dans les algorithmes par apprentissage machine »*, explicite le chercheur. Par exemple, on ne sait pas pourquoi on fait la différence entre un chat et un chien, et finalement, on ne sait pas pourquoi la machine le fait non plus. En IA symbolique, en revanche, c'est comme un modèle scientifique, on a à la fois de l'explicant et de l'explicable.

« Cela veut dire que le processus de computation suivi par le modèle implémenté dans le programme est également interprétable et explicable en lui-même. Il est interprétable car l'ontologie du modèle renvoie à des ensembles d'entités et de propriétés reconnues comme existant réellement dans le système cible (sémantique référentielle) auquel on a accès par ailleurs sous une forme interprétable (sémantique cognitive) »

Cas particulier des modèles à apprentissage machine

L'explicabilité du modèle que l'on recherche doit être fondée autrement pour deux raisons : il ne représente pas forcément un scénario causal d'interactions, et surtout, l'ontologie sous-jacente au modèle et basée sur les données et leur structure, or cette structure peut être complètement inconnue ou fictionnelle.

« Dans le cas des algorithmes à réseaux de neurones, le système met en place des modèles non linéaires reliant les valeurs prédictives et les valeurs prédites : les valeurs prédictives interagissent fortement, donc on ne peut plus parler de simples corrélations », explique le chercheur. *« Et dans le cas de modèle non linéaire à arbres de décision, les étapes élémentaires restent interprétables une à une, mais le processus d'ensemble n'est pas pour autant aisément sensément résumable : il n'est pas compréhensible »* Dans le cas des réseaux de neurones, on met donc en place des « paris métaphysiques sous-jacents », qui sont des hypothèses métaphysiques minimalistes sur le fonctionnement même des algorithmes. (Voir le rapprochement récent entre l'analyse par ondelettes et les réseaux de neurones convolutionnels: Stéphane Mallat, "Understanding deep convolutional networks". Phil. Trans. R. Soc, 2016.)

Ainsi, les modèles par apprentissage machine ne peuvent pas hériter directement leur interprétabilité et leur explicabilité du caractère réaliste et causal des interactions qu'ils modélisent dans leur calcul. Ce défaut fragilise les pratiques de vérification, de validation et de diffusion voire d'appropriation par les utilisateurs. Il remet en cause la confiance dont on parlait plus haut.

Franck Varenne revient ensuite sur les demandes qui émergent.
« Dans la demande d'explicabilité, il y entre en fait souvent aussi une demande de compréhensibilité du modèle. C'est cela qu'on appelle XAI : eXplanable AI. En plus de son explication, on recherche des grands principes unificateurs permettant de penser et représenter de manière unitaire le fonctionnement global, la logique globale du modèle. »

En conclusion :
Outre différentes définitions reprises dans sa thèse, Franck Varenne et Christophe Denis estiment que par leur approche « signal » et le faible rôle donné à la sémantique, les modèles prédictifs à AM

ne prétendent pas représenter de causalité et s'apparentent aux modèles d'analyse de données. Or les modèles à analyse de données classiques reposent sur des hypothèses métaphysiques minimales concernant (les contraintes pesant sur) la structure des signaux qu'ils prennent en compte. Ainsi, ils suggèrent que certaines techniques à apprentissage machine ne semblent pas reposer sur des hypothèses de structure du signal qui soient aussi claires ni interprétables pour nous à l'heure actuelle.
En outre, l'absence de représentation d'une causalité reste à l'origine des points de fragilité de l'apprentissage machine déjà signalés dans la littérature.

Parmi les questions du public, une personne interroge le chercheur sur la traduction de ce modèle explicatif. A qui incombe cette tâche ? Le Développeur, le *data analyst* ? le designer ?
Selon le chercheur, c'est un métier à inventer. Toute la question de l'informatique est qu'elle procède par feuilletage, il faudrait ainsi créer de l'épistémologie appliquée en entreprise.

- Explicabilité et éthique d'applications IA dans le domaine médical

Christophe DENIS et Judith NICOGOSSIAN, Anthropologue.
Pourquoi de l'IA dans la santé ? On est en explosion démographique. Donc face à la croissance de l'espérance de vie, comment faire pour que ces années rajoutées soient des années en bonne santé ? Et autre point, comment soigner ce nombre grandissant de personnes, en réduisant les coûts, et en économisant les ressources.
Le but est d'aller vers une médecine 4P: prédictive (exemple par la génétique), préventive, personnalisée (adaptée au métabolisme, ou à la culture/croyance) et participative, car si le patient est acteur de son parcours de soin, la médecine est plus efficace.
L'IA est le bras armé de la médecine numérique : utilisation des réseaux de neurones pour reconnaissance d'image, éviter les effets indésirables de certains médicaments (faire des simulations…).

Christophe Denis redéfinit alors la santé selon l'OMS, comme un « *état de complet de bien-être physique, mental et social. Il ne consiste pas seulement en une absence de maladie ou d'infirmité* ». En 2006, cette définition de 1946 a été complétée par le fait qu'une opinion publique éclairée et une coopération active de la part du public sont d'une importance capitale pour l'amélioration de la santé des populations D'où l'importance de l'explicabilité de l'IA si elle est vouée à jouer un rôle plus important dans la médecine.

Ivan Illich est lui allé plus loin en affirmant que la colonisation médicale de la vie quotidienne alliène les moyens de soins. l'idée c'est de placer l'humain au cœur des dispositifs, notamment pour favoriser l'autonomie des populations.

On n'a pas attendu les réseaux de neurones pour mettre de l'IA dans la médecine, déjà en 1970 existait un système, le « Rule-Based Expert System » qui permettait d'adapter la posologie des médicaments en fonction des patients. *« Ce système n'a jamais été utilisé, non pas par soucis de performance, mais par une question d'acceptabilité. C était compliqué, cela posait des questions juridiques aussi, et l'acceptabilité du public n'était pas au rendez-vous »*.

Autre exemple, en 1979, Eliza, par Joseph Weizenbaum, était une sorte de chatbot qui simulait une relation entre un psychanalyste et son patient. Basé sur un fonctionnement simple qui reformulait les questions, et Eliza répondait « I understand », lorsqu'elle ne comprenait plus rien.
Déjà des utilisateurs devenaient dépendants vis-à-vis de cela. Il créait un effet de dépendance, appelé Eliza Effect.
Ces exemples indiquent le besoin d'établir une forme de communication particulière entre l'intelligence artificielle et l'humain.

Christophe Denis établit ainsi les talons d'Achille de l'IA : explicabilité, éthique, la prise en compte de cas nouveaux et exceptionnels, et enfin les facteurs humains et l'impact sur le vivant.

L'autre question, c'est l'autonomie du médecin : face à un système fiable à 99,99%, quelle légitimité aura-t-il à dire et verbaliser son doute ? Cela pose des questions d'ordre juridique.

Cependant, rien ne sert de sauter sur sa chaise en criant *« L'éthique de l'IA, l'éthique de l'IA, l'éthique de l'IA ! »*. Christophe Denis tient d'ailleurs à laver l'amalgame récurrent entre moral, éthique et déontologie. La morale énonce des règles de conduite, l'éthique cherche à établir le fondement de ces règles, et la déontologie dicte les obligations que des individus doivent respecter dans le cadre de leur profession.

