Une perspective historique sur l'IA explicable Document préparatoire à un tutorial AFIA juillet 2020

Alain Mille, Rémy Chaput, Amélie Cordier

An historical perspective on XAI / Une perspective historique sur l'IA explicable

Document préparatoire à un tutorial AFIA juillet 2020

Alain Mille, Rémy Chaput, Amélie Cordier

Table des matières

Pourquoi (encore) un tutoriel sur l'explicabilité

Le mouvement international de recherche sur l'« Intelligence Artificielle Explicable » s'est considérablement développé depuis quelques années et la plupart des conférences du domaine proposent maintenant des sessions sur le sujet tandis que des collectifs se forment autour du sigle XAI[1].

L'urgence de ces recherches s'est révélée avec l'extraordinaire impact dans la société des « intelligences artificielles » et plus encore lorsque ces « IA » sont le résultat d'un apprentissage à partir de données massives, et plus encore, si c'est possible, lorsque l'apprentissage exploite les techniques dites de « l'apprentissage profond ».

Le cahier des charges de la littérature

C'est peut-être avec l'annonce de *l'agence des projets de recherche avancé de la défense des USA (DARPA)*, publiée pendant l'été 2016 que l'officialisation d'un domaine de recherche spécifique a été faite [1].

Ce rapport explique les enjeux de cette recherche de la manière suivante :

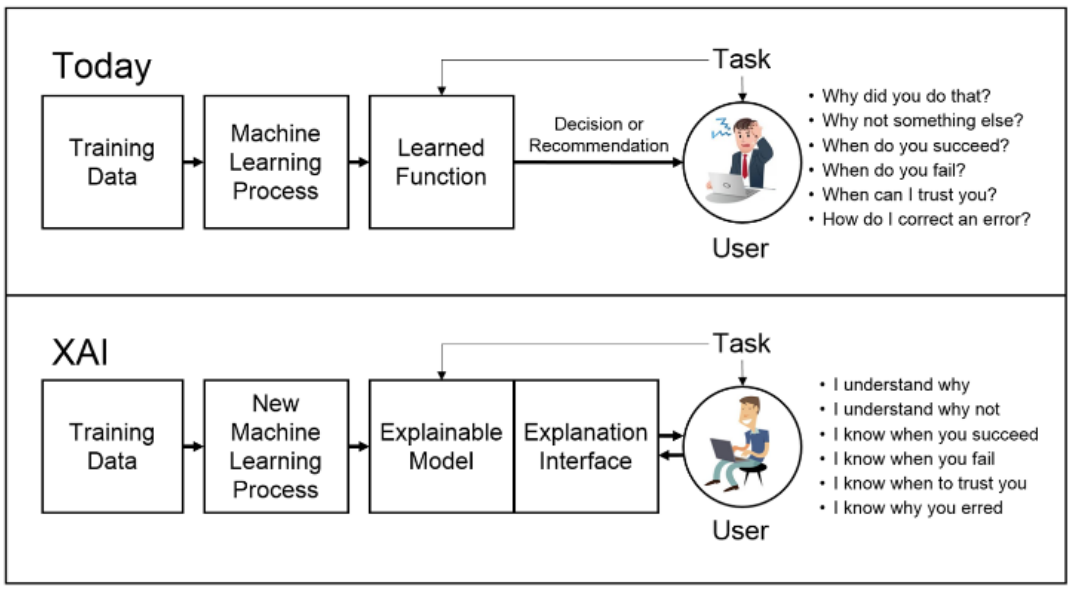


Figure 1: XAI Concept

1 *https://en.wikipedia.org/wiki/Explainable_artificial_intelligence*

Un schéma complémentaire précise des sous-domaines de recherche :

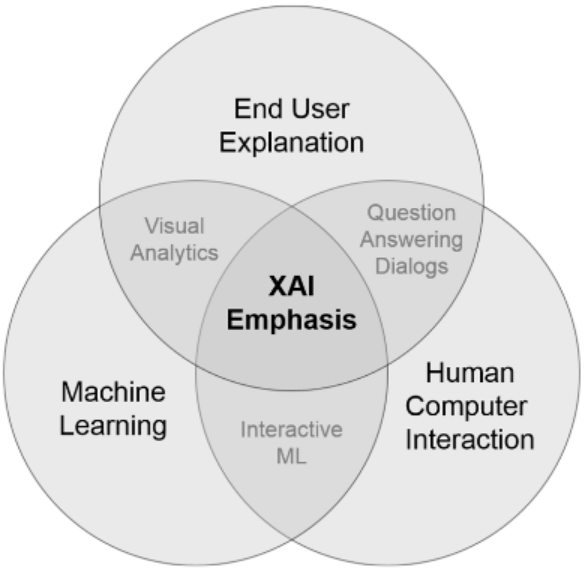


Figure 2: XAI Emphasis

Enfin, des cas d'usage sont utilisés pour illustrer comment ça pourrait être utilisé en pratique :

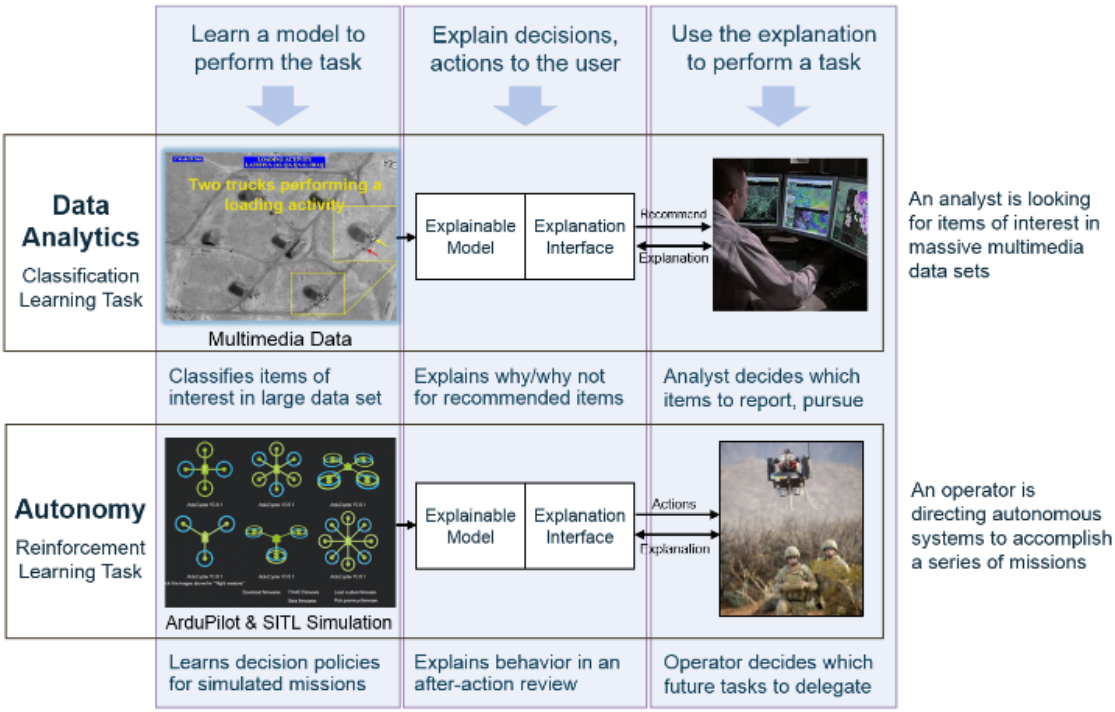


Figure 3: XAI Challenge Problem Areas

La communauté X-AI de l'apprentissage automatique a complété le cahier des charges en listant les ***audiences*** concernées par les explications produites et influençant les différents modèles d'explication à savoir construire. Un schéma illustre cette notion d'audiences :

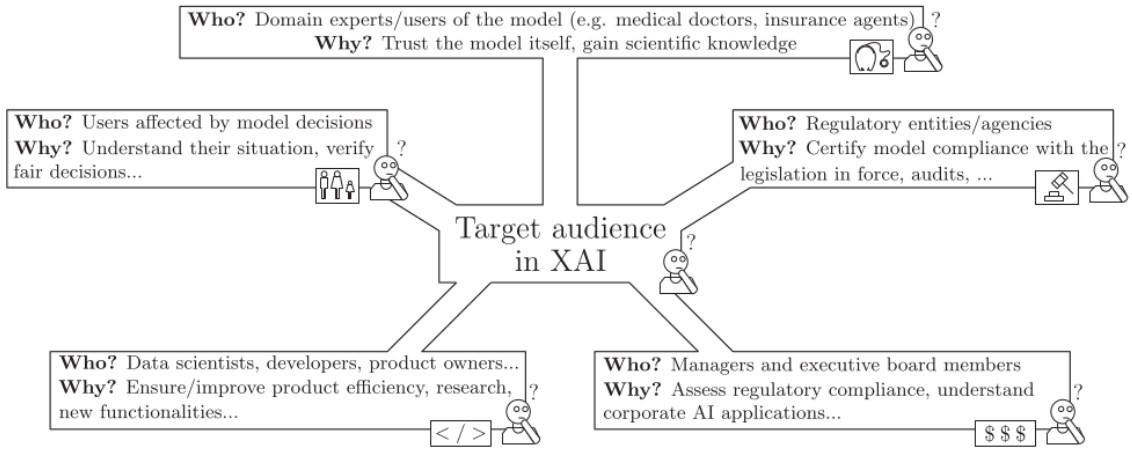


Fig. 2. Diagram showing the different purposes of explainability in ML models sought by different audience profiles. Two goals occur to prevail across them: need for model understanding, and regulatory compliance. Image partly inspired by the one presented in [29], used with permission from IBM.

Illustration 1: Types d'audience

Objectif spécifique du tutoriel

L'objectif de ce tutoriel est d'une part de faire le point sur l'avancement de ce programme de recherche, sur la diversification et la ramification de ces recherches, et bien sûr sur les enjeux actuels après ces quelques années d'effervescence sur le sujet.

Comme nous le voyons dans l'annonce de l'agence américaine, la question porte exclusivement sur l'explicabilité des modèles « appris » à partir de données collectées.

Comme semble le pointer également le terme *audience*, la communauté X-AI considère que les *utilisateurs concernés* sont avant tout *passifs* et qu'il convient de leur présenter des explications qui soient taillées convenablement pour chaque *type* d'audience.

Nous verrons dans ce tutoriel, qu'une *typologie* serait difficile et fatalement normalisante, à son tour et qu'il faut intégrer la question de l'individuation du processus d'explication, que ce soit du coté du dispositif technique numérique que du coté de l'humain concerné par le processus explicatif.

L'explicabilité reste une question centrale et souvent *bloquante* à l'heure des débats sur l'usage de *modèles* formalisés par des informaticiens ou appris à partir de données pour intervenir dans la décision des institutions, des entreprises, des armées, des scientifiques et de manière directe sur les décisions de chacune et chacun.

Cette exigence d'explication par la société s'explique par :

- des raisons éthiques, et il existe toute une littérature sur cette question et très récemment, le parlement européen a publié un rapport spécifique à ce sujet [2],

- des raisons liées à l'établissement de responsabilités, avec en particulier de très nombreuses initiatives dans le domaine de la justice, et en France la revue Dalloz a publié un recensement des publications francophones sur le sujet [3]. La question se pose depuis de nombreuses années dans d'autres pays, en particulier avec l'utilisation de COMPAS[2] aux USA,
- des raisons liées à la sémantique des modèles produits par l'analyse de données pour découvrir des lois du monde : une présentation très récente et particulièrement claire est disponible sur le sujet sur le web par Mihaela van der Schaar[4],
- des raisons économiques : la maîtrise des données massives et le déploiement tout aussi massif des applications de l'intelligence artificielle sont pointées comme les vecteurs les plus prometteurs, avec la transition écologique parfois, du développement de l'économie. Mais, ce marché extraordinaire est lié à la capacité à être d'une part maîtrisé par les entreprises et d'autre part adopté massivement par les utilisateurs. Il faut donc qu'ils soient « dignes de confiance » pour les uns et les autres. Ce sujet est l'objet d'un rapport récent pour le MIT [5].

Nous ajoutons, que des dispositifs techniques numériques capables d'intégrer des processus d'explication *empathiques* avec les besoins des utilisateurs participent activement à la formation et à l'éducation par l'usage.

En effet, est-il utile de rappeler qu'avec l'usage d'algorithmes d'apprentissage pour créer des modèles de décision ou de recommandation, la question de l'Intelligence Artificielle se pose de manière renouvelée et qu'elle devient un enjeu social majeur, quelle que soit la forme qu'elle prend lorsqu'il s'agit de remplacer un *jugement* humain par un *calcul d'optimisation.*

La ***délibération*** de ces jugements, qu'ils interviennent dans le cadre des institutions judiciaires ou non, nécessite la capacité de *s'expliquer.* L'explication n'est donc pas uniquement la capacité à fournir des informations sur la manière dont le jugement a été *calculé* et à partir de quelles *données,* mais aussi le processus même de délibération explicative. Dans quelle mesure l'utilisateur, quel que soit son type et en particulier l'utilisateur concerné directement par un jugement, peut-il mettre en *délibération* le *jugement calculé,* c'est-à-dire la norme retenue pour constituer les critères d'optimisation ?

