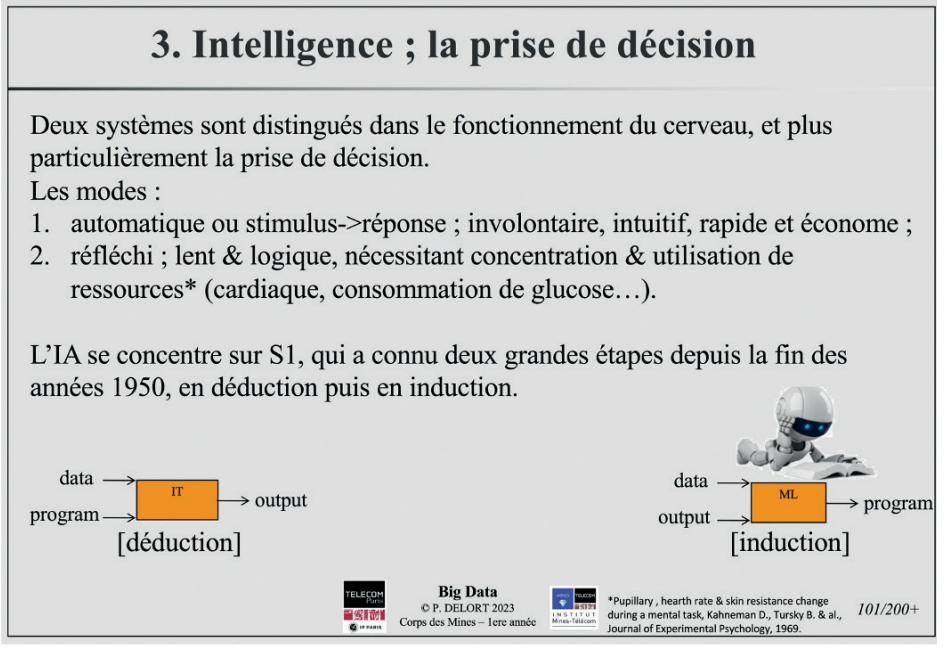
Conférence : seconde vague de l'Intelligence Artificielle, entre performances et explicabilité

1er février 2023

Pierre Delort B 85, DSI, Professeur Invité & PhD à MinesParis, a proposé aux Alumni cette conférence, en format présentiel et distanciel, autour des notions de Big Data, d'Intelligence Artificielle explicable (XIA) et de responsabilité. Pour ce faire, il a mobilisé de nombreux exemples issus de domaines distincts avant d'engager un débat sur les notions de responsabilité et d'exploitabilité, fondamentales sur ce sujet d'XIA.



Platon et Aristote ont posé que les deux formes principales de raisonnement sont la déduction et l'induction.

Dans le raisonnement déductif, une conclusion est atteinte en appliquant des règles générales (prémisses), réduisant le domaine du discours par des étapes logiques, de la règle générale au cas particulier (conclusion) : les hommes sont mortels, Socrate est un homme, donc Socrate est mortel.

Dans le raisonnement inductif, les prémisses fournissent des arguments forts, mais pas des preuves de la conclusion. Nous passons des faits à des règles et les mathématiques permettent de mesurer l'incertitude pesant sur ces règles, dépendant notamment des faits sur lesquels ces règles sont basées. Partir d'un modèle correspond à adopter un raisonnement déductif, partir des données est adopter un raisonnement inductif, avec une incertitude à apprécier. Ainsi, en concluant, par induction, que les animaux (canari, tortue, chien…) sont mortels, la proportion de vieux canaris ou de jeunes tortues dans l'échantillon fera que l'espérance de vie estimée sera très différente.

Ces deux éléments, déduction et induction, sont à l'œuvre dans le cadre de l'IA. Plus précisément, le *knowledge-based system* applique la déduction tandis que le *machine learning* applique l'induction[1].

Les premiers pas de l'IA (déductive) dans les années 1960-1990 ont été marqués par des innovations emblématiques, notamment un système mis au point par Schlumberger d'assistance à l'exploration pétrolière, incorporant les connaissances d'Al Gilreath, un découvreur de pétrole de légende. Des « *knowledge engineers* » du MIT le suivirent durant un an alors qu'il interprétait les résultats de mesure, lui demandant de décrire comment il atteignait ses conclusions, et formalisant ceci.