Avec le projet moralmachine.mit.edu, disponible sur internet, où le but est de récupérer énormément de données sur le dilemme du tramway, on se rend compte qu'il est très compliqué de mettre en place un système moral sur une machine. D'une part car cette morale devrait dépendre géographiquement de là où la machine est utilisée, et d'autre part, de manière fonctionnelle, *« car comme on est sur une machine apprenante, on ne peut pas garantir l'établissement d'un système moral sur une machine »* précise le chercheur. Une manière de briser cette symétrie humain/machine, serait d'établir et d'injecter du hasard, comme l'expliquait Alexei Grinbaum. Cela rejoint des travaux en cours, où le but est de ne pas faire croire qu'il y a réciprocité ou symétrie dans les applications.
Concernant l'explicabilité, Christophe Denis pose la question de la transparence, et met en garde sur le fait que la transparence peut aussi avoir des écueils (voir le livre : *Tyrannie de la transparence)*.

L'explicabilité :

Christophe Denis revient ensuite sur la notion d'explicabilité et souligne l'importance mais aussi l'ambiguïté du fait que l'explication fournie doit être adapté à la connaissance de l'utilisateur et la finalité de l'utilisation de l'application. Il faut alors trouver un compromis. La transparence doit ainsi, elle aussi, varier en fonction du contexte.

En reprenant ensuite certaines conclusions des travaux de l'école de Palo Alto, il souligne que l'IA et notamment les systèmes communicants de type Chatbots, ne sont pas dans un système linguistique précis. Certains humains veulent être trompés par leur sens. Et un système d'IA, énonciateur, n'est pas sujet de l'énoncé. Cela pose des questions, car l'utilisateur est ainsi coupé de la communication non verbale, car le système n'a pas de finalité propre. Peut-il recréer cette communication non verbale ?
Il termine alors en reprenant Lévinas: la relation est asymétrique à autrui, je ne dois pas attendre de réciprocité. Le moi, devant autrui, est infiniment responsable.

En conclusion Christophe Denis revient sur la « convivialité » de l'IA en citant Ivan Illich. « *L'outil reste convivial dans la mesure où chacun peut l'utiliser sans difficulté, aussi souvent qu'il le désire. Il ne nécessite pas de diplôme pour s'en servir et doit répondre à trois critère s: il doit être générateur d'efficience sans dégrader l'autonomie personnelle, il ne suscite ni esclave ni maître, et il élargit le rayon d'action personnel.* »

Dans les questions du public, un membre du groupe Mérieux, estime qu'à ce jour, lorsqu'on parle d'IA avec les professionnels de santé, ils ne parlent pas du tout de ces priorités. Les professionnels de santé attendent autre chose des technologies, leur discours ne se concentrent pas sur ce qui a été présenté. Et l'auditeur incite le monde académique à aller voir davantage de professionnels de santé pour se reconnecter à leur besoin propre.

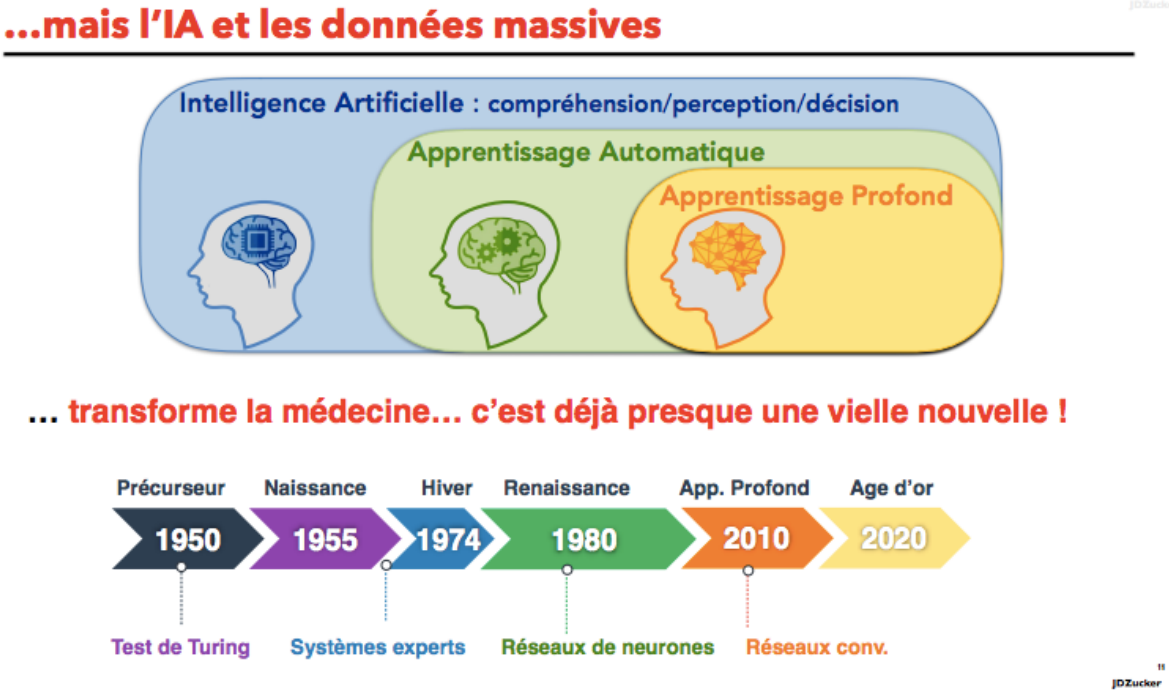
- Apprentissage Artificiel et Interprétabilité pour la médecine de précision

Jean-Daniel Zucker, DR IRD

Jean Daniel Zucker commence en introduction par un hommage à Jacques Pitrat, pionnier de l'intelligence artificielle symbolique qui est décédé le 14 octobre.

Il commence par définir l'interprétabilité: le degré avec lequel un utilisateur de la machine peut faire sens des décisions qui ont été prises.

Dans un premier temps, le chercheur veut rappeler que la médecine est imprécise, c'est à dire que bons nombres de patients, avec un médicament ont au mieux un bénéfice, parfois pas de bénéfice, et au pire, rien. Et dans quelle mesure, on peut tenter d'établir une médecine de la précision. En passant par une meilleure stratification des individus, pour mieux adapter les soins. Parlons davantage de précision, plutôt que de personnalisation, car les médecins ont toujours réalisé de la médecine personnalisée. Mais en analysant davantage la génétique des individus, nous pouvons aller plus en détail. Il rappelle sur ce point le livre « 10% human » (ce serait davantage 50%) qui établit que nous ne sommes composés de millions de gènes bactériens dans notre métagénome. Mais que nous sommes plus complexes que simplement « humains ». Il rappelle ensuite que le lien entre intelligence artificielle est profond et ne date pas d'hier, mais que les systèmes à apprentissage profond amène un nouvel âge d'or dans la transformation de médecine.



Il précise même que le médecin et l'IA atteignent un meilleur pourcentage de réussite de diagnostic que l'IA seul ou le médecin seul. *« Mais tous les jours que ce soit en dermatologie, en ophtalmologie, gastroentérologie, cardiologie, des cas où les systèmes à apprentissage profond, grâce à sa capacité de reconnaissance de forme, améliore considérablement les performances des médecins ».*

Selon le livre d'Eric Topol, *Deep Medicine,* le but n'est pas de supprimer les médecins, mais de leur libérer du temps pour passer plus de moments avec leurs patients. En supprimant des tâches plus mécaniques, ou qui ne font appel qu'à des capacités mémorielles.

L'interprétabilité

Le droit, et notamment le RGPD, indique qu'une prise de décision algorithmique doit être accompagnée d'un « droit d'explication ». La notion d'explicabilité est donc un impératif légal dès lors qu'un algorithme, quel qu'il soit, entre en considération dans une prise de décision.
En outre, une prédiction explicable est souhaitée par les médecins, à partir du moment où un modèle doit être validé avant d'être déployé en routine. C'est le principe de la confiance. On s'aperçoit alors que cette confiance dépend du contexte. Ce n'est pas la même chose pour analyser un cheveux que de faire fonctionner des machines de réanimation.