Nous allons tenter d'instruire cette approche en examinant d'abord les fondements philosophiques de l'explication, en passant en revue l'état de l'art actuel, en le confrontant aux objectifs poursuivis et en les éclairant sous l'angle de la capacité à soutenir des processus de délibération.

2*https://en.wikipedia.org/wiki/COMPAS_(software)*

Introduction générale : Intelligence artificielle et explicabilité

Intelligence artificielle : de quoi parlons-nous ?

L'histoire de la notion même d'intelligence artificielle se confond avec celle de la notion du mécanique d'abord (robots mécaniques, mythologie de l'humain artificiel), puis du computationnel avec Türing. L'intelligence artificielle a pris le statut d'objet de recherche à l'occasion de la fameuse conférence de Dartmouth College, en 1956. Récemment, depuis 5 à 10 ans tout au plus, le terme Intelligence Artificielle échappe à ce mouvement initial lié à la recherche ou à l'entreprise pour apparaître massivement dans les média et être maintenant désiré ou craint de manière massive. Nous, chercheurs en IA, n'avons pas toujours compris pourquoi, à la radio, dans les médias, on parlait d'Intelligences Artificielles, au pluriel, quand souvent nous n'y voyions que des algorithmes. La confusion algorithme-IA semble aujourd'hui admise largement.

Passé le moment d'incrédulité, nous comprenons pourquoi il est en fait correct de considérer de nombreux algorithmes comme des intelligences artificielles. En effet, nous réalisons que les dispositifs techniques numériques actuels s'intéressent à réguler des fonctions cognitives qui, jusqu'à il y a peu, étaient réputées hors du champ de l'automatisation. Les dispositifs techniques numériques actuels s'emparent de la cognition, se déclinant en recommandation comportementale, orientation de la décision, décidant de ce qui est bien ou non pour l'utilisateur, jugeant des capacités cognitives lors des apprentissages, classant inlassablement les profils cognitifs au profit de stratégies non explicitées, mais s'engageant, très souvent, à tout faire pour le bien-être des utilisateurs. Nous trouvons dans cet état de fait, la justification de la nécessité de proposer des voies pour l'explicabilité de ces intelligences artificielles, que tout un chacun peut installer sur son terminal numérique, et en premier lieu actuellement son smartphone. A ces usages massifs, s'ajoute la perspective des robots compagnons, de la voiture autonome, de l'humain augmenté... La question de l'explicabilité se pose alors de manière plus directe et sensible, et ces développements relèvent encore plus essentiellement de la cognition, et l'avenir du marché associé est souvent présenté comme considérable. La société du bien-être est la justification la plus courante de ces développements dans l'offre du numérique. La diffusion massive d'objets interconnectés en un internet des objets (Internet of things), facilite cette mise en œuvre en ajoutant au raisonnement-calcul, des capacités sensori-motrices par extension et même par procuration de l'utilisateur. C'est dans ce contexte mouvant et où le marketing l'emporte souvent sur la rigueur que nous situons ce tutorial sur l'explicabilité de l'IA.

Explicabilité, Explication, Processus d'explication

L'explicabilité d'un phénomène est liée à la capacité à l'expliquer. Il n'y a pas d'études philosophiques particulière sur cette notion, mais la question de l'explication habite les travaux des philosophes depuis longtemps sous le terme général de théorie de l'explication.

La théorie de l'explication s'est d'abord focalisée sur la manière d'établir des **causes** pour expliquer l'observation de tel ou tel phénomène. La question principale est POURQUOI. Cette question reste vive, en particulier dans le monde numérique quand il s'agit d'établir des « lois » scientifiques à partir de l'observation massive de données collectées.

Toutefois, au tournant du 20ème siècle, le concept d'explication se diversifie en s'intéressant à des facettes différentes du processus explicatif.

L'approche causale rigoureuse, nécessite de pouvoir construire une chaîne causale complète qui à partir d'observations non contestables infère des conséquences avec un mécanisme déductif validé. Si, *quand Ca est observé, il y toujours Co observé, mais pas l'inverse, alors Ca peut être considéré comme la Cause de la Conséquence Co.* Pour l'établissement de cette chaîne causale, il faut que les phénomènes soient observables et les observations irréfutables. Au début du 20ème siècle, les progrès scientifiques majeurs, en physique notamment ont été réalisés sans qu'il soit possible d'observer directement la causalité. L'observation modifiant le phénomène ou n'étant possible qu'indirectement par des effets indirects. Les instruments de mesure observent les effets qui confirment ou infirment les lois établies pour expliquer des effets (observés) et prévoir d'autres effets (attendus) si la loi est valide. Le chercheur propose une loi qui permet d'expliquer des effets, ce qui décrit une « réalité » non observable selon les méthodes habituelles.

Cette nouveauté a profondément modifié l'étude de la théorie de l'explication qui se décline alors selon une orientation réaliste (empiriste) ou épistémique.

Pour une approche réaliste (empiriste), une explication est une description littérale de la réalité externe -> une forme de récit contextualisé de ce qu'il se passe -> tout ce qui est décrit est concret et observable.

Pour une approche épistémique , l'explication sert à faciliter la construction d'un modèle empirique cohérent, c'est à dire un modèle qui ne rentre pas en contradiction avec l'observation, sans en faire une description littérale contextualisée. Il s'agit alors d'une forme "logique" au sens d'un modèle qui explique le mieux l'effet, observable à partir de causes qui ne sont pas directement observables en situation réelle.

Au delà de cette dualité, certains philosophes ont alors complété la théorie de l'explication en y intégrant l'étude du processus d'explication. Par exemple, la philosophie du langage se focalise sur la compréhension entre individus. Ou encore l'approche basée sur les sciences cognitives considère que l'explication est purement cognitive et résulte d'une représentation mentale liée à l'activité et aidant cette activité.

Nous listons ci-dessous les principales approches que nous reteons. Ces éléments sont tirés de l'Internet Encyclopedia of Philosophy [6]

Approche historique et scientifique : la causalité :

"The event under discussion is explained by subsuming it under general laws, i.e., by showing that it occurred in accordance with those laws, by virtue of the realization of certain specified antecedent conditions" (p.152). Définition tirée de l'article séminal de Hempel et Oppenheim [7]

Développements contemporains / théorie de l'explication

1. **Explication et "réalisme causal"** Il s'agit d'un raffinement de la théorie historique par Salmon[3] qui introduit une approche probabiliste pour établir une causalité. Il s'agit toujours de production de règles causales permettant de répondre à la question POURQUOI ? C'est une approche privilégiée au niveau scientifique quand le processus causal observé est bien le « vrai » et pas un pseudo processus où l'on garde dans l'observation des informations qui, bien que corrélées car liées à l'activité observée, ne sont pas causales.
2. **Explication et empirisme constructif** (Bas van Frassen[4]) Bas van Frassen conteste le fait qu'une théorie scientifique et ses modèles associés soient des « explications » complètes. Il indique que tout au plus, l'explication donnée par le modèle est liée aux données qui ont permis de l'établir. L'explication est valide dans le cadre de ces données. Il s'agit d'une inférence conditionnelle, et les probabilités utilisées pour établir les lois sont bayésiennes par nature.
3. **Explication et philosophie du langage ordinaire** (Peter Achinstein[8]) L'explication est une tentative de fournir des éléments à une autre personne qui pose la question d'une certaine manière. La question n'est pas seulement POURQUOI, mais peut être QUI, QUOI, COMMENT, OÙ...Tout acte pour répondre à une telle question est un acte d'explication. La reformulation dans le registre de la personne qui pose la question est alors nécessaire. *S formule U avec l'intention que cette formulation rende Q compréhensible en produisant la connaissance de la proposition exprimée par U comme réponse correcte à Q.* Cette approche est pragmatique tout en cherchant à rester « correcte » (avec la validité de l'état de l'art). La critique est souvent que l'explication soit satisfaisante tout en étant trop vague.
4. **Explication et sciences cognitives** [9], [10] -> l'approche de Holland concilie les approches symboliques et neuro-scientifiques en considérant qu'une explication est recherchée lorsque les "schemes" habituels sont incapables de gérer une situation "inconnue". C'est la "surprise" qui fait que l'on se pose une question : pourquoi, comment, qui, où, quoi...). Une investigation part de la mise en évidence d'un "erreur" d'interprétation de ce qui est automatiquement venu à l'esprit. Plus récemment, en 2019, dans Zachery et al le processus d'explication a été étudié en démontrant la validité de cette intuition en montrant à quel point l'explication était développementale, située, orientée par la personne qui cherche l'explication.

Ces approches théoriques sont importantes à avoir en tête quand il s'agit d'explorer la bibliographie récente sur l'explication à l'ère du Web et de l'Intelligence Artificielle. Nous reprenons ici une série de papiers publiés sous le titre générique « Explaining Explanation » et coordonnée par Robert R-Hoffman.

Part I : Theorical Foundations [11] insiste sur l'importance centrale de l'explication causale, tout en mettant en évidence les mécanismes qui animent le processus de l'explication.

3 *In Stanford Encyclopedia of Philosophy : https://plato.stanford.edu/entries/wesley-salmon/*

4 *Wikipedia : https://fr.wikipedia.org/wiki/Bas_van_Fraassen*

- *Causal reasoning plays a central role in our mental models about how events transpire and what will happen if we intervene. Our mental models hinge upon knowledge and beliefs we summon to make sense of events. We might even define our mental models as the causal network that we understand to be operating to make things happen.*
- *Causal reasoning is central to decision making; the causal models people hold determine the way they recognize and categorize situations and the kinds of mental simulation they will perform to evaluate courses of action.*
- *Causal reasoning is central to replanning, diagnosing why a plan might be going poorly and considering what needs to be altered.*
- *Causal reasoning is central to coordination, anticipating how individuals' actions will affect the team's sactivities.*
- *Causal reasoning is central to anticipatory thinking, using our mental models to prepare ourselves for possible events, particularly low-probability high-impact events.*

  Le processus d'explication serait lié aux mécanismes:

  - d'abduction -> l'inférence hypothétique telle que définie par Peirce[12]. Ce serait le plus souvent le « début », l'hypothèse à vérifier, bien souvent par induction ;
  - de retrospection -> voir si des explications du passé peuvent fonctionner vérification contrefactuelle ;
  - de prospection -> s'interroger sur ce qui pourrait arriver, en relation avec la prédiction.

Part II : Empirical Foundations[13]

l s'agit d'un article général sur la notion d'explication dans un raisonnement causal (par exemple en science, pour l'explication d'un phénomène, mais aussi pour expliquer une décision).

Les causes ont été typées par David Hulme[5] selon quelques caractéristiques :

- la propension, c'est à dire une cause "naturelle" et connue pouvant expliquer les choses
- la réversibilité, c'est à dire le fait que si la cause disparaît à quel point l'effet disparaitrait-il ?
- la covariation, c'est-à-dire la coïncidence des causes et des effets (corrélation).

On pourrait ajouter la manipulabilité, c'est-à-dire le fait que si on modifie la cause, l'effet se retrouve modifié d'une certaine manière

L'explication est donc considéré comme le fait de fournir les relations causales entre un effet et d'autres éléments qui en seraient les causes. Le contexte peut modifier considérablement les choses et il doit donc être précisé (les causes universelles causant des effets universels sont rares).

Les auteurs ont interviewé 10 spécialistes en logistique, renseignement, commande et contrôle. Ils constatent que même *quand les choses ne se passent pas comme prévu mais que l'objectif*

5 *(D. Hume, A Treatise of Human Nature, anonymous publisher, 1739–1740;reprinted by New Vision Publications, 2007.)*

poursuivi est atteint, il n'y a en général pas de recherche d'explications causales sur la perturbation observée.

Une question principale est donc de comprendre la raison principale qui fait que l'on va chercher des explications.

Les auteurs établissent une table[Illustration 2] des types de buts « naturels » du raisonnement causal :

Table 1. Some natural purposes of causal reasoning.