Le système comportait un moteur d'inférence, 90 règles et une station de travail avec une Interface Homme Machine (IHM) spécifique. Les composants IA, moteur de règle et base de connaissance pesèrent 30% de l'effort et l'interface graphique fut, comme souvent, l'élément le plus important. Le facteur limitant de cette IA dite aujourd'hui symbolique reste le temps des experts en création et en maintenance du corpus de règles, et son avantage est l'explicabilité totale de l'algorithme en ses règles.

Le *machine learning* (IA inductive) fait, quant à lui, l'objet de développements marqués depuis les années 2010, dans un contexte d'accroissement des capacités de traitement, de transmission et de stockage de données. À ce jour, l'IA est très bien représentée par les réseaux de neurones profonds : le *deep learning*, incorporés notamment dans les *Transformers*, dont GPT-3.

Leurs performances sont étonnantes. Deux exemples emblématiques ont été présentés à la conférence, l'un dans le domaine de la vision (avec Yann Lecun[2], reconnaissance à la volée d'objets présentés à la caméra d'un ordinateur portable) et l'autre de l'audio; lors d'une conférence au Collège de France, Stéphane Mallat[3] montre la performance d'un algorithme identifiant et séparant deux discours simultanés, rendant chacun d'eux parfaitement intelligible.

L'importance de l'explicabilité

Pierre Delort aborde la notion d'explicabilité d'un point de vue opérationnel à partir d'un exemple médical. Un travail de comparaison de différentes technologies – régressions, système à base de règles, réseau de neurones – a été réalisé sur l'orientation à l'admission d'un hôpital (entre soins intensifs et soins ambulatoires) de personnes souffrant de pneumonie. Il s'est avéré que les réseaux de neurones affichaient les meilleures performances globales (AUC de 0,86 sur la courbe entre taux de faux positifs et taux de vrais positifs) alors que la régression logistique (AUC de 0,77 donc globalement bien moins performante) a été retenue.

Pourquoi ?

Les réseaux de neurones renvoyaient systématiquement à leur domicile les personnes asthmatiques alors qu'une pneumonie est bien plus dangereuse dans ce cas. Les algorithmes avaient été entraînés à partir de données selon lesquelles les personnes atteintes d'asthme étaient directement envoyées en soins intensifs, ce qui leur donnait un bon niveau de survie. En conséquence, la machine avait appris que l'asthme favorisait la survie des personnes souffrant de pneumonie, ce qui est… complétement faux. La non explicabilité du réseau de neurones a empêché les médecins de réaliser le danger, ici pour les asthmatiques, de l'algorithme.

(1) Qui peuvent être mêlés, cf. Model Output Statistics par exemple de la Météorologie.
(2) Universitaire et dirigeant du laboratoire d'IA de Meta.
(3) Mathématicien.

Maintenant, la meilleure performance globale du réseau de neurones relativement aux régressions logistiques pourrait être un argument pour son adoption… si, envers les asthmatiques, la responsabilité des médecins n'était pas en jeu ; ils *auraient* dû savoir la sur-léthalité affligeant les premiers en cas de pneumonie et donc *auraient* dû les envoyer directement en soins intensifs, comme à l'accoutumée.

La responsabilité des personnes, physiques ou morales, apparaît ainsi importante pour l'utilisation des algorithmes, ce qui suppose donc leur explicabilité.

Plusieurs modes d'explicabilité font l'objet de recherches dont nous citons deux catégories, les valeurs de Shapley et les explications génératives d'image.

Les valeurs de Shapley, fondées sur la théorie économique du même nom, proposent une explicabilité tout à la fois globale et locale (par occurrence). Leur utilisation peut produire des figures séduisantes. Cependant le volume de calcul nécessaire rend l'emploi de raccourcis indispensable, conduisant parfois à des aberrations « they can assign non-zero attributions to features that are not even referenced by the model [4] ». Elles sont ainsi à utiliser avec précaution..

Procédant non de manière introspective mais justificative, les explications génératives [5] expliquent en quoi le résultat d'un système visuel est compatible avec des éléments de cette image et elles génèrent des phrases détaillant en quoi des éléments visuels de l'image sont compatibles avec le classement attribué par l'algorithme.

D'autre solutions sont explorées. Yann Lecun établit un excellent parallèle avec les médicaments dont le fonctionnement est parfois totalement obscur. Leur utilisation est validée par une Autorisation de Mise sur le Marché (AMM), incluant des démarches statistiquement valides de test et d'identification d'effets secondaires. Yann Lecun avance qu'il pourrait être pertinent de réaliser le même genre d'étude pour les algorithmes que personne ne comprend. Cela est judicieux, notons seulement qu'une AMM pharmaceutique prend jusqu'à une dizaine d'années. Est-ce réaliste pour des algorithmes basés sur des données d'affaire et souvent de validité brève ?