La question des biais des données est également évoquée, surtout lorsque par exemple des algorithmes de classification ont davantage été entraîné sur des hommes blancs que sur des femmes à la peau foncée. *« Cela pose une injustice flagrante, et il y a beaucoup de travaux à l'heure actuelle pour développer des algorithmes* fair » précise le chercheur.
Autre limite : les attaques par pixel. En modifiant un pixel dans mon image, on trouble fortement l'algorithme, qui peut reconnaitre une grenouille dans une voiture. Sinon les attaques adversariales, où en rajoutant du bruit sur l'image, un panneau stop peut être interprété par l'IA comme un panneau « cédez le passage » … Ce qui est une erreur gravissime.

De même ces attaques peuvent aussi avoir lieu en médecine et tromper le diagnostic, avec pourtant une probabilité de 100%.

En IA, on doit trouver un compromis entre la précision et l'interprétabilité. C'est ce qui nous amène à la notion de boîte noire, car les arbres de décisions, moins précis, sont plus facilement interprétables. Et c'est l'inverse avec les réseaux de neurones.

Comprendre l'interprétabilité ?
Le chercheur expose ensuite les différents moyens de comprendre les boîtes noires. Cela peut être fait a posteriori. Ou encore localement par des perturbations. La présentation de ces méthodes est disponible ici.

Ensuite, en détaillant une technique de diagnostic pour la cirrhose du foie, le chercheur arrive à démontrer que l'on peut avoir des modèles interprétables, en connaissant le nombre de données qui ont utilisés pour réaliser le diagnostic, sur différentes espèces de bactéries.
Idem, le chercheur passe rapidement sur un modèle d'optimisation clinique des résultats après des opérations de chirurgie bariatrique, pour vaincre l'obésité.

Conclusion :
Il existe deux approches sur l'interprétabilité. La première qui dit qu'en médecine et en justice pénale il faut arrêter d'utiliser des boîtes noires. De l'autre, que les boîtes noires (réseaux de neurone principalement) permettent d'optimiser grandement les résultats.
Ainsi les progrès de l'IA et du Deep Learning posent des questions éthiques sur son adoption en médecine : équité/confiance/transparence/interprétabilité. Ainsi, l'IA doit aider les cliniciens à être

plus efficace mais l'interpretabilité joue un rôle clé pour la confiance, la compréhension des biais, et contribuer à la recherche de l'étiologie des maladies. Il faut donc développer les IA explicables (XAI).

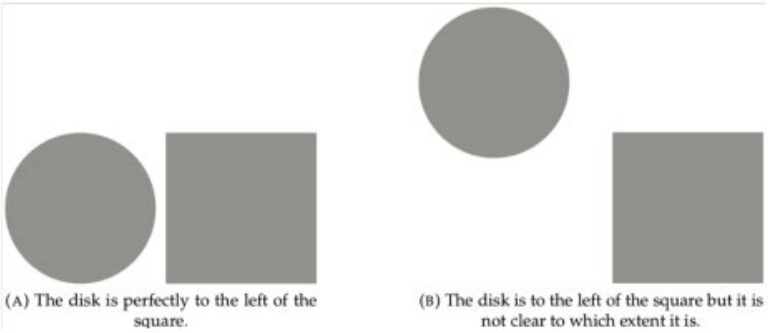
- Interprétation d'images médicales : quelques travaux vers l'explicabilité

Céline Hudelot, Laboratoire MICS - CentraleSupélec

Il existe plusieurs types d'approches dans l'interprétation des images médicales. Les approches de types « symbolique », utilisée dans les années 80-90, basées sur un système d'experts, où le système reproduit leur raisonnement. C'est un système par connaissance explicite. Idem pour les systèmes probabilistes (année 90). A base d'atlas, c'est un modèle statistique construit à partir de connaissances expertes mais dont les paramètres sont appris à partir de données.
Enfin, on retrouve le système « *orienté données* », à partir des années 2000, où c'est un modèle à base d'apprentissage et d'apprentissage profond. Il nécessite des bases d'apprentissage de bonne qualité et de grande taille. Les données font office de connaissance implicite, en revanche les labels (catégories) sont des connaissances explicites.

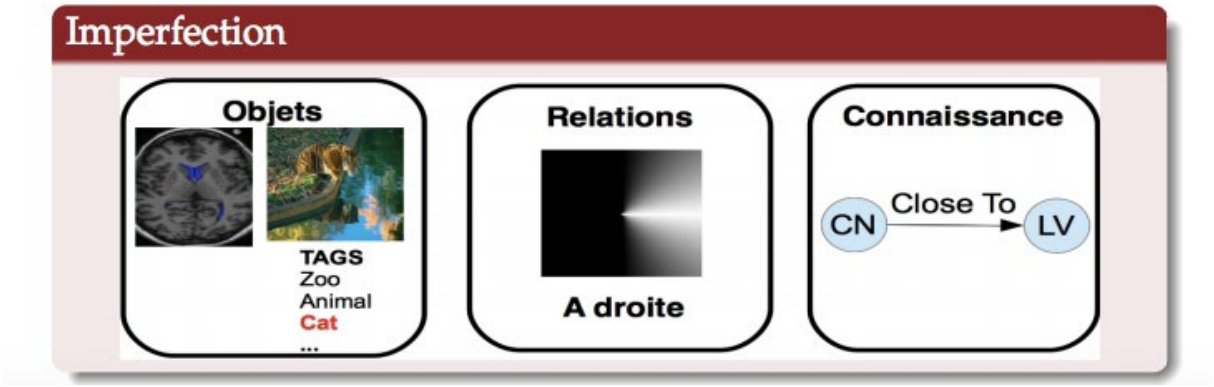
Ces derniers systèmes peuvent apporter des succès. Par exemple, dans une étude parue dans Nature en 2017, une équipe a pu classer les cancers de la peau par reconnaissance d'image basée sur de l'apprentissage profond. Cependant, cela nécessite d'avoir des données labellisées de bonne qualité. L'expérience s'est basée sur 129450 images labellisées selon 2032 classes. Cela demande donc au préalable l'apport de connaissances expertes avec une taxonomie hiérarchique des pathologies et la construction du jeu de données de départ. « *En outre, cette expérience est discutable car il n'y a pas de collaboration entre les experts. Le jeu de données est constitué de manière individuelle et l'on prend ensuite la moyenne des prédictions* », estime la chercheuse de CentraleSupélec. Tout ceci est, pour elle, très discutable.

Ce sont donc des approches de types « boîtes noires », incapables d'expliquer leurs décisions de manière compréhensible et reconstructible par un humain. Et qu'on peut facilement tromper ou biaiser leur résultat. Et si leur résultat se trompe, nous ne sommes pas capable de comprendre pourquoi. Cela pose un souci dans la nature même de la tâche qui est une prise de décision. « *Au-delà de la sortie d'une prédiction, nous attendons des éléments de preuves, explicatifs* », indique Céline Hudelot. Elle revient ensuite sur les approches de l'interprétabilité. Soit on construit dès le départ des modèles interprétable, soit on passe par des systèmes de types boîtes noires, au sein desquels on va chercher à expliquer par la suite.



(A) The disk is perfectly to the left of the square.

(B) The disk is to the left of the square but it is not clear to which extent it is.



La chercheuse se concentre avant tout sur l'explication, c'est à dire trouver des éléments textuels qui ont permis au système de converger vers cette prise de décision.
Ce qui ressort des travaux effectués jusqu'à ce jour, c'est l'énorme importance de la connaissance experte à l'origine du système. Le libellé des images par les modèles, par exemple. *« Il y a un moment, où on a demandé à des médecins de noter les images, de les interpréter pour construire le modèle de départ »*, raconte la chercheuse.