Type	Ipsative ("self") causal reasoning	Projective ("other") causal reasoning
Prospective	Reasoning about the future (forecasting)	Reasoning about what someone else thinks will happen
Interventive	Natural experiment or anecdote	Deliberate experimental action to probe the cause-effect relation or test some theory
Inspective	Comprehending the present (nowcasting)	Reasoning about what someone else thinks is happening
Retrospective	Reasoning about past events (hindcasting)	Reasoning about what someone else thinks has happened
Reflexive	Reasoning about one's own reasoning, for example, "Why is this difficult?"	Reasoning to influence someone else's reasoning, for example, deception
Continuous	When do I have an account? What is the stopping rule?	Reasoning to prevent someone else from engaging in causal reasoning
Corrective	Recognition that there is an explanatory gap Reasoning about what went wrong in one's causal reasoning When do I change my explanatory account? How do I know when to change it? Responsive gap filling (response to encountering a black swan)	Recognition that there is an explanatory gap Reasoning about what went wrong in someone else's causal reasoning Responsive gap filling (response to encountering a black swan)
Protective	Reasoning to achieve a justification or rationalization of one's actions, to provide a rationale (for example, "cover your butt")	Reasoning to achieve a justification of rationalization of someone else's (or some organization's) actions, to provide a rationale (for example, "cover your butt" and "scapegoating")

Illustration 2: Tableau des types de buts naturels pour le raisonnement causal

Deux options : soit on se pose des questions sur la base de **l'observation de sa propre expérience,** soit on se pose de question sur la base de **l'observation d'un autre**. Les auteurs relèvent dans leur enquête que de manière de plus en plus importante les personnes concernées tentent de comprendre et raisonner sur les buts, les actions et les processus de systèmes complexes intelligents, et en premier lieu ceux qui réalisent des tâches associées à l'IA.

Un tableau [Illustration 3] est dressé pour lister les "clichés" relevés sur les processus d'explication causale :

Dans la même section, une étude sur les modes d'explications préférés montrent une importance significative de la culture des personnes interrogées.

Table 3. Myths about causal explanation.

Myth	Reality
Correlation does not imply causation.	Of course it does. It does not require a determination of causation, but it is often the beginning of a fruitful investigation.
Logic is the exclusive basis for the analysis of causal reasoning.	Perhaps in philosophical investigations, but in real-world settings the evidence for causation is usually too ambiguous to permit valid deductive inferences.
The analysis of physical causation is the model for understanding all forms of causation.	Of all the events about which humans reason, speculation about causation more often involves indeterminate questions for which there will never be any closure on the single or "real" cause.
The scientist is the ideal model for causal reasoning.	Much of the research on causal reasoning involves scientists learning to overcome reductive tendencies to oversimplify cause-effect relationships. However, in natural settings some degree of simplification is necessary to cope with complexity, and furthermore, scientific standards are too restrictive.
Causal reasoning means finding the one true cause for an effect.	This might be true for studies of physical causation, but it is not true for natural settings involving indeterminate causes. The quest for a single "root" cause must be a distortion and an oversimplification, although people often seek single causes as a desirable simplification.
The "effect" to be explained is usually clear.	In natural settings, people often revise the description of the effect as the causal investigation continues.
Causal reasoning is to be described as a process having clear-cut beginnings and endings.	Quite often, it does not.
The property of "being an explanation" is a property of statements.	Clearly, it is a complex interaction.

Illustration 3: Tableau des clichés sur le raisonnement causal

Part III : The causal landscape[14]

L'auteur propose la notion de réseau causal pour **représenter** les causes conditionnelles d'un événement dans une situation complexe.

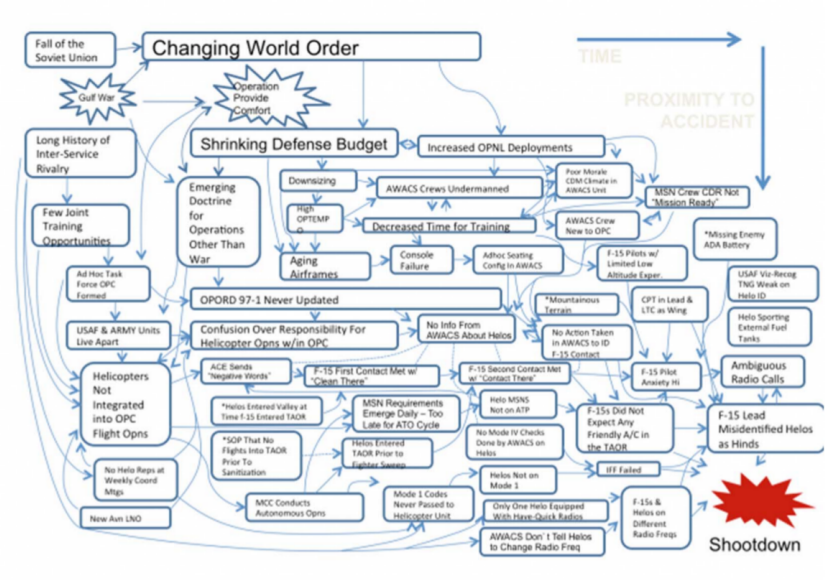


Figure 1. Snook's causal network for the Black Hawk shootdown (reproduced with permission from *Friendly Fire: The Accidental Shootdown of US Black Hawks over Northern Iraq*; Princeton University Press, 2002).

Illustration4: Réseau causal d'un tir

Un réseau causal[6] est un réseau bayésien dont les relations représentent des causalités.

6 *https://en.wikipedia.org/wiki/Bayesian_network#Causal_networks*

Dans cette partie, l'auteur illustre l'utilisation d'un tel réseau pour expliquer un « tir ami » réalisé par deux avions militaires américains sur deux hélicoptères américains dans des conditions de visibilité et d'accès à l'information normales. Le tir « ami » a causé 21 morts.

Le diagramme explicatif d'un tir est complexe et très difficile à utiliser pour comprendre la situation elle-même [Illustrations 4 et 5]

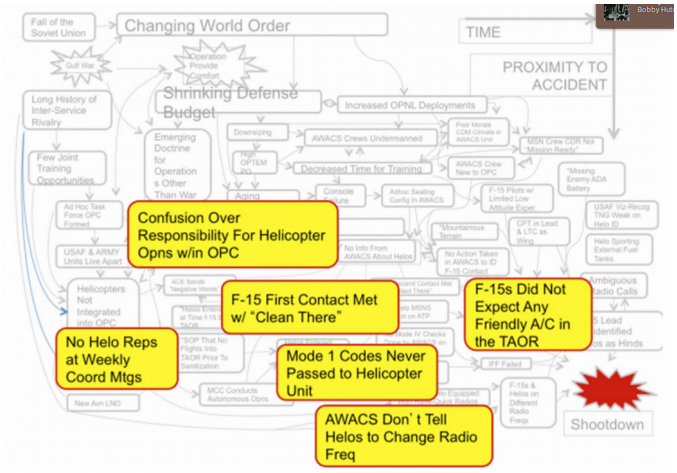


Figure 2. Snook's causal network highlighting the key nodes based on the scalar analysis. (Reprinted with permission.)

Illustration 5: Les nœuds clés identifiés dans la situation précise du tir

Part IV : A deep dive on deep learning [15]

L'article rappelle l'approche causale pour expliquer un modèle [Illustration 6] (selon J.A. Overton[7]) en l'appliquant au cas du Deep Learning et en distinguant différentes informations d'explications globales au processus d'apprentissage :

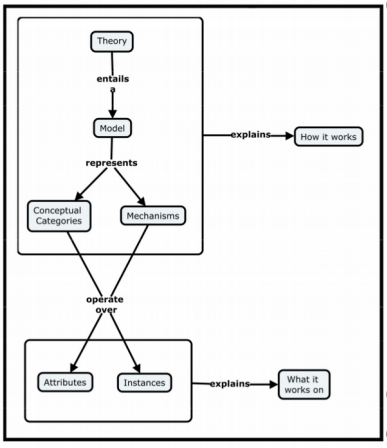


Figure 1. A process for global explanation.

Illustration 6: Processus d'explication globale

- L'information sur les données utilisées pour l'apprentissage.
- L'architecture et les algorithmes.
- La description des performances au niveau fonctionnel.
- Les modes de défaillance et les conditions limites.

en distinguant le processus d'explication globale, générique de l'explication locale, spécifique et exploratoire, lié aux capacités d'observation [Illustration 7].

7 J.A. Overton, « ''Explain'' in Scientific Discourse », Synthese, vol. 190, no. 8, 2013, p.1383-1405.

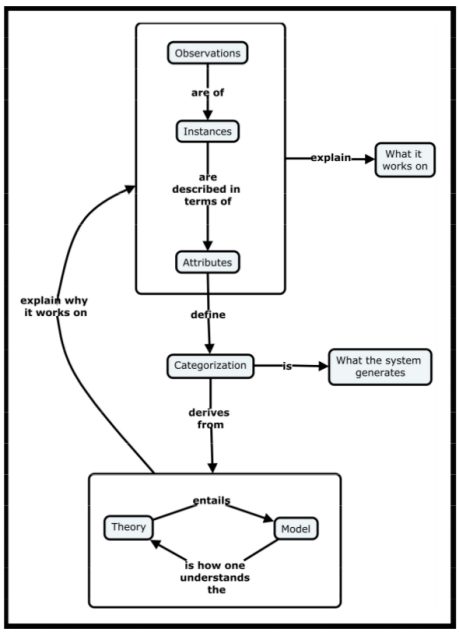


Figure 2. The process for local explanation.

Illustration 7: Processus d'explication locale

Si les explications globales sont relativement stables et documentées, les explications locales son bien plus nombreuses, situées et difficiles à caractériser : À quoi sert… ?, Pourquoi... ? Qu'est-ce qui… ? Comment… ? Que se passe-t-il si… ? Pourquoi pas… ? Qu'est qui arriverait si… ? On dit que c'est X, mais pourquoi on ne dit pas que c'est Y.. ?

Les méthodes habituelles sont contrefactuelles et contrastives : il s'agit de pouvoir tester différentes configurations pour tenter de comprendre la « logique » du modèle sous-jacent.

L'explication est alors une exploration et plusieurs méthodes inspirées des méthodes de test d'hypothèse[8] ou de test de discrimination perceptuelle[9] sont listées (avec des exemples exclusivement dans le domaine des images) :

- Recherche des co-variations : on fait changer une variable dans les données pour observer les effets de modification de cette variable (choix de filtres par exemple)
- Recherche de cohérence des résultats : on utilise la connaissance ontologique de ce qui est classé pour enlever des éléments de repère connus pour reconnaître tel ou tel élément (on enlève le bec par exemple pour un oiseau).
- Recherche par différences : on part d'un élément bien classé et on lui joint un élément qui n'a rien à voir avec son classement (un oiseau et une pizza par exemple). Quelle fusion peut déclencher une mauvaise classification ? Si la classification est la même, qu'est qui a permis de ne pas la modifier…
- Association des deux précédentes recherches : on fabrique des jeux d'éléments considérés comme « proches » par l'algorithme et des jeux d'éléments considérés comme très éloignés. Les propriétés différentes ou communes sont alors repérées.
- Recherche par ajustement : on modifie graduellement un élément pour détecter ce qui provoque une erreur de classification.

Nous verrons dans la suite du tutoriel que cette approche causale, bien qu'insistant sur l'importance du processus ne permet pas de prendre en compte le processus d'explication avec un utilisateur « final » d'un dispositif de type Intelligence Artificielle.

8 *J.S. Mill, A System Of Logic, Ratiocinative And Inductive, 2002 edition, University of the Pacific Press, 1843*

9 *G.T. Fechner, Elemente der Psychophysik, Breitkopf und Härtel., 1860*

Cet article s'appuie sur l'étude réalisée en 2014 sur ce qui peut tromper facilement un algorithme de Deep Learning [16]

Explaining Explanation for « Explainable AI » [17]

Les précédents articles ont été repris pour réaliser un article spécifique à la question de l'intelligence artificielle. Les mêmes auteurs ont plus particulièrement traité des capacités explicatives de l'IA symbolique dans une synthèse sous forme de « méta-revue »[18].

Les auteurs y pointent les éléments qui font « une bonne explication » dans le cadre de dispositifs de type Intelligence Artificielle. Ces éléments constituent une sorte de contextualisation théorique à la série d'articles présentée plus haut. Nous reprendrons plus lon à notre compte ces éléments pour définir les capacités à exhiber dans un dispositif technique numérique conçu pour l'explicabilité de ses régulations.

Après avoir rappelé rapidement les notions d'explications dans le domaine de la psychologie et de l'IA, les auteurs listent des *concepts clés* à retenir pour étudier la notion d'explication :

1. Expliquer est un processus continu : Les humains sont motivés pour « comprendre les buts, l'intention, la conscience du contexte, les limitations de la tâche, [et] les bases d'analyse du système pour vérifier si on peut lui faire confiance »[19]. Il s'agit donc de permettre à un humain d'explorer la chaîne explicative de façon à augmenter sa capacité à *apprendre en utilisant.*
2. Expliquer est un processus co-adaptatif : « Les explications améliore la coopération, la coopération permet la production d'explications pertinentes » [20]. Ce n'est plus seulement à l'utilisateur (humain) de s'adapter mais aussi au dispositif qui doit fournir l'explication.
3. Déclencher l'explication : réfléchir à quels sont les déclencheurs d'un besoin d'explication, qui puisse se baser sur le phénomène de « surprise » devant une attente non satisfaite par exemple.
4. Faciliter l'auto-explication : l'auto-explication est le fait de trouver soi-même l'explication avec ou sans aide. C'est l'explication qui fait le plus progresser la connaissance de l'utilisateur. L'exploration personnelle est souvent associée à ce concept.
5. L'explication comme étant une exploration : le récit de l'exploration participe de l'explication, en montrant le cheminement des questions, des réponses, de la complétion des informations pour répondre aux questions [21]
6. Contraster les situations qui s'expliquent différemment : les exemples de situations qui s'expliquent différemment permettent d'illustrer les éléments différents concrets qui illustrent l'explication. A l'explication s'ajoutent les « cas » illustrant telle ou telle situation différente. [22], [23]

Explication et Explicabilité : définitions retenues pour le tutoriel

Explication : processus d'enquête personnelle ou collective pour répondre à une ou plusieurs questions de compréhension d'un phénomène : pourquoi, comment, quoi, quand, où. Nous appellerons ce processus, le processus d'explication.