Maintenant l'explicabilité porte une contrepartie. Un programme de recherche de la Defense Advanced Research Projects Agency (DARPA, une agence de recherche du ministère de la défense US) met bien en évidence la tension entre explicabilité et performances des algorithmes. Par exemple, les arbres de décision affichent un très bon degré d'explicabilité, mais un assez faible niveau de performance, alors que c'est l'inverse pour les réseaux de neurones profonds. Les champs de recherche de ce programme incluent :

- comment produire des modèles plus explicables ?
- comment concevoir des interfaces d'explication ?
- comment comprendre les prérequis psychologiques pour des explications effectives ?

En conclusion, Pierre Delort propose d'ouvrir la réflexion sur le thème de la responsabilité devant l'inexplicable. Plus précisément, quelle responsabilité est-il possible de demander à un acteur qui prend des décisions sur la base d'algorithmes que personne ne comprend (GPT-3 compte 175 Milliards de paramètres, ChatGPT qui en est une spécialisation sur l'interaction humaine, 20 Md de paramètres) ?

La responsabilité doit-elle incomber à l'individu décisionnaire ou à l'organisation (en tant que personne morale) à laquelle il appartient, tel a été le sujet d'une mission de consulting de Pierre Delort avec une banque ? Le sujet étant de déterminer, en fonction de l'activité et de la culture des équipes, comment est répartie, et acceptée, la responsabilité entre les entités fournisseur de données d'entraînement, créateur de l'algorithme et utilisateur de cet algorithme, dépendant de l'observabilité de l'algorithme (souvent faible cf. ci-avant) et des mécanismes de sa maintenance, d'identification et correction des dérives…

Congruente avec le sujet de responsabilité, et conditionnant l'emploi et donc l'utilité de l'algorithme, une explicabilité doit souvent être complétée d'interprétabilité et d'exploitabilité, comme une récente mission de l'auteur en management de transition (Direction Scientifique d'une jeune société) l'a montré.

Le sujet cité consistait à présenter, parmi les 200 000 clients, essentiellement TPE, en Multi Risque Professionnelle d'un assureur de premier plan, les 2 ou 3% les plus appétents en up ou cross selling (vendre plus de garantie du même produit d'assurance ou vendre d'autres produits) aux équipes commerciales.

Il s'est trouvé que fournir, avec la liste des clients à contacter, les variables indépendantes de plus forte importance pour l'appétence en up/cross selling ne suffisait pas, même avec des résultats statistiques corrects (AUC de 0,73 optimisée sur le F1 Score) et satisfaisant les actuaires du Marketing de l'assureur, clients de la société.

L'acceptabilité (dont par la chaîne hiérarchique du réseau commercial) et les premiers résultats ont en fait dépendu, au-delà de la fourniture d'explications (un exemple simple, la jeunesse du dirigeant), d'éléments d'interprétabilité (par ex. un taux endettement personnel élevé) ainsi que d'exploitabilité (la protection offerte en perte d'exploitation) dans le discours commercial.

Ainsi lorsqu'un algorithme s'insère dans un système social, son explicabilité est un facteur déterminant de son succès, et la tension existante entre performances et explicabilité laisse plusieurs champs ouverts, dont mixer IA inductive et déductive, insister sur l'IHM et les prérequis psychologiques, ceci reprenant des enseignements de l'IA symbolique…

Les progrès très récents des technologies d'IA (facilité de choix des hyper paramètre, méthodes de Tuning, Auto Machine learning, économie des étiquettes de l'apprentissage par renforcement, algorithme proposés sur le Cloud en utilisation…) facilitent la génération d'algorithmes performants et donc souvent non-explicables. Dès lors le défi devient la bonne insertion de ces algorithmes dans les mécanismes de décision, de responsabilité et d'action des entreprises.

Pierre Delort

est B85, Professeur Invité & PhD à MinesParis, est DSI et président de l'Association Nationale des Directeurs des Systèmes d'Information. Il débuta sa carrière dans le consulting, puis pris plusieurs positions de responsabilité en organisation d'entreprise, puis Direction des Systèmes d'information, direction des Données et Direction Scientifique. Il est l'auteur de « Le Big Data », coll. Que Sais-je ? et intervient en entreprise en consulting et management de transition.

(4) Sundararajan M. & Najmi A. The Many Shapley Values for Model Explanation, Google, 2020.
(5) Darrell T & al, Generating Visual Explanations, Berkeley & Max plank Institute, 2016.