Un des facteurs importants dans la reconnaissance d'image consiste en la reconnaissance de relations spatiale entre des objets. *« Cela se base sur des travaux cognitifs, ou pour reconnaître des éléments, nous procédons par analyse des liens spatiaux entre différentes parties de ces objets »*, explique-t-elle. Cette relation est stable, et nous permet de prendre en compte les imperfections de l'interprétation. Elle donne alors un exemple avec un rond à gauche d'un cercle. Dans l'analyse des relations spatiales, il est possible d'établir des relations de degrés. On parle alors de relations floues.

Céline Hudelot expose alors les travaux qu'elle a pu mener pour construire des relations spatiales « floues » entre différents objets sur des images médicales. Elle est partie pour cela d'une hypothèse : la pertinence de la relation spatiale est sujette à la fréquence de cette relation. Cela lui permet alors d'extraire de ses analyses les relations spatiales pertinentes sur les images.
Elle expose alors les outils algorithmiques qui permettent de faire ces extractions. La vidéo est disponible ici.

En conclusion, la chercheuse insiste sur le fait que quel que soit l'approche dans la recherche d'interprétabilité en IA, la connaissance experte est fondamentale sur la base de données de départ, et souligne aussi l'importance du raisonnement. Notamment en imagerie médicale, où les analyses sont aujourd'hui faites sur la base de discussions entre experts.

- L'explicabilité, LA clé pour le déploiement d'IAs dans l'entreprise

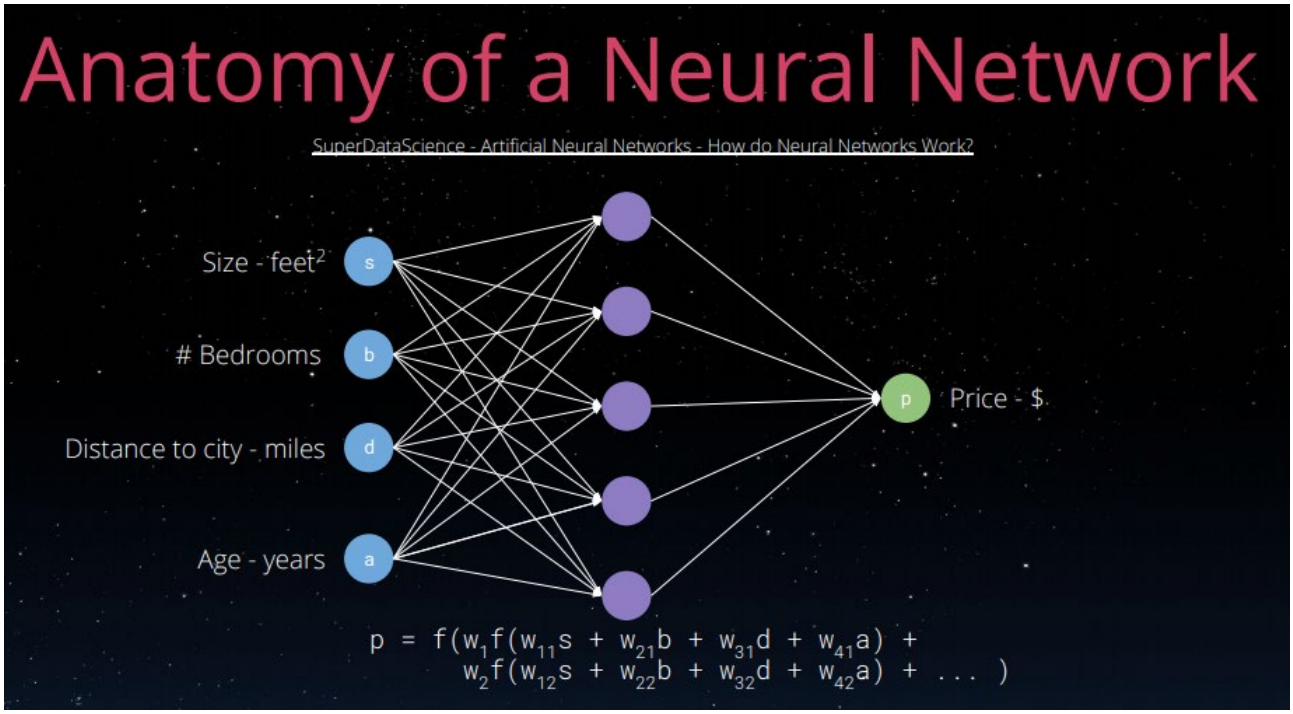
Clodéric Mars, Fondateur de Craft IA

La vidéo de la présentation est disponible ici

Le principe de Craft IA, est d'aider les entreprises à développer des IA explicables, et à mettre en place de l'explicabilité dans leurs projets d'IA.

Clodéric Mars reprend rapidement les bases de l'explicabilité : répondre à la question du pourquoi, dans la communication, l'importance du contrefactuel (pourquoi le système n'a pas fait cela ?) et enfin, le fait qu'il faille adapter l'explication à l'utilisateur.

Le grand souci provient de l'utilisation de réseaux de neurones, qui sont des fonctions de combinaisons linéaires de fonctions, qui sont elles-mêmes des combinaisons linéaires de fonctions. *« On mélange ainsi tous les paramètres, et il devient alors très difficile de savoir quel paramètre est imbriqué*



avec quel paramètre », décrit le chercheur.

Pourquoi expliquer en entreprise ?

Dans un premier temps, mettre en place des IA explicables permet d'augmenter l'acceptabilité de ces nouvelles technologies. *« Les employés vont avoir peur. Soit, ils vont craindre pour leur emploi, et se dire qu'on va les remplacer, ou alors ils ne croient pas à l'outil, dans le sens où ils ont appris leur métier, et c'est impossible qu'une machine le fasse à leur place »*, analyse l'entrepreneur. Donc il faut embarquer les gens dans la construction d'une IA explicable, pour mieux démystifier la technologie, et leur permettre de comprendre ce que fait réellement l'IA. Ils vont pouvoir faire des remarques, et vont

apporter de la valeur à la construction de l'IA également. *« Pour cela il existe des modèles, des outils : des visualisations, des modèles d'exploration ou de simulation »,* précise Clodéric Mars.
Dans les modèles Seq2seq par exemple, qui sont des algorithmes qui permettent de travailler sur des séquences de mots, très utilisés en traduction, des outils existent pour « ouvrir le capot » et tenter de comprendre pourquoi telle séquence a été rapprochée de telle séquence. Ces outils sont utilisables par des personnes qui n'ont pas forcément d'expérience pointue dans le domaine.

L'autre point rejoint la confiance. Les systèmes parfois, se trompent complètement. Le chercheur donne l'exemple de son chauffe-eau « intelligent », censé optimiser sa consommation d'eau, qui parfois se vautre complètement. *« Si l'on sait dans quel cas le système a tendance à se tromper, je peux mettre en place des garde-fous pour vérifier ou contrebalancer ces écarts de fonctionnement »*, précise-t-il. Ou encore d'expliquer pourquoi il s'est trompé, ce qui en termes de responsabilité est important.

L'explication peut porter sur tel ou tel paramètre. C'est ce qui est utilisé par la société Backwen, qui cherche à faire de la détection de fraude, pour tenter de savoir l'impact d'un paramètre sur la prise de décision. Un outil permet de déterminer et de comprendre quel paramètre a davantage compté dans la prise de décision, et donc de mieux cerner celle-ci.