Explicabilité d'un dispositif numérique de type *Intelligence Artificielle :* capacité de ce dispositif à soutenir un processus d'explication en fournissant les moyens aux utilisateurs de comprendre les modèles et les données exploitées par ces modèles lors de leur mise en œuvre pour une utilisation spécifique.

Les « bonnes propriétés » nécessaires pour un dispositif technique numérique explicable de type Intelligence Artificielle

Dans l'introduction, nous avons précisé ce que nous entendions par Intelligence Artificielle et par Explicabilité, mais il est temps de définir la notion de **dispositif technique** (numérique de type Intelligence Artificielle) : la notion de **dispositif technique** intègre non seulement les logiciels et matériels mobilisés, mais aussi les organisations en amont (recherche, conception), en production (distributeur d'application ou de services, installateur, hébergeur, …) et en aval (utilisateurs individuels et collectifs, organisations d'utilisateurs, organisations de formation et recherche action, …). D'une manière générale, un dispositif technique, implicitement le plus souvent, mobilise toutes les *parties prenantes* de sa mise en œuvre. Ce sont ces parties prenantes qui sont concernés par les processus d'explication que les dispositifs techniques doivent donc accompagner sans pouvoir les modéliser complètement pour autant. Les capacités explicatives d'un dispositif reposent en effet également sur des organisations techniques et humaines, et le type de rapport qu'elles entendent entretenir entre elles. La recherche à réaliser concerne alors aussi bien l'informatique, que les sciences cognitives, le design, et les sciences humaines en général.

Dans le tableau suivant, nous nous livrons à l'exercice classique en recherche d'établir un certain nombre de *bonnes propriétés* qu'un dispositif technique numérique en IA devrait exhiber pour être satisfaisant. Cet exercice réduit la question à une liste de propriétés, alors que nous avons affirmé que le processus d'explication était très contextualisé. C'est donc plutôt des éléments permettant de guider la recherche que des éléments couvrant complètement le sujet.

Capacité à :
[DF] Décrire formellement : Une description technique et formelle du modèle à l'œuvre pour proposer une décision, une recommandation, un diagnostic, ….
[DFS] Décrire formellement à des fins **Scientifiques Il s'agit d'une variante plus exigeante de [DF].** C'est le cas des modèles *du monde* construits à partir de données pour en établir des *lois*
[DAD] Décrire l'apprentissage et les données d'apprentissage : Il s'agit de décrire le dispositif

d'apprentissage et les données mobilisées pour alimenter ce dispositif.
[SM] Simuler le comportement du modèle. Le simulateur peut être décliné pour différents types d'utilisateurs, mais doit permettre à l'utilisateur de « jouer » avec le modèle avec des données proposées par l'utilisateur.
[DP] Décrire Pédagogiquement : L'explication et sa documentation dépendent de ce qu'attend l'utilisateur. D'une certaine façon, c'est ce qui est pris en compte par la notion d'Audience, mais en généralisant à la capacité à s'adapter au registre de compréhension de l'individu.
[OE] Opérationnaliser les éléments d'Explication : Il s'agit ici de rendre visible, documentée et commentée la chaîne explicative pour faciliter l'apprentissage et la discussion des régulations faites sur une tâche en cours. L'opérationnalisation consiste à rejouer les motifs explicatifs dans des conditions similaires, avec adaptation, ou normalisées, avec module pédagogique. Le dispositif possède alors des capacités **d'individuation** issues des interactions. La mise *en intelligence* est à ce prix sans doute.

Focus sur les approches symboliques

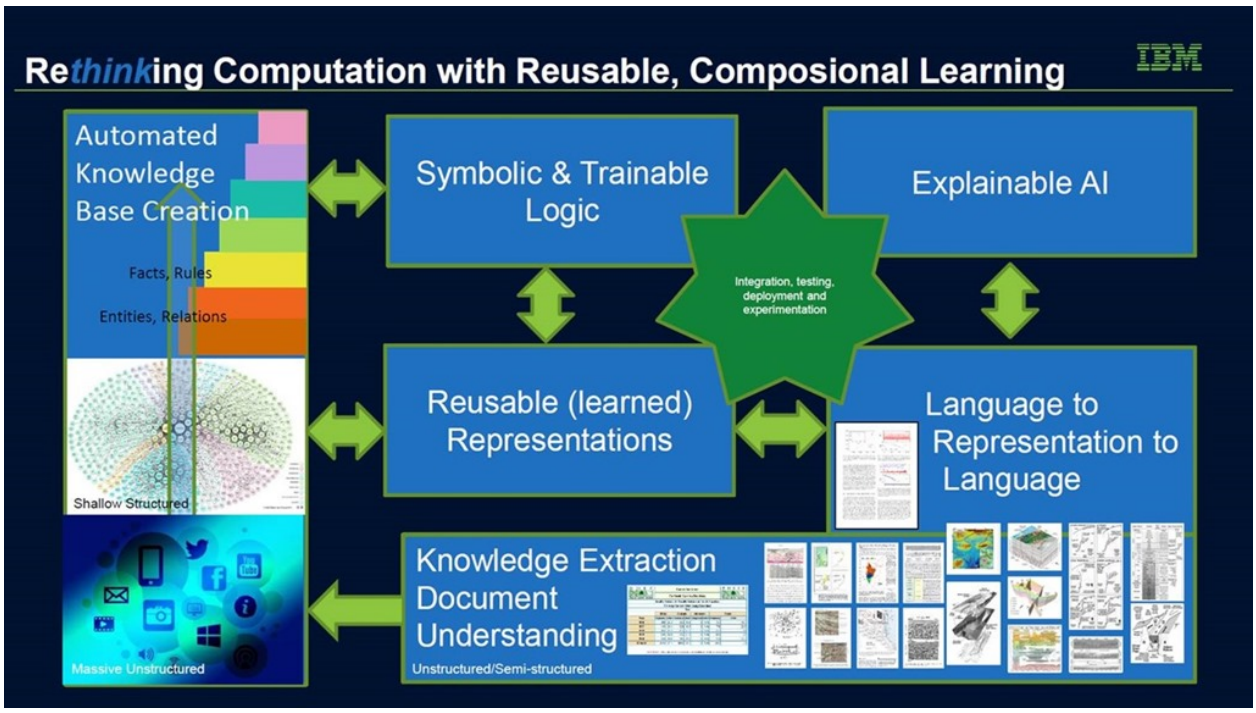
Les approches symboliques sont présentées comme « évidemment » explicables car elles utilisent des représentations *symboliques* des connaissances, permettant un accès facile à l'humain[10], puisque s'exprimant dans un registre symbolique construit par lui.

Pionnières de l'IA, les approches symboliques sont devenues des GOFAI (Good Old Fashion Artificial Intelligence), et sont réputées être supplantées par les approches numériques.

En pratique, les approches sont très étroitement associées, en particulier quand il s'agit de fournir des explications, avec le passage obligé à un niveau symbolique compréhensible dans le registre des connaissances de l'humain. L'image ci-dessous[11] est par exemple proposée par IBM pour associer les approches.

10 *Voir par exemple un article de The Conversation https://theconversation.com/comprendre-lintelligence-artificielle-symbolique-104033*

11 Image tirée du Blog d'Olivier Ezratti : https://www.oezratty.net/wordpress/2018/que-devient-ia-symbolique/ consulté le 2 juillet 2020



Les systèmes experts

Les capacités de raisonnement logique et essentiellement de raisonnement déductif ont fait le succès des *systèmes experts* et un engouement étonnant s'est emparé des médias de l'époque et l'Intelligence Artificielle était alors considérée comme le futur proche de la société. Toute une ingénierie des systèmes expert s'est développée pour les générer à partir de descriptions propositionnelles ou prédicatives symboliques. La *base de connaissance* ainsi constituée est exploitée par un *moteur d'inférences* adapté à la représentation de ces connaissances. L'ensemble des connaissances s'exprime sous une forme symbolique et typiquement sous la forme de *règles* qui peuvent dont être affichées, dont l'enchaînement peut être tracé et donc, d'un certain point de vue, le *raisonnement* est explicable. Toutefois, la période d'euphorie a été relativement courte, une dizaine d'années. Un article de 1995 analyse les raisons de l'échec des systèmes experts [24]. Une section de cet article s'intéresse à la question de l'explicabilité de ces systèmes et l'analyse paradoxalement comme faible et comme une des raisons de leur échec. Cette difficulté à accéder à l'explication et surtout la grande difficulté à maintenir le système quand il se trouvait en situation nouvelle, ont provoqué le désintérêt économique pour cette approche. Le système Mycin[12], très souvent utilisé pour illustrer la notion même de systèmes experts est un célèbre système d'aide à la décision médicale. Mycin démontrait des qualités de diagnostic supérieures aux médecins et même à des collectifs de médecins. Toutefois, il ne fut utilisé que pour les formations. L'incapacité des utilisateurs médecins d'intégrer leurs propres savoirs et connaissances contextuelles dans le dispositif les a éloignés du bénéfice de la qualité du diagnostic, qu'ils ne contestaient pas par ailleurs. Les systèmes experts existent toujours pour des expertises très stables, faiblement sensibles au contexte

12 *http://www.shortliffe.net/Buchanan-Shortliffe-1984/MYCIN%20Book.htm*

d'utilisation, et par exemple des systèmes experts évolués utilisant la réutilisation de l'expérience passée (Raisonnement à partir de cas) ou modérés par des approches probabilistes (c'était le cas de Mycin) ou utilisant les représentations floues. On se souvient peut-être des appareils photographiques proposant un réglage automatique basé sur des représentations en logique floue.dans les années 90. Ce problème lié à la faiblesse des mécanismes d'explication n'a été que faiblement considéré comme sujet de recherche, à l'exception notable de quelques études de l'époque qui en démontraient l'importance, comme [25] dès les années 1985.

Aujourd'hui comme hier, on ne peut pas demander à son appareil photographique pourquoi il a choisi tel ou tel réglage et, si on souhaite reprendre la main pour faire un réglage manuel, toutes les connaissances encapsulées dans le modèle de réglage automatique ne nous seront pas proposées !

De très nombreux mécanismes sont aujourd'hui basés sur des sortes de systèmes experts, fondés sur des connaissances explicites et symboliques agrémentées ou non de modérateurs probabilistes, flous ou possibilistes. Ils sont utilisés pour assister les opérations les plus courantes dans les entreprises mais aussi dans les équipements ménagers, mais, à notre connaissance, ils ne sont pas plus capables d'intervenir dans un processus explicatif que les modèles issus de l'apprentissage numérique.

L'ingénierie des connaissances (symboliques)

Une des grandes difficultés pour la réalisation de systèmes experts, qui se sont ensuite appelés *systèmes à base de connaissance* est la modélisation de la connaissance experte. La modélisation se concrétise sous une *représentation* compatible avec le mécanisme de raisonnement qui exploitera les dites connaissances. C'est Alan Newell qui en 1982 [26], eut l'idée de proposer de séparer la représentation des connaissances des différents types d'exploitation de ces connaissances. Une très importante communauté de recherche s'est alors développée autour de cette question avec ses conférences nationales (IC et EGC en France par exemple) et mondiales (Knowledge Engineering) par exemple. Cette approche devait permettre aux *experts* de gérer les connaissances sans être des spécialistes informatiques et de permettre non seulement le raisonnement mais le partage et la diffusion des connaissances.

L'objet assurément le plus connu sorti de ces travaux est *l'ontologie.* Une ontologie est une description plus ou moins formelle d'un domaine de connaissance. En pratique, c'est un graphe de nœuds concepts reliés entre eux par des relations exprimant leur sémantique réciproque. Cette fois, on peut imaginer que l'on a l'outil absolu pour expliquer les connaissances mobilisées dans un raisonnement ? C'est ainsi qu'un domaine économique s'est développé dans les années 2000 autour du concept de *Gestion des connaissances* (dans l'entreprise surtout). La communauté de l'ingénierie des connaissances est maintenant très active dans le domaine du web, avec les promesses du *web sémantique.* Le lecteur intéressé pourra manipuler directement de tels graphes en visitant Dbpedia[13].