Chez Dalkia, dans un outil d'aide au diagnostic, l'algorithme indique quel paramètre a été pris en compte. *« Ce qui est intéressant, c'est que l'explicabilité n'a pas valeur de confiance mais de productivité. Comme l'outil est un assistant pour l'expert, si ce dernier n'a pas d'explication, il doit refaire tout son travail, et donc l'outil n'aide à rien. Il doit tout reprendre »,* explique le chef d'entreprise.

En outre, l'explicabilité peut permettre au système de collaborer avec d'autre système, pour certifier une prise de décision par exemple. *« Si on a une modélisation du fonctionnement de l'algorithme, on peut le connecter à un autre outil qui peut travailler et échanger des informations sur lui »* ou même avoir une rétroaction sur le dit système de départ. il donne l'exemple de Total Direct Energie, qui utilise ce type de système sur la consommation des foyers.

Clodéric Mars revient ensuite sur l'évaluation des systèmes explicables, qui reste encore un champ complexe de la recherche.

En conclusion, selon l'entrepreneur, il reste encore beaucoup de choses à faire. Il relativise l'approche dualiste de l'interprétabilité, en « Boîte noire » ou en modèle 100% interprétable. Mais pour sa part, il existe une troisième voie qui est une approche hybride, dans laquelle on peut laisser un algorithme non explicable faire ses tâches (pour détecter quelque chose sur une image par exemple) et faire travailler ensuite un algorithme qui, lui, sera explicable et fera le lien entre les choses détectées.

Conclusion:
Selon lui, l'explicabilité n'est pas une contrainte. Au contraire, elle va permettre de mettre en place de meilleures IA. D'après son expérience très peu d'entreprises sont capables de mettre en production des IA qui n'ont pas dépassé le deuxième niveau d'explicabilité, c'est à dire une compréhension détaillée du fonctionnement du système.

- Les limites des algorithmes de Deep Learning dans le domaine de l'imagerie médicale

Issam IBNOUHSEIN Quantmetry

La présentation est disponible ici.

Issam Ibnouhsein introduit sa présentation en recadrant la perception que l'on peut avoir des algorithmes. Il ne faut pas faire de confusion entre l'algorithme et le contenu de l'algorithme lui-même, et ce qui en est fait. D'une part, c'est plus souvent la fabrication de l'algorithme que l'algorithme lui-même qui doit être discuté, ou encore ce que l'on fait de la sortie de l'algorithme, et l'importance que l'on donne à la décision prise par l'algorithme. Se pose la question de la transparence. Donner les lignes de codes d'un algorithme est-il suffisant en termes de transparence ?

Deux types de propriétés peuvent être définies dans la transparence.

Une famille de propriétés normatives extrinsèques :
La loyauté : un algorithme est loyal si la fonctionnalité affichée auprès de l'utilisateur est identique à la fonctionnalité connue du fournisseur
L'équité : un algorithme est équitable si son fonctionnement ne provoque pas d'effets discriminants à l'égard d'une partie de la population.

Une famille de propriétés épistémiques intrinsèques :
L'intelligibilité : un algorithme est intelligible s'il est possible de comprendre son comportement dans l'état de l'art scientifique.
L'explicabilité : un algorithme est explicable s'il est possible de faire comprendre son fonctionnement à un utilisateur (sans expertise scientifique).

Parmi toute ces notions, la plus fondamentale est l'intelligibilité, c'est sur elle que se base ensuite les autres.

Il faut en outre faire la distinction entre l'intelligibilité de la procédure et des sorties. « *On peut bien comprendre une procédure, d'un point mathématique ou intuitif, sans bien comprendre une sortie donnée* ». La loi reste encore floue sur ce point.

Pour communiquer auprès du public, c'est l'intelligibilité des sorties qui est le plus important.

Dans les procédures bureaucratiques, les décisions sont compositionnelles, c'est à dire qu'elles se découpent en paramètres simples. « *A la Caf lorsqu'on vous refuse une aide, on vous dit que votre salaire est supérieur au critère, mais on ne vous refait pas l'historique des paramètres juridiques: citoyen européen, majeur etc.* », explicite l'entrepreneur. Donc l'avantage des procédures bureaucratique est qu'elles peuvent être bien expliquées auprès des employés et du grand public.
D'ailleurs, la compréhension de systèmes globaux est parfois compromise. « *Dans les entreprises, certains gros projets informatiques ne sont pas compréhensibles par un être humain, et c'est ce qui est source de lenteur ou de bug dans ces projets* » estime Issam IBNOUHSEIN.

Ainsi, la question n'est pas de savoir si l'on peut faire une explication globale du système, mais davantage des explications brèves et simples de parties du système.

Le chercheur revient ensuite sur l'explication d'un réseau de neurones. La présentation est disponible ici.
La difficulté d'explication vient du fait qu'on ne sait pas forcément dans quelle dimension on est en termes d'utilisation de paramètres (à la différence des arbres de décision par exemple).

Cette non explicabilité est un frein à la progression en entreprise, par exemple, des banques qui ont des algorithmes de « *score crédit* » ont développé des procédures plus puissantes par apprentissage machine, mais ne sont pas capables de les industrialiser, du fait même de sa non explicabilité. Et la direction préfère se référer à des algorithmes moins puissants, mais mieux connus ou explicités.

En médecine, les médecins mettent la société de plus en plus au défi d'expliquer le système derrière.

Conclusion

Il ne faut pas opposer systématiquement AM et autres familles d'algorithmes. Et ainsi ne pas tomber dans une catégorisation simpliste des modèles en « modèles opaques » et « modèles transparents ». Il faut voir des catégorisations plus complexes.

En outre, un travail doit encore être réalisé pour affiner les méthodes d'explicabilité par extrait des sorties de modèles pour pouvoir croître en taille dans les procédures. Enfin, cela reste un champ scientifique jeune et en pleine évolution…

Empathie, conscience et créativité de l'IA

- Explicabilité, obligation ou option ? Etat de l'art juridique

Alain BENSOUSSAN, Cabinet Lexing Alain Bensoussan

L'avocat commence par introduire le fait que l'éthique ne sera sans doute pas suffisante pour réguler le marché. Restant « inframorale », et l'IA étant partout, la société aura besoin d'une assise juridique.

D'un point de vue juridique, l'intelligibilité d'un système se regarde sous deux angles : le mécanisme technique en lui-même, et les décisions qu'il prend (les faux vrais, les vrais faux, et les choix qui sont faits par rapport à la finalité du système). Le juriste s'interroge ainsi sur l'intelligibilité du système, mais aussi de ses conséquences. Il faut alors évaluer l'impact, des décisions prises.

Concernant l'explicabilité, le juriste revient sur le principe de licéité. L'article 22 du RGPD et la loi informatique et liberté de 1979 parle de l'intelligence artificielle. « *Elle est traitée à deux niveaux, cela peut faire deux fois plus de sanctions en cas de manquement!* », ironise-t-il. Les deux lois parlent du principe de transparence (article 5.1 A) et du principe de compréhension (Article 14G). « On est sur du droit de base. Il faut alors indiquer à la personne la logique et les conséquences, donc l'intelligibilité sociale », détaille-t-il. Cet ensemble de règle s'applique à toute l'algorithmie de l'IA.

L'avocat revient ensuite sur l'exemple particulier de la voiture autonome, qu'il indique comme étant un permis de tuer « Elle tue moins que les humains, mais oui, elle tue quand même », philosophe-t-il. La présentation est disponible ici.
Ce qu'il faut retenir, c'est qu'aujourd'hui, au niveau du droit, lorsqu'un conducteur rentre dans une voiture autonome, il n'est plus responsable, en cas de victime. Ce qu'il se passe, c'est que la responsabilité glisse du créateur au certificateur de l'algorithme. « *C'est ce qu'il se passe dans la plupart des pays* », conclut-il. Il revient ensuite sur le principe de loyauté. Dans le cadre de l'intelligibilité, plus c'est intangible, plus il faut fournir des informations sur la méthode, sur la maîtrise par le concepteur, de son algorithme et lutter contre les biais, avec en cas de conséquences graves, une analyse d'impact et un registre d'erreurs. Enfin, le juriste ouvre en estimant qu'il faut passer d'une IA régulée par l'éthique par une IA régulée par le droit, et milite pour l'ouverture d'un commissariat à l'IA.

- Pourquoi l'IA ne peut avoir ni empathie, ni créativité ni conscience

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

Le but n'est pas de parler d'explicabilité, mais de montrer pourquoi l'IA n'a pas d'empathie. Réponse simple : c'est un système basé sur des algorithmes. Mais il faut aller plus loin. Ces programmes informatiques sont nés de la question : qu'est ce qui est calculable. Or Kurt Gödel a démontré l'indécidabilité des mathématiques, donc de la logique, et qu'en réaction à cette démonstration, les mathématiciens de l'époque se sont posés la question de « qu'est ce qui est décidable ? » c'est à dire calculable. La question de la décision, c'est la question de savoir si un énoncé est vrai ou faux. C'est ça la décision.

Alonzo Church et Alan Ruting on répondu chacun de manière différente à cette question. Le premier en affirmant que les fonctions calculables sont les fonctions récursives (qui font appel à des expressions qui se contiennent et dont le calcul va se terminer). Et le deuxième par une métaphore : la machine de Türing. Or ces machines ont une propriété qui est que les séquences de calculs qu'elles mettent en oeuvre se terminent. Tous les ordinateurs sont basés sur le concept de Türing. Donc l'intelligence artificielle est basée sur le modèle de la machine de Türing.

Donc l'IA n'est rien d'autre qu'un système qui est censé faire un calcul qui se termine, qui a une fin. Donc l'intelligence artificielle n'est rien d'autre, in fine, qu'un système voué à faire des calculs finis. Raja Chatila revient alors sur l'origine du mot « Intelligence artificielle », né en 1956, par des chercheurs qui souhaitaient se faire financer un séminaire de deux mois lors du Dartmouth College: John McCarthy, Marvin Minsky, Nathaniel Rochester et Claude Shannon. Le principe était de simuler les caractéristiques de l'apprentissage par des machines. Cela voulait dire que le propre de l'intelligence est d'apprendre, « et les mots « décrit de manière tellement précise que » indiquait qu'ils parlaient d'algorithme ou d'ensemble d'algorithmes », estime le chercheur. Raja Chatila insiste ensuite sur la notion de « simulation » essentielle dans la définition de départ. La tentative consiste à l'origine, à trouver comment ces machines peuvent manipuler le langage.
Donc « l'intelligence artificielle est avant tout un programme de recherche qui fournit des résultats, fort heureusement souvent efficaces, qui nous permettent de les mettre en applications, mais parfois beaucoup moins efficaces », résume le chercheur. Mais l'idée fondamentale est d'estimer que l'on peut réduire tout système d'apprentissage par des algorithmes. Est-ce que c'est vrai ? Je n'en sais rien. Cela n'a pas été démontré.

Le chercheur revient ensuite sur la notion de récompenses dans les chaînes de Markov. Des fonctions récursives, dont la récompense est définie par le créateur de l'algorithme. La présentation détaillée est ici.

Il schématise ensuite les réseaux de neurones comme des fonctions de fonctions de linéraires, dont le but est d'optimiser les poids afin de réduire les erreurs de sorties. « *Le problème de l'explicabilité est ici: je ne sais pas lier la valeur de tel ou tel poids, en fonction de la valeur de sortie, du fait du trop grand nombre de poids* », pointe-t-il. Car l'optimisation est globale et non dans les détails, localement. Or une multitude de systèmes aujourd'hui, qui ne sont pas de l'IA, fonctionnent de la même manière,

sans qu'on puisse expliquer pourquoi, en optimisation globale. Mais le système a été vérifié, testé. Et dans le cas de l'IA nous sommes dans le cas d'un système ouvert, avec une dépendance vis-à-vis des données. Ce qui complique encore plus la donne.

Raja Chatila évoque ensuite de nombreuses erreurs des IA qui prouvent que le système ne comprend rien à ce qu'il fait. Les exemples sont disponibles en vidéo ici.

« Comment faire confiance à un système qui peut commettre des erreurs aussi évidentes ? Pourquoi vous, vous pouvez différencier un chat et un chien sans regarder 1000 photos de chaque ? Car vous, vous le savez. Et comment vous le savez ? Car vous n'êtes pas un système algorithmique » s'interroge-t-il. Il définit ainsi le méta-raisonnement : *« La conscience de soi est fondée sur l'interaction avec le monde et sur une capacité d'auto-évaluation et de méta-raisonnement »*
L'intéraction sensori-motrice est source de projection et connaissance. Cette projection est source d'empathie.

Raja Chatila conclue en rappelant des assertions démystifiant les intelligences artificielles :

« - Les décisions prises par des systèmes informatiques se situent à un niveau calculatoire et sont inscrites dans les algorithmes conçus par les êtres humains.

- *Toute "création" par un système informatique est le résultat d'un processus calculatoire.*
- *Les systèmes d'IA n'ont aucune autonomie de décision. Ils optimisent des fonctions.*
- *Les systèmes d'IA sont syntaxiques : ils n'ont aucune sémantique et ne comprennent pas ce qu'ils font. »*

Selon lui l'IA a une perception faussée dans l'opinion publique du fait de l'utilisation excessive de termes métaphoriques relatifs aux humains, et de l'exagération par les médias des résultats des algorithmes.

Dans le débat qui suit entre eux deux, Alain Bensoussan, commence en parlant de la personnalité juridique. Selon lui, on peut définir une personnalité juridique particulière. *« Si toutes les personnes sont humaines, toutes les personnes en droit ne sont pas des humains »,* estime-t-il. Par exemple, les enfants ou les malades psychiques ne sont pas soumis aux mêmes régimes et aux mêmes responsabilités juridiques, comme les entreprises, par exemple. Ils peuvent ne pas faire de faute. Donc selon lui, on peut définir un cadre spécial pour les intelligences artificielles, sans se demander pour autant si elles ont une conscience.

Selon lui, il y a une différence entre la responsabilité et la faute. Ce n'est pas le même champ. Il faut passer dans le champ des dommages.

L'avocat, en reprenant le rapport de la Cnil, sur la loyauté du fabricant notamment, re-milite pour un commissariat à l'intelligence artificielle.

S'ensuite un débat de position sur la définition de l'algorithme d'apprentissage suite à une question du public sur les algorithmes métiers. Deux positions très arrêtées se font alors face sur le cadre à donner à l'apprentissage.

Ensuite, une question arrive sur le cadre juridique de la ligne 14, système automatisé, apprenant via des mises à jour régulières.

L'avocat parle alors de la définition de l'accidentologie à 10 puissance -8 près. Est-ce raisonnable ?

Tout réside dans le registre des erreurs acceptables ou non. Et tenir le registre de l'ensemble des erreurs et des failles est extrêmement important. Le Cnil d'ailleurs pointe cette caractéristique.

Suite à une question du journaliste, Alain Bensoussan recadre alors son domaine. Selon lui, est-il acceptable de laisser les humains conduire et avoir environ 4000 accidentés par an, quand on sait de manière probabiliste qu'avec une IA vous êtes à moins de 10. C'est donc de l'acceptabilité sociale donc on parle. « *en droit, on n'est pas objectif dans l'approche, mais subjectif dans les risques* », tranche-t-il.