La *déclaration* des connaissances par celles et ceux qui les ont reste difficile et de plus en plus de graphes de connaissances sont *automatiquement* créés en exploitant les pages

13 https://wiki.dbpedia.org/

documentaires disponibles sur le web. La base de connaissances est alors *apprise* avec des méthodes d'apprentissage à partir de ces corpus textuels [27].

Focus sur les approches numériques

Parmi les approches d'Intelligence Artificielle, les approches numériques consistent à faire apprendre automatiquement un modèle des données à partir de données qui lui sont fournies. À partir de ce modèle, il est alors possible de faire des prédictions sur des données similaires (dans le cas d'une classification ou régression), ou bien d'agir dans un environnement (dans le cas d'une politique d'action). C'est typiquement dans ce cas qu'une régulation est appliquée au niveau « méta » sans être explicitée dans le modèle des données, même si on a paramétré l'apprentissage avec le biais de la régulation « méta ».

Les récents travaux dans ce domaine du *Machine Learning* ont montré à la fois des résultats impressionnants, dépassant les capacités humaines à réaliser la même tâche (voir l'article de média [28] pour des exemples ou plus particulièrement l'article [29] pour le domaine de la reconnaissance d'image) et en même temps capable de faire des erreurs qui seraient triviales pour un humain. Par exemple, dans l'illustration 8 (issue de [17]), le modèle est capable de prédire correctement une image floue, mais en revanche le résultat est aberrant en rajoutant quelques artefacts sur l'image, ou en accentuant les traits de l'image. Il est intéressant de remarquer que, dans le cas des artefacts, le modèle prédit (à tort) avec un score de confiance assez élevé ; ce dernier ne paraît donc pas suffisant pour conforter un opérateur humain dans sa décision de faire confiance à la machine ou de déléguer son jugement.

Figure 1. Some examples of Deep Net classification.

Illustration 8: Exemples d'erreurs de classification

Dans les approches numériques, le processus d'explication peut servir deux objectifs, selon l'audience considérée :

1. Permettre à un opérateur humain qui utilise un modèle numérique obtenu par apprentissage automatique de comprendre le « raisonnement » qui a mené à cette décision, afin de juger si ce raisonnement lui paraît crédible (donc acceptable), ou non.

2. Permettre aux concepteurs, superviseurs ou régulateurs de comprendre les causes d'une erreur de décision, afin de fournir des données palliant à ce problème pour ré-entraîner le modèle.

L'objectif 1. est particulièrement important dans le cas d'une décision avec des conséquences (ou enjeux) importantes. Citons par exemple le cas d'un juge devant décider de relâcher un détenu ou de le remettre en prison [30]. Ce juge pourrait s'appuyer sur un modèle prédictif qui, à partir des informations disponibles sur le détenu (son historique de crimes et délits, son comportement en prison, etc.), informerait le juge sur le risque de récidive dans le cas où le détenu serait relâché. Un modèle n'intégrant pas de processus d'explication et possédant un biais aurait des conséquences néfastes immédiates, en influençant le juge pour garder en prison les personnes d'une certaine race, malgré un risque réel inférieur à l'estimation prédite pour des personnes relâchées. Un dispositif numérique intégrant un processus d'explication serait en mesure de fournir au juge non seulement une prédiction sur le risque de récidive, mais également une explication du raisonnement qui a mené à cette prédiction Quelles sont les variables importantes dans l'ensemble des données disponibles ? Quels cas similaires dans le jeu de données ont été utilisés ?. Le juge peut alors évaluer ces raisons et détecter un biais (ou non), et ainsi approuver (ou non) la décision de la machine.

La grande diversité des modèles numériques disponibles appelle une grande offre de mécanismes pour produire de l'explication dans ces modèles. [31] propose de catégoriser ces mécanismes en 4 types de problèmes à résoudre différents, que nous décrivons par la suite.

Problème type 1 : Explication du modèle

Considérant un modèle numérique généré à partir d'un jeu de données et non facilement interprétable (aussi appelé « opaque » ou « boîte noire » - *black box model*), le problème consiste à produire une explication dite « globale » de ce modèle opaque, sous la forme d'un 2ème modèle numérique (dit « transparent », par opposition au modèle « opaque »), qui doit être interprétable et compréhensible par les humains. Ce modèle transparent est appris à partir du modèle opaque par rétro-ingénierie (c'est-à-dire en ne considérant que les données d'entrées X et la sortie Y du modèle opaque, sans s'intéresser au processus de calcul interne).

Ce modèle transparent est construit avec pour objectif de maximiser sa ressemblance avec le modèle opaque (i.e. les décisions doivent être les mêmes) et en conservant des performances similaires (i.e. les décisions sont proches du *ground truth* établi). Toutefois, ils diffèrent fondamentalement par leur complexité d'interprétation des modèles dits « opaques ».

Un exemple de type de modèle utilisable comme « transparent » selon les auteurs est le classifieur par règle de décisions (voir l'illustration 9). L'explication de la décision est alors produite par lecture de ce modèle, supposé transparent pour un humain.

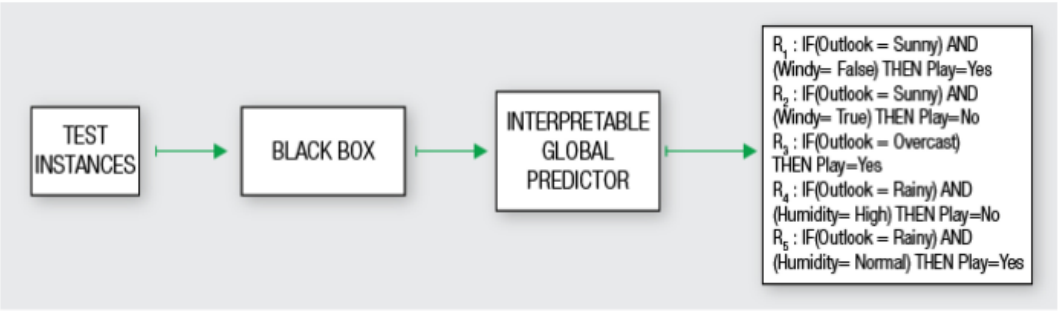


Fig. 6. Model explanation problem example. Starting from test instances in X, first query the black box, and then extract an interpretable global predictor from $\{X, b(X)\}$ in the form of a decision rule classifier.

Illustration 9: Schéma de production d'un modèle proxy global

Cette approche souffre d'un inconvénient majeur : le modèle transparent est produit sans considérer la méthode de calcul du modèle opaque. On peut espérer, avec un jeu de données suffisamment grand, que les règles de décision produites soient plus ou moins cohérentes avec les calculs qui ont été effectués par le modèle opaque pour parvenir à cette prédiction, mais nous n'avons pas de garantie. De plus, la question de l'interprétabilité d'un tel modèle (supposé suffisante dans la plupart des travaux) dépend en réalité de la taille d'un tel modèle et de l'audience, i.e. l'humain qui « lira » ce modèle. En effet, un arbre de décisions de plusieurs centaines de nœuds ne semble pas facile à interpréter ; de plus, selon les variables utilisées, il est possible qu'un utilisateur non-expert dans le domaine d'application ne soit pas capable de comprendre les décisions, bien qu'il soit effectivement capable de « lire » le modèle.

Notons qu'un modèle similaire, LIME [32] propose une alternative permettant d'obtenir une explication plus fidèle. LIME (*Local Interpretable Model-agnostic Explanations*) a également pour but de produire un modèle « proxy » interprétable à partir d'un modèle opaque, mais propose une démarche « locale » plutôt que « globale » - c'est-à-dire en ne considérant comme données d'entrée que les instances voisines d'un certain cas d'intérêt. L'hypothèse sous-jacente est qu'il est plus aisé de construire un modèle fidèle au modèle opaque sur un sous-ensemble de cas plutôt que sur la totalité. Cela soulève tout de même la question du « voisinage » : comment déterminer un voisinage intéressant autour d'un cas donné ? Combien de voisins faut-il considérer ?

Ce problème est fréquemment traité dans la littérature, et repris sous d'autres noms par d'autres travaux, citons par exemple [33] qui le nomme « Explanation by simplification ».

Ce problème se range dans une catégorie plus générale, nommée « post-hoc explainability » [34], regroupant les approches dont le but est de produire une explication après avoir obtenu un modèle par apprentissage. Le modèle est généralement considéré comme une boîte noire (comme c'est le cas dans ce problème précis) dont les caractéristiques ne sont pas utilisées ; cela a pour conséquence de fournir une méthode indépendante du type de boîte noire (i.e. *Model-Agnostic*), une propriété intéressante pour généraliser à une variété d'approches numériques, mais d'un autre côté la fiabilité de l'explication vis-à-vis du calcul réellement effectué par le modèle n'est pas garantie, comme nous l'avons mentionné.

Problème type 2 : Explication du résultat

Toujours en considérant un modèle opaque appris à partir d'un ensemble de données, ce problème consiste à produire une explication sur un résultat précis (i.e. un unique *y* tel que *f(x) = y*, avec *f* le modèle et *x* une instance précise parmi le jeu de données).

Fig. 7. Outcome explanation problem example. For a test instance x, the black box decision $b(x)$ is explained by building an interpretable local predictor c_I, e.g., a decision rule classifier. The local explanation $\varepsilon_I(c_I, x)$ is the specific rule used to classify x.

Illustration 10: Schéma de production d'un modèle proxy local

De la même manière que pour le problème 1, il est possible de générer un modèle « proxy » tel que des règles de décision (illustré par 10). D'autres mécanismes sont possibles, par exemple en modifiant l'une des variables d'entrée x_i et en observant l'impact sur la sortie *y* (méthode par contre-factuel).

Les cas d'applications de ce problème sont typiquement ceux visés par la question du « droit à l'explication » que l'on peut trouver dans le RGPD de l'Union Européenne (Préambule 71 [35]) :

> En tout état de cause, un traitement de ce type devrait être assorti de garanties appropriées, qui devraient comprendre une information spécifique de la personne concernée ainsi que le droit d'obtenir une intervention humaine, d'exprimer son point de vue, **d'obtenir une explication quant à la décision prise** à l'issue de ce type d'évaluation et de contester la décision.

Cela s'applique particulièrement au cas du juge que nous mentionnions précédemment : un individu jugé comme « dangereux » par un système automatique devrait avoir le droit de comprendre les raisons qui ont mené à cette décision ; les règles « globales » qui s'appliquent à l'ensemble du jeu de données sont superflues dans son point de vue.

Il est à noter que ce problème d'explication du résultat est également appelé « explication locale » par [33]. L' « explication par l'exemple » est une autre approche, similaire bien que formulée différemment, utilisée notamment par [36]. Cette méthode consiste à extraire depuis le jeu de données des cas (exemples) qui se rapprochent du cas étudié ; cela est similaire à l'une des façons que nous avons (en tant qu'humains) de produire des explications : présenter une illustration afin de mettre en valeur un point important.

En reprenant l'exemple du juge, il serait possible d'expliquer une décision en fournissant des cas similaires ayant déjà été jugés (et qui ont pu être évalués post-jugement).

La « jurisprudence » permet en effet d'adapter les explications d'un cas complet pour un cas à prédire. Dans le domaine du raisonnement à partir de cas, il existe une importante littérature sur le thème Explanation Based Learning, reprise actuellement d'ailleurs dans la communauté CBR

→ voir les travaux de Kass, encore régulièrement cités… → https://www.sciencedirect.com/science/article/pii/B9781558600362500199

De même que pour l'explication du modèle (l'explication « globale »), ce problème est généralement considéré dans la catégorie « post-hoc ». Certaines méthodes considèrent en effet le modèle comme une boîte noire et n'utilisent aucune de ses propriétés pour tenter d'expliquer la décision (e.g. la méthode « explication par l'exemple », qui ne repose que sur le jeu de données). La question de la fidélité de l'explication par rapport au calcul réellement effectué se pose donc.

Problème type 3 : Inspection du modèle

Contrairement aux deux problèmes précédents, celui-ci ne s'intéresse que peu aux décisions produites par le modèle étudié ; l'objectif est de fournir une représentation (visuelle ou textuelle) de certaines propriétés du modèle (voir l'illustration 11).

Ces propriétés peuvent correspondre, par exemple, à la sensibilité du modèle face à un changement dans les attributs, ou encore l'importance d'un attribut par rapport aux autres (ce que l'on peut rapprocher de la méthode *feature relevance explanation* dans [33]) ; l'identification des composants internes d'un modèle responsables de certaines décisions (ou contextes de décisions) est également une méthode envisageable par rapport à ce problème.

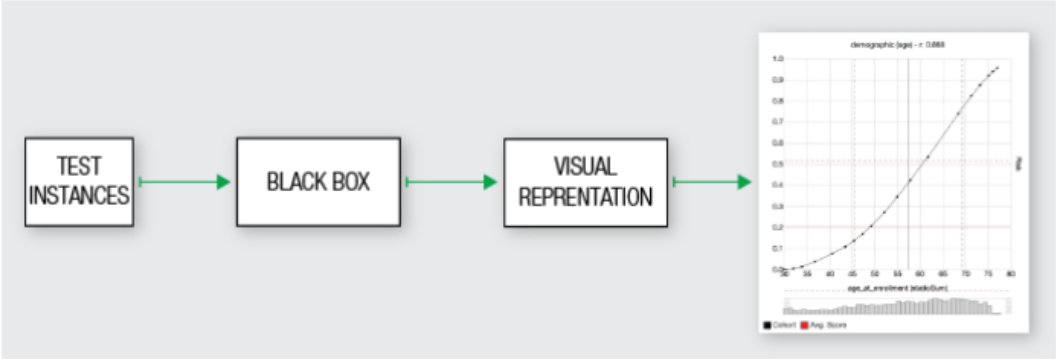


Fig. 8. Model inspection problem example. Query the black box on test instances X, and then extract a sensitivity analysis plot.

Illustration 11: Visualisation graphique des propriétés d'un modèle opaque

Ce problème offre l'avantage, par rapport aux deux précédents, de mieux comprendre le fonctionnement interne d'un modèle boîte noire ; par exemple, nous pourrions identifier qu'une variation de la variable « âge » a pour conséquence un changement de décision de la part du modèle sur la plupart des cas considérés ; il semble donc que la variable « âge » soit un élément clé du processus de décision.

En revanche, cela ne permet pas de fournir directement d'explication sur une décision donnée : en observant le cas d'entrée X et la décision Y, nous pourrions avoir l'intuition que la variable « âge » a le plus influé sur la décision, mais cela ne nous permet pas de comprendre *pourquoi* cette variable est considérée importante par le modèle ? Quel est le raisonnement ou calcul qui a mené à mettre autant de poids sur cette variable et pas les autres ?

Problème type 4 : Conception d'un modèle transparent

Ce problème diffère fondamentalement des trois autres, dans le sens où nous ne considérons pas de modèle boîte noire : le but est d'apprendre directement un modèle considéré comme transparent. Les problématiques d'explication du résultat, de la décision, ou de l'inspection du modèle sont donc déjà résolues (le modèle étant transparent et donc, dans cette hypothèse, facile à lire, interpréter et comprendre).

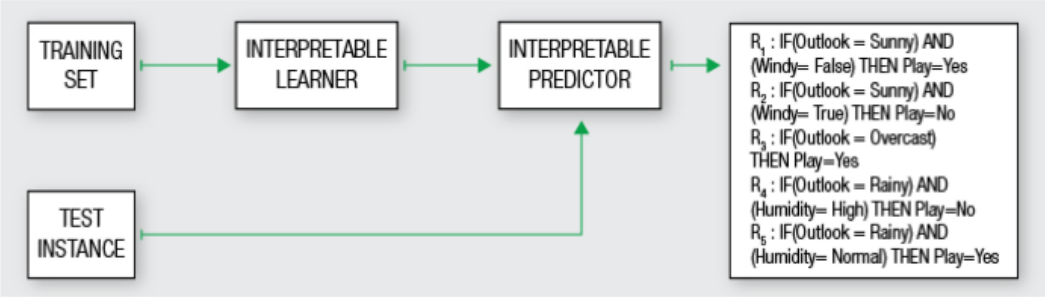


Fig. 9. Transparent box design problem example. A decision rule classifier learned from a training dataset is globally interpretable predictor. Moreover, the rule that applies on a given test instance is a local explanation of the predictor's decision.

Illustration 12: Production d'un modèle transparent

[33] propose une liste de modèles, y compris numériques, qui sont considérés comme transparents :

- Régression linéaire / logistique
- Arbre de décisions
- K-Plus proches voisins
- Rule-based Learners
- General Additive Models
- Modèles Bayésiens

Ce problème s'oppose directement à la catégorie « post-hoc » dont font partie les 3 autres problèmes ; il est également nommé « Interprétabilité Intrinsèque » par [34]

Comme nous l'avons déjà mentionné dans le Problème 1 (qui consiste à produire un de ces modèles « transparents » à partir d'un modèle « opaque »), la notion d'interprétabilité est souvent considérée comme acquise dès lors que l'on utilise un de ces modèles, dans la littérature étudiée. Or, un arbre de décisions de plusieurs centaines de nœuds n'est pas forcément facile à interpréter (bien qu'interprétable car il suffit de tracer le chemin selon les règles à chaque nœud). De même, ces modèles sont souvent interprétables par les chercheurs ou experts du domaine qui conçoivent l'application (car ils ont des connaissances techniques), mais pas par les utilisateurs finaux qui sont bien souvent dépourvus de ces connaissances.

Nous pouvons citer [37] qui recommande l'utilisation de ces modèles transparents dans le cas de décisions avec des enjeux importants (comme l'exemple du juge que nous avons repris plusieurs fois au fil de cette section).

La question de la complexité

Les problèmes que nous avons présentés (et les méthodes qui en relèvent) se focalisent sur la production d'une explication (ou d'un modèle interprétable), sans considérer le processus d'explication dans son ensemble.

En effet, une explication n'a de sens que dans le contexte du récepteur ; en particulier, une explication n'est efficace que si elle est compréhensible par son audience. Nous avions déjà mentionné le fait qu'un modèle « interprétable » n'est pas facile à interpréter s'il contient trop d'informations ; [38] propose d'aller plus loin en quantifiant ces informations (qu'il nomme des *cognitive chunks*). Cela permet ainsi de quantifier l'explicabilité d'une méthode à partir des *cognitive chunks* qu'elle produit. Les auteurs proposent une formule en fonction de :
- Ni = le nombre de *cognitive chunks* en entrée
- No = le nombre de *cognitive chunks* en sortie (dans la représentation de l'explication)
- I = l'interaction (corrélation, causalité, etc.) entre les différents *cognitive chunks*
- w1, w2, w3 = des poids pour ajuster chaque partie de la formule selon le cas d'application considéré

$$E \stackrel{\text{def}}{=} \frac{w_1}{N_i} + \frac{w_2}{N_o} + w_3(1-I)$$

Le chunking est un découpage efficace de l'information pour la mémoriser (travaux sur la mémoire de Miller en 1956!) c'est donc plus pratique pour présenter les choses « reconnaissables » à un utilisateur selon sa propre expérience (mnésique). Comme il s'agit dans les travaux de « cognitive chunk » de fabriquer des « chunks génériques », on ne voit pas comment ils pourraient convenir à la mémoire individuelle de chaque utilisateur ? Ce qu'il semble, c'est que le fait d'avoir une *formule* permet de *mesurer* la « qualité » d'une explication est que c'est le Graal recherché dans les étapes en amont de l'utilisation. Le nombre de variables (5) semble vraiment petit par rapport à la combinatoire des situations à expliquer ?

La question de l'audience

En plus de la question de la complexité de l'explication, l'explication (ou plutôt le processus explicatif) doit s'adapter à l'audience. En effet, il existe différents acteurs ayant chacun un point de vue différent sur les modèles numériques d'apprentissage et qui ne poursuivent pas le même but ; l'explication ne devrait donc pas être la même.

Afin de traiter cette question de l'audience, [39] propose d'utiliser des « scénarios » comme base de réflexion pour concevoir des systèmes explicables (ou contenant un processus explicatif). Ces scénarios se focalisent sur ce dont les personnes ont besoin de comprendre à propos de ces systèmes afin de pouvoir agir dessus. Cette réflexion prend le contre-pied de ce que nous avons présenté jusque-là, en réfléchissant avant tout sur ce dont les acteurs humains ont besoin au lieu de considérer ce que les systèmes peuvent offrir comme explication. Cela n'est pas sans rappeler la démarche de conception centrée sur l'utilisateur et notamment la technique des *persona*.

Une typologie des différentes audiences possibles (en les catégorisant notamment selon leurs buts et attentes) est proposée par [33] (voir la table ci-dessous et les illustrations 1 et 13) :

Audience	**Dans quel but**
Experts du domaine Utilisateurs du modèle	Faire confiance au modèle Gagner de la connaissance scientifique
Utilisateurs affectés par le dispositif numérique	Comprendre leur situations Vérifier que la décision est juste
Chercheurs Développeurs Chef de projet	S'assurer de l'efficacité du produit L'améliorer Ajouter des fonctionnalités
Manageurs Comité de direction	S'assurer que le produit respecte les normes Comprendre les applications potentielles
Agences ou entités de régulation	Certifier que le produit respecte la législation Passer un audit

Table 1
Goals pursued in the reviewed literature toward reaching explainability, and their main target audience.

XAI Goal	Main target audience (Fig. 2)	References
Trustworthiness	Domain experts, users of the model affected by decisions	[5,10,24,32–37]
Causality	Domain experts, managers and executive board members, regulatory entities/agencies	[35,38–43]
Transferability	Domain experts, data scientists	[5,21,26,30,32,37–39,44–85]
Informativeness	All	[5,21,25,26,30,32,34,35,37,38,41,44–46,49–59,59,60,63–66,68–79,86–154]
Confidence	Domain experts, developers, managers, regulatory entities/agencies	[5,35,45,46,48,54,61,72,88,89,96,108,117,119,155]
Fairness	Users affected by model decisions, regulatory entities/agencies	[5,24,35,45,47,99–101,120,121,128,156–158]
Accessibility	Product owners, managers, users affected by model decisions	[21,26,30,32,37,50,53,55,62,67–71,74–76, 86,93,94,103,105,107,108,111–115,124,129]
Interactivity	Domain experts, users affected by model decisions	[37,50,59,65,67,74,86,124]
Privacy awareness	Users affected by model decisions, regulatory entities/agencies	[89]

Illustration 13: Question de l'audience visée par le processus d'explication (selon Barredo et al.)

On pourra trouver un exemple détaillé dans[40].

Focus sur le Deep Learning

Les travaux utilisant des approches de Deep Learning se rapprochent des travaux numériques de manière générale ; en particulier, nous retrouvons les mêmes problèmes que ceux mentionnés précédemment. Ainsi, les solutions présentées pour ces problèmes, e.g. l'explication par production d'un modèle transparent, sont également utilisées pour les approches de Deep Learning.

Toutefois, le nombre extrêmement élevé d'éléments numériques impliqués, c'est-à-dire les *paramètres* du réseau – plusieurs millions pour les modèles les plus récents – rendent

l'interprétation du modèle produit plus complexe encore. Bien que le modèle soit représentable par une fonction mathématique, les propriétés statistiques extraites du jeu de données, donc la connaissance produite, sont abstraites dans ces millions de paramètres et donc difficilement accessibles, ne serait-ce qu'aux concepteurs du modèle, sans parler des utilisateurs non-experts, des décideurs, des régulateurs, etc.

Ce nombre de paramètres, ainsi que les influences entre ces paramètres, de par l'architecture en « couches » connectées entre elles, rend la question de la fidélité, i.e. à quel point l'explication fournie est correcte vis-à-vis du calcul réellement effectué, d'autant plus importante.

Explication post-hoc et agnostique

Les méthodes purement post-hoc, qui se situent après la génération du modèle, et qui n'utilisent pas d'informations sur le modèle sont sensiblement les mêmes que présentées précédemment, dans « Problème 1 : Explication du modèle ». En effet, le modèle en lui-même n'est pas utilisé, seules les entrées et sorties (la prédiction ou action) sont considérées.

Comme nous l'avions déjà mentionné, la question de la fidélité se pose avec ces approches ; les calculs effectués par le modèle de Deep Learning peuvent contenir des biais non détectés par ces méthodes post-hoc.

Pour rappel, un exemple de modèle « transparent » est un arbre de décision avec des règles du type « IF (Outlook == Sunny) AND (Windy = False) THEN Play = yes ». Typiquement, ces modèles ne sont pas capables de « passer à l'échelle », c'est-à-dire que si l'on considère des dizaines voire des centaines de données en entrée, le modèle sera difficile à générer et d'autant plus à interpréter.

Or, les modèles de Deep Learning, en leur qualité d'approximateurs de fonction universels, sont particulièrement adaptés et de ce fait fréquemment utilisés à l'utilisation de nombreuses données en entrée. Nous détaillons plusieurs cas dans la section suivante.

Explication sur de nombreuses entrées

Deux sous-domaines du Deep Learning sont particulièrement concernés par l'utilisation de nombreuses données d'entrées : premièrement, la Vision par Ordinateur, i.e. les travaux qui traitent des données de type image ou vidéo, dans lequel les pixels d'une image composent les données fournies ; deuxièmement, le Traitement Automatique du Langage, i.e. les travaux qui utilisent du texte en langage naturel, dans lequel des mots, phrases, paragraphes ou documents entiers composent les données d'entrée.

Notons que les images sont le plus souvent traitées à l'aide de CNNs – *Convolutional Neural Networks* – tandis que le texte est quant à lui souvent traité à l'aide de RNNs – *Recurrent Neural Networks*. Les caractéristiques de ces réseaux de neurones peuvent être exploitées afin de créer un processus d'explication, ou, *a minima*, un élément explicatif. Ainsi, plusieurs travaux proposent notamment de visualiser les neurones à l'intérieur, par exemple les poids ou motifs d'activation afin de représenter graphiquement le calcul opéré par le réseau dans son ensemble.