- Enjeux et defis du déploiement de l'IA dans la santé

David Gruson, groupe Jouve, professeur Jacques Lucas, Conseil national de l'ordre des médecins et Delphine Jafaar, du cabinet d'avocats Jafaar.

Le premier intervenant, David Gruson revient d'abord sur l'algorithmique génétique en détaillant le fonctionnement d'un champignon : le blob. Il insiste sur le fait qu'il est intéressant que les algorithmes génétiques se retrouvent aujourd'hui à aider les chercheurs à encore mieux comprendre la génétique. Joli retournement.
En outre, il tempère les progrès de l'IA en estimant que les plus grandes avancées aujourd'hui se limitent pour la plupart à la reconnaissance d'images par apprentissages machines. Pour les autres domaines, la réalité reste pour le moment en deçà des attentes.

Jacques Lucas se voit ensuite poser la question sur la disparition des médecins, suite aux avancées de l'IA. « *Vous faites allusions à divers succès de librairie comme La médecine sans médecin, ou du livre de Laurent Alexandre, La guerre des intelligences, tout cela nourrit des fantasmes et des fantasmagories qui nous éloignent de notre propre sujet. J'ai souvent l'habitude de commencer par une question que posera l'intelligence artificielle et que posent déjà les intelligences artificielles faibles qui est la question de l'acceptabilité sociale.*

En dehors des robots tueurs qui vont se révolter contre leur créateur - le mythe de la création est derrière - on voit bien que les métiers médicaux vont être affectés. » Jacques Lucas parle ici des métiers qui ont recours à l'imagerie : radiologie, dermatologie (pour la reconnaissance de nodule par exemple) ou la cardiologie où l'IA sera une aide aux diagnostics… Mais il évoque aussi en dehors de ces questions les robots sociaux, pour les enfants autistes, ou dans la maladie d'Alzheimer qui sont de véritables plus thérapeutiques. Ces robots qui viennent s'immiscer dans la relation thérapeutique entre le médecin et le patient. « *De combien de radiologues aura-t-on besoin quand l'IA viendra aider au diagnostic ?* » Ces médecins devront apprendre un nouveau métier qui ne consistera pas tant à diagnostiquer le nodule, mais à connaître la conduite à tenir.
Ces médecins devront alors changer leur formation pour s'adapter à ces nouvelles données : ils devront aussi apprendre à travailler avec l'intelligence artificielle. Ce qui remet d'ailleurs en cause le serment d'Hippocrate où les médecins doivent apprendre la médecine aux enfants s'ils désirent l'apprendre et donc transmettre le savoir de génération en génération.
Pour illustrer son propos, Jacques Lucas cite la révolte des canuts, où l'arrivée des métiers à tisser a profondément changé le métier et a amené les tisserands à se révolter. Ce qui, selon le médecin, pose la question de l'acceptabilité sociale. « *Car on ne fait pas le bonheur des gens contre leur propre sentiment,* estime-t-il. *Il y a donc une œuvre pédagogique très importante à effectuer.* »
Cette acceptabilité sociale se pose autant pour les médecins que pour les patients. Car pour faire travailler l'intelligence artificielle, il faut des données médicales qui doivent garder leur caractère secret. « *Une personne se construit aussi et surtout sur des secrets* », estime-t-il.

Peut-on garantir dans une grande masse de données, et dans un travail effectué sur ces données, l'anonymat de la personne qui a été la productrice de ces données. La réponse est évidemment non. « *Donc il doit y avoir des processus pour éviter l'identification, mais vis-à-vis de la population il faut aussi être transparent et utiliser les bons termes* », pointe-t-il. Deux positions antagonistes se font alors face, presque paradoxale: comment garantir le secret médical, tout en utilisant les données médicales pour faire avancer la recherche? « *C'est un débat qui ne peut pas se cantonner à un débat d'initiés* », conclut-il.

Delphine Jafaar prend alors la parole pour rebondir sur cette idée, et notamment l'actualité en cours sur le procès du Médiator. Elle détaille alors le processus par lequel Irène Frachon a établi son étude de cas. Elle a dû effectuer des croisements avec des données d'autres patients pour obtenir des résultats basés sur un échantillon plus acceptable statistiquement. « *Elle a obtenu ces données non anonymes. Elle les a appelés ensuite, ces patients, et elle a croisé alors au fur et à mesure ses bases de données. Ensuite, le croisement au niveau des données de l'assurance maladie, s'est fait de manière anonyme* détaille-t-elle. *Même si personne ne lui a reproché cette méthode qu'elle a détaillée d'ailleurs dans son livre.* » Mais si Maître Jafaar détaille cette méthode, c'est pour rappeler cette tension. « Les patients n'ont pas estimé que la violation ici, de l'anonymat des données personnelles ne leur a été préjudiciable », conclut-elle. Voilà qui illustre donc la tension entre les deux principes énoncés par Jacques Lucas. Personne n'a remis en cause la violation du principe juridique.

Jacques Lucas reprend alors la parole, pour évoquer les thèses de médecine et affirmer qu'il faut peut-être concevoir le secret médical d'une manière différente et adaptée à cette nouvelle société qui se dessine.
Même si la notion de consentement peut être questionnée, car Jacques Lucas précise que l'on a accepté énormément d'applications, ou d'enregistrements sur un site internet, sans jamais lire les conditions générales.

« Nous abdiquons sur notre liberté, en faisant confiance à une entité. Mais on ne sait pas vraiment à qui nous faisons confiance, conclut-il. *Je fais évidemment référence au Gafam et au BatX chinois, dont les valeurs ne sont pas calquées sur les nôtres. »*

David Gruson est alors interrogé sur le concept de garantie humaine, qui établit qu'un médecin, quand il décide de faire confiance à une intelligence artificielle, le fait en tant qu'humain. Toute son explication est disponible ici.

Delphine Jaffar, ensuite, pour détailler les questions et les fonctionnements du droit au niveau des données de santé, reprend une figure créée par le professeur de droit Jean Rivero en 1962. Il avait créé un personnage, Huron, qui s'interrogeait de manière très naïve sur le rôle du conseil d'État.
Delphine Jafaar, elle, décrit un petit Huron - son fils ? - au pays des algorithmes.

« Notre jeune Huron, « assis au pied de son hêtre sombre, dont une feuille parfois, détachée par
le vent, vient poser sur son épaule l'amorce d'une épitoge noire », a donc quitté les rives fleuries de son fleuve pour venir s'enquérir des règles encadrant le traitement des données massives de santé, plus particulièrement en France »

Le texte entier, très bien argumenté par l'avocate, et qui fait mouche par sa rhétorique est disponible dans son intégralité ici.

L'objectif de la métaphore consiste à poser la question de la dimension collective des données personnelles. Et donc de définir le bénéfice que peut avoir la collectivité à l'agrégation d'un grand nombre de données personnelles. Cette définition limitée par le consentement, évidemment, mais aussi par l'information de l'utilisation de la donnée. Et non pas seulement celle du recueil de la donnée. Savoir comment est utilisée votre donnée est aussi important que savoir qu'on vous la prend.