Les CNNs notamment sont souvent visualisés à l'aide de cartes de saillance, ce qui permet (en reliant ces cartes aux images du jeu de données d'entrée) de « voir » quelles zones de l'image ont participé à la décision (voir illustration 14 pour un exemple).

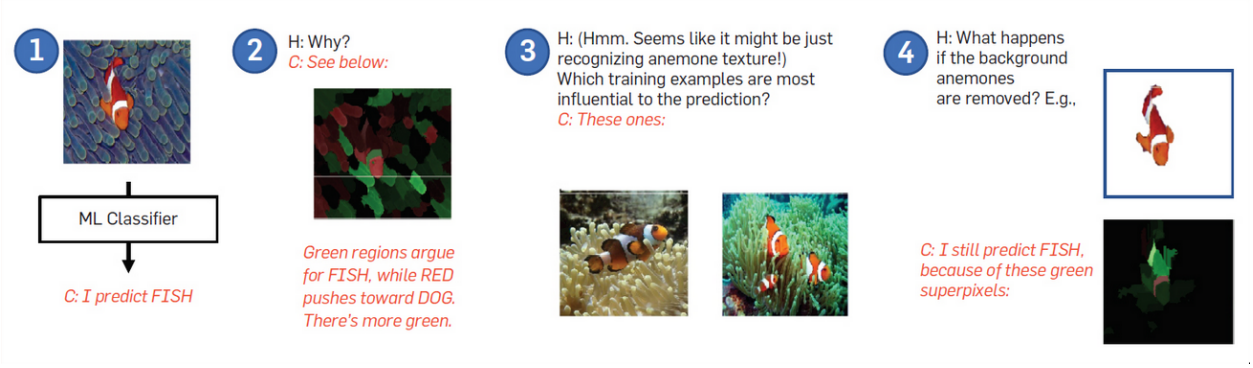


Illustration 14: Exemple fictif de processus d'explication interactif avec, entre autres, une visualisation des neurones ayant participé à la classification (image tirée de [43]). Une légende (ou explication textuelle) est également générée et permet le dialogue avec un opérateur humain.

Il est également possible de visualiser les motifs d'activation (selon un aspect temporel) des neurones, principalement dans le cas des RNNs, mais également dans le cas des CNNs qui traitent des vidéos plutôt que des images.

Ces éléments d'explication sont intéressants mais devraient être intégrés dans un processus explicatif, en intégrant les problématiques déjà mentionnées, telles que la question de l'audience ou la capacité d'interaction de l'utilisateur.

En effet, il peut être nécessaire par exemple de demander à reformuler l'explication ou encore à proposer une explication alternative, mais portant sur le même sujet. Dans l'illustration 14, supposons que l'utilisateur ne soit pas convaincu par l'explication (2) : il y a en effet beaucoup de régions en rouge dans l'image. L'utilisateur peut-il demander une autre raison à cette même classification ? Ou demander, par exemple, le compte du nombre de pixels rouges et verts dans cette explication-ci ?

Boîtes à outils

Suite à la recrudescence du nombre d'articles portant sur l'Explicabilité en IA et la prépondérance des techniques de Deep Learning, plusieurs entreprises ont récemment proposé des outils « clé en main » pour intégrer des notions d'Explicabilité dans leurs dispositifs techniques.

Ces solutions sont utiles pour les développeurs afin d'intégrer simplement et rapidement des éléments explicatifs, mais restent limitées (pour l'instant ?), notamment dans leur traitement de l'audience et de l'interaction. Nous détaillons par la suite deux de ces boîtes à outils pour donner l'exemple : TensorFlow *What-If* et IBM *AIX360*.

TensorFlow What-If

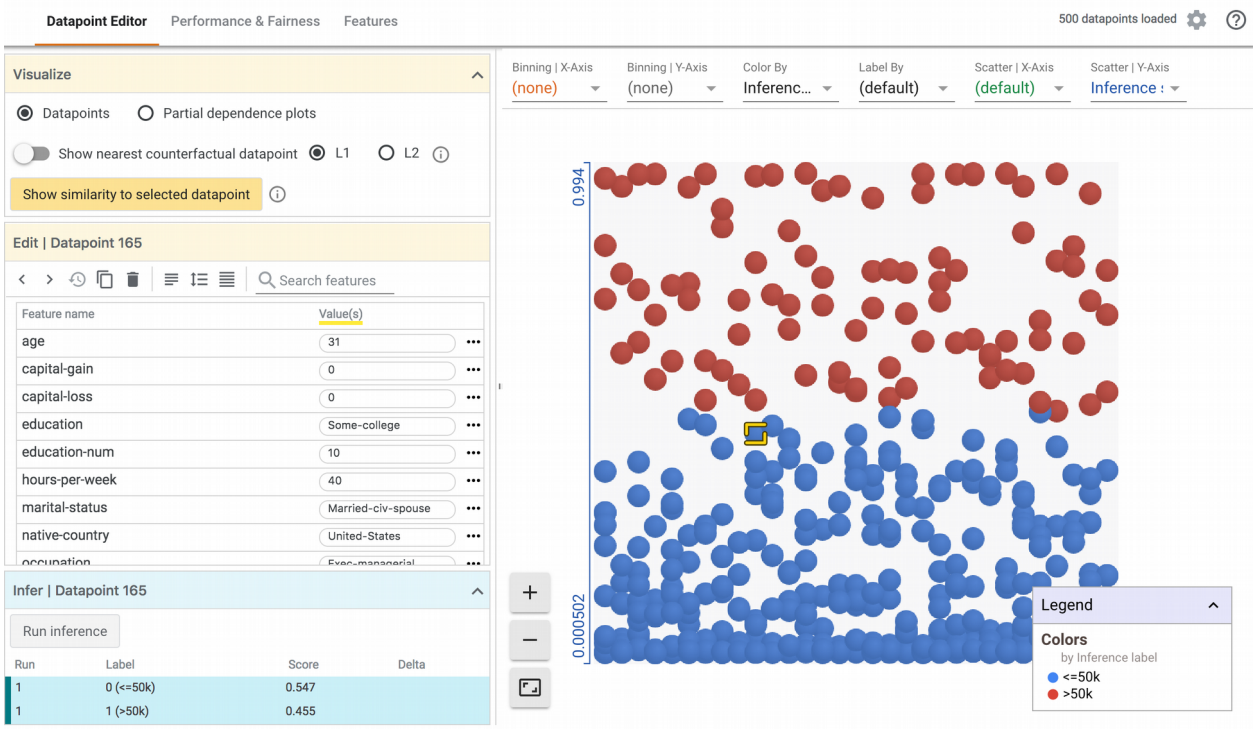


Illustration 15: L'interface du What-If Tool, permettant de modifier temporairement les variables d'un point parmi le jeu de données. Image tirée de https://www.tensorflow.org/tensorboard/what_if_tool

TensorFlow est une des bibliothèques les plus utilisées pour les modèles de Deep Learning ; en plus de faciliter la création de réseaux de neurones, TensorFlow inclut un outil de visualisation du modèle, nommé TensorBoard. À l'origine, TensorBoard permettait de visualiser les couches du réseau de neurones sous forme de graphe, de modifier certains paramètres, et d'observer les métriques telles que la précision de la classification, l'erreur moyenne, etc.

Récemment, TensorBoard a été augmenté d'un nouvel outil : *What-If Tool* (WIT), permettant de mieux comprendre le modèle. Cet outil permet principalement aux développeurs d'explorer le fonctionnement interne de leur modèle, par exemple en permettant de modifier temporairement une donnée en particulier et d'observer le résultat sur la prédiction du modèle, i.e. en utilisant des contre-factuels.

Ainsi, WIT a été conçu avec un objectif d'équité en particulier : l'outil permet, pour reprendre l'exemple du Juge, de prendre un cas parmi tous les cas dans le jeu de données, et d'explorer, par modification de ses variables, ce qui fait pencher le jugement d'un côté ou de l'autre, ou encore de voir les cas similaires. On peut objecter que l'équité, bien qu'étant un objectif important, n'est pas le seul objectif de XAI...

Le principal avantage est la capacité d'exploration : l'humain contrôle et décide de ce qu'il veut voir pour se construire une représentation mentale du modèle. Toutefois, l'inconvénient est le manque d'adaptation à l'audience. Cet outil est en effet proposé principalement aux développeurs : l'interface graphique (voir illustration 15) présente seulement la liste des caractéristiques de chaque donnée « en brut ».

Le système IBM AIX360

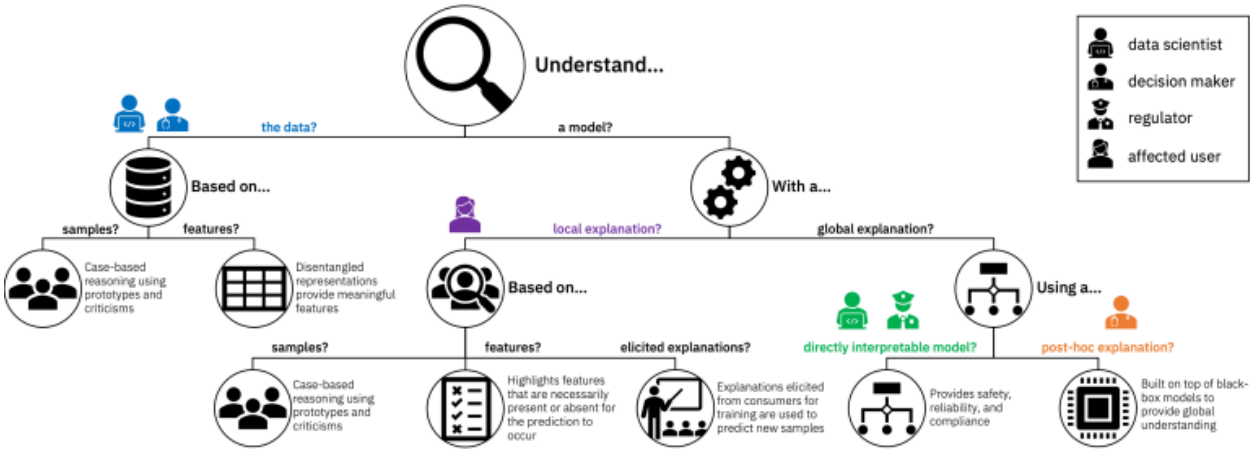


Illustration 16: La multitude de façons de permettre la compréhension, pour des audiences différentes, selon IBM. Image tirée de https://www.ibm.com/blogs/research/2019/08/ai-explainability-360/

AI Explainability 360 [41], contrairement à WIT, contient plusieurs algorithmes, ayant chacun un but différent. Le schéma sur l'illustration 16 montre les choix que l'on peut faire pour sélectionner un algorithme, par exemple si l'on souhaite comprendre le modèle sous forme d'explications locales, donc spécifiques à un ou quelques exemples, ou globales, donc relatives à l'ensemble des données disponibles. La question de l'audience est également abordée en considérant quatre types d'utilisateurs. Ces utilisateurs correspondent à ceux que nous avons mentionné précédemment, à l'exclusion des « experts du domaine » qui ne sont pas présents dans ce schéma.

En revanche, l'application de démonstration (http://aix360.mybluemix.net/data) montre peu voire pas d'interaction pour l'utilisateur. Par exemple, le client d'une banque voyant sa demande de prêt refusée a accès aux variables importantes à changer, mais ne peut pas faire une simulation pour observer l'impact du changement d'une variable ou d'une autre.

Focus sur les approches émergentistes

Le terme *émergentiste* est lié au fait que le DTN-IA repose sur des approches en général bio-inspirées. L'algorithme cherche à construire des comportements *situés* pour *satisfaire* des critères donnés. Ces critères peuvent être fournis par le régulateur du DTN-IA afin d'obtenir un

résultat satisfaisant pour lui, ou, dans les approches développementales, le critère à satisfaire est intrinsèque à l'agent constitué par le DTN-IA.

Ce sont les capacités d'interaction avec l'environnement qui sont conçues pour atteindre les comportements *satisfaisants* attendus ou les comportements caractéristiques de l'agent à motivation intrinsèque.

Le lecteur intéressé pourra consulter le rapport réalisé par un groupe de travail de l'AFCET et de l'AFIA qui a tenté de caractériser la notion d'émergence et l'a utilisé pour présenter le domaine des Systèmes Multi-Agents. L'article est signé par M.R. Jean, non sans malice ![42]

La question de l'explicabilité est bien sûr très difficile dans ce cas, puisque les comportements émergent du dispositif à partir d'éléments nombreux et sont contingents à la situation rencontrée. Un article contemporain du précédent fait le point sur cette question de l'explication dans les systèmes émergentistes dans la revue Intellectica [44].