David Gruson revient quant à lui sur l'explicabilité de la médecine : *« la médecine n'est pas forcément une science expliquable. C'est une pratique, et l'inexplicabilité fait partie intégrante de la recherche médicale, d'ailleurs*, argue-t-il. *Si on cherchait à rendre tout explicable, on bloquerait inévitablement l'innovation. »* Ce n'est pas pour autant qu'il défend une responsabilisation des robots, qui conduirait davantage, selon

lui, à une déresponsabilisation des professionnels qu'à une responsabilisation des machines. Ce qui ne serait évidemment pas à l'avantage des patients.
Il donne ensuite l'exemple du monde des pratiquants du domaine bucco-dentaire, qui a mis en place une organisation de diagnostic, revue ensuite par un groupe d'experts « *indépendants de la start-up initiatrice de la technologie* », détaille-il. Ils ont donc mis en place une organisation qui correspond à la volonté de toutes les instances actuelles, qui est de ne pas perdre le contrôle.
Ce principe repris dans l'article 11 bientôt voté au Sénat sur le projet de loi sur la biotéthique est donc anti-règlementariste. Surprenant pour le droit français, le plus serré sur les données de santé, les données personnelles et leurs utilisations. « *Mais si on ne veut pas subir des effets d'importation de solutions d'IA dont on ne pourra pas garantir le caractère éthique, il faut avoir un droit très ouvert, sur le développement de l'IA elle-même, sous supervision d'un mécanisme de garantie humaine* », conclut-il.

- L'IA sera-t-elle capable de faire des « expériences de pensée » ?

Etienne Klein, CEA

La totalité de sa présentation est disponible ici.

Etienne Klein commence par quelques remarques sur l'esprit critique. Ce dernier ne peut agir que s'il connaît les explications et les remarques de celui ou celle qui s'apprête à critiquer. Il ne saurait donc s'exercer contre des boîtes noires contenant des algorithmes. Mais expliquer pour un physicien n'est pas pareil qu'expliquer pour un biologiste. La polysémie de « l'explication » doit donc se retrouver dans « l'explicabilité ».
De même, il revient sur la traduction historique du terme intelligence de l'anglais en français. « *Si on insistait davantage sur cette différence de sens entre les deux termes anglais et français, beaucoup de débat éthique ne se poseraient pas de la même façon.* »
Idem, il estime que nous projetons sur les machines des mots qui ne leur correspondent pas. « Décisions » : dire qu'une machine décide, c'est déjà une erreur, car elle n'est pas capable

d'expliquer le cheminement logique qui a amené à telle ou telle décision.

Si l'intelligence artificielle est révolutionnaire, les questions qu'elles posent ne sont pas neuves. Il indique par exemple l'extrait d'un texte de Diderot, sur la ré-appropriation du savoir par le peuple, qui selon Etienne Klein, pourrait tout à fait être remis dans la bouche des défenseurs de l'internet libre et de l'Open Source.

Il revient ensuite sur la transparence et les termes de l'opacité de la connaissance scientifique, qui se différencie du savoir scientifique. « *Tout le monde sait que la Terre est ronde, mais qui sait expliquer*

comment nous sommes arrivés à cette connaissance ? » interroge-t-il. En résumé, l'IA est-elle ou sera-t-elle capable d'interroger ce qu'elle sait ?

Cette question ravive les querelles des anciens et des modernes dans les cursus scientifiques.
Dans un développement sur le cursus des écoles d'ingénieurs et l'interdisciplinarité, le chercheur - philosophe introduit sa thèse : la physique ne s'est pas construite sur un ordonnancement des données. *« Ce n'est pas la bureaucratie des apparences »*, déclame-t-il. Elle s'est construite en laissant peu de place aux indications provenant des datas. Elle s'est même construite contre les données empiriques.
Alexandre Koyré explique que le pari de la physique consiste à expliquer le réel par l'impossible. Il développe alors l'exemple de Galilée qui a établi des lois contre le réel. Le principe d'inertie, par exemple, personne n'a jamais vu cela. Les lois physiques ont été trouvées par des gens qui ne se contentaient pas de lire le réel mais qui tentaient de regarder le phénomène, mais qui ont trouvé des stratagèmes intellectuels pour que le monde dé-coïncide d'avec ce qu'il nous montre. Donc avec des expériences de pensées, des fictions qui permettent de tenir le monde à distance.

Il développe ces thèses en vidéo ici.

Galilée a ainsi procédé en considérant qu'Aristote a raison. Mais il va procéder par contrefactuel, en cherchant à voir si Aristote va au bout.

Un autre exemple est cité par Einstein sur la lumière qui lui permettront d'établir que la lumière est un invariant des équations de Maxwell. C'est par des expériences de pensées qu'il a déterminé aussi la théorie de la relativité générale.

Depuis presque 80 ans, il n'y a pas eu de grandes théories physiques. Alors même que l'on n'avait pas de données à l'époque. Et pourtant les gens ont mis sur pieds un formalisme toujours aussi actuel. *« On apprend même à l'école qu'une théorie physique est menacée par tout ce que l'on a comme données »*, s'insurge-t-il. Alors que ces théories et ces formalismes ont résisté au temps.

En 1915 on ne savait rien sur l'univers. Et aujourd'hui quand on fait de la cosmologie ont utilise toujours les équations d'Einstein.

Il cite alors un article de Chris Anderson de 2008, sur la fin de la théorie. *« Le déluge de données va rendre la méthode scientifique obsolète »*, où il développe qu'à partir des nombres pris dans les données, la science n'aura plus à formuler des hypothèses, et donc à s'appuyer sur des théories scientifiques.
Mais les algorithmes extraient dans les données des lois générales, et prolongent le passé. Les algorithmes privilégient l'induction et font confiance à une uniformité de la nature.
Mais corrélation ne veut pas dire causalité. Or nous lisons la corrélation nous lisons une causalité. Et il y a selon le philosophe, une énorme pédagogie à faire sur le sujet notamment auprès des journalistes.

Le chercheur énonce alors plusieurs exemples où la physique n'aurait rien pu faire en se basant sur un stock de données : comme les étoiles. Nous n'aurions rien pu comprendre au fonctionnement des étoiles rien qu'en analysant les données du spectre de la lumière des étoiles. Idem pour le boson de Higgs. Les données nous disent que les particules élémentaires ont des masses nulles. Or c'est justement parce que ces masses nulles contredisent les équations trouvées sans données que l'on a découvert le boson de Higgs. « *Mais les lois contredisent les phénomènes. En tout cas nous obligent à réinterpréter les données* », synthétise-t-il.

Il termine en proposant une expérience de pensée. Imaginons que nous ayons toutes les données de l'univers, tout. Arriverait-on à déterminer les équations d'Einstein ? « Il y aurait dans ces données les ondes gravitationnelles, sachant que nous les avons trouvées parce qu'elles avaient été prédites par les équations », explique-t-il. C'est parce que nous avons eu les équations que nous avons eu les données. « *La question que je pose, c'est est-ce que cela fonctionne dans l'autre sens ?* »

Idem pour l'inertie. « *Un algorithme est-il capable de contredire les données et d'imaginer deux corps tombant à la même vitesse dans le vide ?* » philosophe Etienne Klein.

Il conclut en citant Einstein en 1933 : « *Aucune méthode inductive ne peut conduire au concept fondamentaux de la physique. L'incapacité à le comprendre est la plus grave erreur philosophique de nombreux penseurs du 19e siècle. (il pensait aux positivistes qui contestaient l'existence de l'atome sous prétexte que l'atome n'était pas dans les données expérimentales).*
L'expérience et toutes les données qu'elle nous livre peut nous orienter dans le choix des concepts mathématiques à utiliser mais il n'est pas possible qu'elle soit la source d'où il découle. C'est dans les mathématiques que réside le principe vraiment créateur. En un certain sens, je la tiens pour vrai que la pensée pure et compétente pour comprendre le réel. »
il conclut alors d'une question: « *l'IA forte pourrait-elle venir contredire la pertinence de cet avis* ».