Il est possible d'expliquer le comportement de l'algorithme *émergentiste* par la description du modèle choisi pour que l'agent DTN-IA puisse *construire* ses comportements de manière adaptée à ses objectifs et dans le contexte interactionnel dans lequel il est plongé. Ces algorithmes sont d'ailleurs bien documentés dans la littérature et sont *génériques* d'un type d'approche émergentiste (Multi-Agents réactifs, génétiques, évolutionnaires, par renforcement, développementaux,…).

Mais expliquer le comportement qui émerge est une toute autre affaire. En effet, ce qui caractérise l'émergence c'est son imprédictabilité (même quand ce qui émerge est un résultat attendu), et donc ne pouvant pas s'expliquer *a priori.*

Pour *comprendre* ce qui advient, il faut *tracer* les modifications qui interviennent *à l'intérieur de l'agent* et mettre ces traces dont la sémantique est strictement technique en relation avec l'observation humaine du comportement lui-même tracé, mais cette fois dans le registre symbolique de l'humain.

L'humain reformule le comportement avec des phrases du type *« le chemin de fourragement se construit et est emprunté par les agents qui gagnent en performance dans leur tâche »...*ou encore « *l'agent évite maintenant les obstacles et a appris à reconnaître les ressources qui lui sont nécessaires... ».* Les traces externes sont en général faciles à comprendre par définition, mais les traces internes sont très souvent *ad hoc* et imaginé pour rendre compte de l'évolution des structures internes de l'agent. L'exemple suivant [Illustration 17]de traces internes et externes est tiré de [45].

Nous pouvons considérer que l'explication à partir de ce qui se passe en interne est pour le moins *délicate.* Les explications données à l'extérieur empruntent leur vocabulaire à la manière dont un humain décrirait un organisme « naturel ». De la même façon, dans le cas d'un dispositif à base de *fourmis*, c'est le comportement de la *fourmilière* qui devient prédictible dans la durée et l'observation permet d'expliquer comment cette fourmilière résout un problème qui lui est posé.

Dans ces systèmes, ce sont le plus souvent des simulateurs comportementaux qui permettent d'accéder à des éléments d'explication, le plus souvent sans chercher à *expliquer* le fonctionnement interne détaillé.

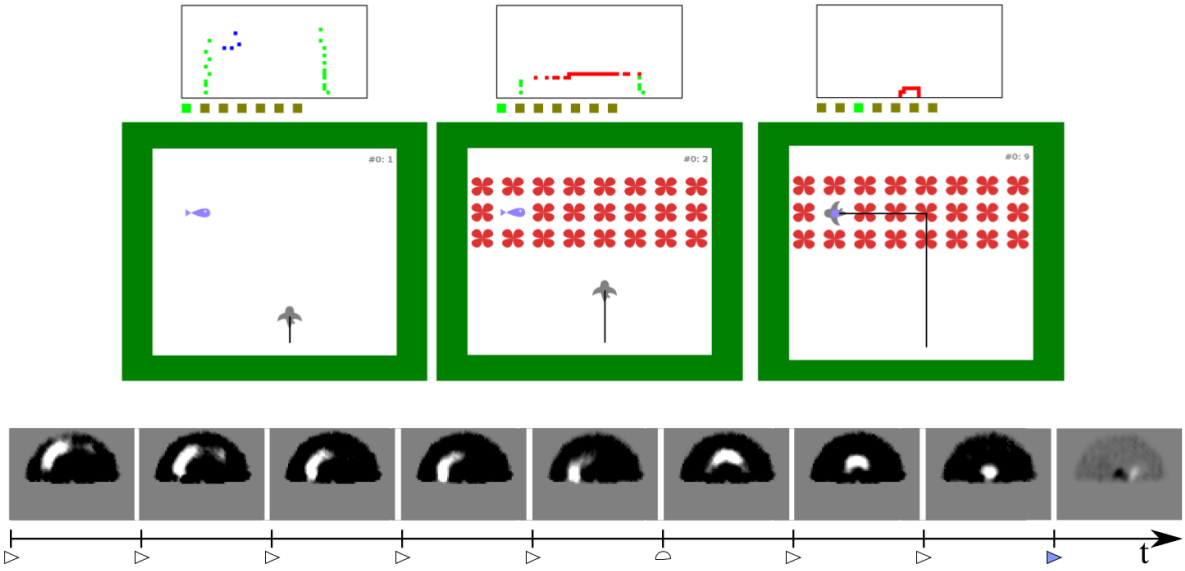


Figure 27. We let the agent discover the prey (left), then we hide the environment with algae (middle). The agent moves through algae with an efficient path (only one rotation) and reaches the prey (right). Bottom: the estimated position of the prey at each enaction cycle. The estimated position is obtained by adding place signatures of places that characterize the position of the object. The white blobs show the positions in which the object instance is likely to be. We can observe, on the last enaction cycle, that the agent *believes* that another prey is present on its right side (while the position of the eaten prey is considered to be empty). This means that during signature learning, the agent experiences several times the situation where two preys are adjacent.

Illustration 17: Exemple de tentative d'explication rapprochant les traces internes d'un agent aux traces d'observations externes, pour un agent autonome à motivation intrinsèque.

Synthèse de l'état de l'art

Nous avons, dans l'introduction de ce document, replacé l'explication dans un processus continu (exploratoire), avec une co-adaptation entre l'humain et la machine et listé un ensemble de capacités que nous considérons comme importantes pour un dispositif technique « explicable ». Nous les citerons ici et nous les reprendrons pour proposer un agenda de recherche revisité dans le domaine de l'explication des IA.

Les approches symboliques

Comme nous l'avons vu dans notre focus sur ces approches, la capacité de description formelle est présente par définition. En pratique, ces descriptions sont très rarement utilisées, en dehors du cycle de conception à des fins de mise au point par exemple. La Simulation du Raisonnement est facile car les connaissances modélisées sont exploitées par un *raisonneur.* mais dans la pratique seuls les concepteurs le font pour la mise au point. L'émergence du web sémantique modifie la donne et les ontologies disponibles sur le Web (notamment DBpedia) permettent l'exploration de grandes bases de données en exploitant les relations avec une sémantique permettant aussi le raisonnement. La faible pénétration du web sémantique dans les usages massifs du web laisse penser que l'explication du fonctionnement n'est pas suffisante pour l'appropriation par tout un chacun.

Ce sont ces capacités d'individuation du dispositif technique qui permettent son appropriation par les utilisateurs dans leurs différents contextes d'explication.

Les approches numériques & Deep Learning

Une bonne partie des recherches menées actuellement s'attachent à fournir une Description Formelle et Pédagogique des dispositifs. De la même façon, la communauté de recherche est mobilisée sur la Description des Données et de l'Apprentissage, ne serait-ce que parce que les publications des travaux l'exigent à des fins de comparaison et de reproductibilité.

La Simulation du modèle, bien qu'applicable à ces approches car lorsque le modèle F est obtenu, il « suffit » de fournir une entrée X et d'appliquer le modèle pour observer F(X), par calcul direct, ne semble pas souvent proposée. Certaines approches de vérification utilisent une sorte de simulation en produisant automatiquement, un quelconque indicateur sur les données d'entré, par exemple en bruitant une des variables sur chaque instance d'apprentissage et en observant la sortie ; Il est par contre très rarement prévu que l'utilisateur final, dans le cadre des sa tâche, puisse explorer par lui-même le modèle alors qu'il fait partie du processus d'explication en tant que demandeur de l'explication. Quelques travaux econnaissent l'importance de cette propriété et proposent une interface d'explication entre le dispositif intégrant le modèle et l'humain : cette interface peut permettre à l'humain d'explorer le modèle en fournissant ses propres données et en observant la réponse.

Si ces dispositifs sont parfois étudiés dans les laboratoires, force est de constater qu'ils ne sont pas déployés avec les applications dans la société. La plupart des capacités explicatives ne sont plus présentes. La cause invoquée est celle du secret industriel et commercial : les entreprises font payer leur service et il leur semble donc dangereux de fournir un Description Formelle ou une Description de l'Apprentissage qui pourrait permettre à des concurrents de répliquer leur modèle et/ou de l'améliorer. La Description des Données est toutefois présente, pour des raisons réglementaires (RGPD par exemple), mais se limite le plus souvent à une liste des données utilisées comme instance d'entrée du modèle.

Le passage au symbolique n'étant pas encore garanti, les capacités des dispositifs à faciliter son appropriation par une individuation du mécanisme d'explication ne sont pas abordées dans la communauté : quelques travaux proposent de distinguer des types d'audience spécifique mais ne proposent pas d'adapter le processus d'explication à des types d'audience différents. Cette limitation est reconnue comme centrale par des travaux récents, et nous proposons de la pousser dans l'agenda actuel de ce champ de recherche.

Les approches émergentistes

Les approches émergentistes, à l'exception de rares travaux déjà anciens [43], [44], n'a apparemment pas la question de l'explication à son agenda de recherche. Le mot-clé *explication* n'apparaît pas dans les conférences internationales comme nationales de la communauté SMA par exemple qui pourtant est toujours très active pour le développement de systèmes d'optimisation, de systèmes distribués de résolution de problème, etc. Dans les tutoriaux sur l'ingénierie de ces systèmes, la question de l'explication n'apparaît pas. Certains

pensent même que par définition, un modèle émergent ne s'explique que par sa propre « histoire » et donc ne peut donner lieu à explication [42].

Le faible impact dans la société des solutions d'IA basées sur des approches émergentistes peut s'expliquer en partie par cette faible capacité à installer la connivence par une mise en intelligence du dispositif numérique et de l'humain ?

Agenda de recherche ? Vers des DTN-IA fonctionnant *en intelligence* avec l'utilisateur.

L'intelligence propre des dispositifs IA développés est liée à une tâche qui est aussi celle des utilisateurs d'une manière ou d'une autre. L'autonomie est donnée pour l'exécution d'une partie de la tâche, à condition que les régulations choisies pour cette délégation soient parfaitement exécutées, indépendamment du contexte et de l'environnement d'exécution. *Exécuter une tâche humaine* en y intégrant un tel dispositif exige de travailler *en intelligence* avec lui plutôt que de considérer que l'intelligence du dispositif est autonome de l'humain. Une autonomie complète supposerait d'ailleurs que l'humain n'y soit pour rien, ce qui, pour un système artificiel, laisse perplexe.

Comme nous l'avons compris dans cette façon de présenter la question de l'explicabilité en IA, la recherche elle-même devrait s'ouvrir à d'autres disciplines que l'informatique de l'IA, et ne pas se limiter à la communauté de recherche pour s'ouvrir aux acteurs impliqués dans la mise en œuvre des *dispositifs techniques numériques de l'IA (DTN-IA)*

Revenons sur les concepts clés [voir page 14 pour le détail], présentés en synthèse de la littérature sur l'explication [18], et qu'il convient d'avoir en tête pour définir une *bonne* explication :

C1 Expliquer est un processus continu

C2 Expliquer est un processus co-adaptatif

C3 Déclencher l'explication

C4 **Faciliter l'auto-explication**

C5 **L'explication comme étant une exploration**

C6 Contraster les situations qui s'expliquent différemment

À partir de la littérature et sur la base des concepts énoncés ci-dessus, nous avons proposé une grille de *bonnes propriétés* sous la forme de *capacités* du processus d'explication.

Nous allons les reprendre ici et indiquer à quel point ces capacités sont déjà opérationnelles, à l'étude ou à mettre dans l'agenda de la recherche.

Capacité à :	**État de la recherche** •  recherche active avec des résultats publiés régulièrement dans des sessions ou ateliers nombreux. •  recherche en développement avec de premiers résultats. A développer dans l'agenda. •  recherche marginale et à mettre dans l'agenda
[DF] Décrire formellement : Une description technique et formelle du modèle à l'œuvre pour proposer une décision, une recommandation, un diagnostic, ….	 La capacité à la description formelle est par nature disponible pour les dispositifs s'appuyant sur des approches symboliques et, dans une moindre mesure, pour les dispositifs numériques capables de réaliser une ingénierie inverse, globale ou locale pour générer une sorte d'arbre de décision exploitant des variables symboliques dans le registre de compréhension de l'utilisateur. Ces recherches sont essentiellement mono-disciplinaires. Capacité de base.
[DFS] Il s'agit d'une variante plus exigeante de [DF]. C'est le cas des modèles *du monde* construits à partir de données pour en établir des *lois*	 Il existe un débat sur la nature des modèles du monde *découverts* à partir de données massives à l'aide de méthodes de l'apprentissage numérique. Le passage du modèle numérique à sa formulation en équation symbolique avec des variables identifiées dans le monde observé est la question principale. La connexion avec les chercheurs des disciplines scientifiques concernées est obligatoire par définition. Une recherche de niche sur la question est active. Il est tout à fait possible qu'un effet de bord intéressant de ces recherches soit de fournir autre chose qu'un arbre de décision pour expliquer les modèles appris pour réaliser des fonctions cognitives ? Le concept C5 doit être gardé en tête. La collaboration des disciplines concernées par les découvertes de loi est naturellement obligatoire.
[DAD] Décrire l'apprentissage et les données d'apprentissage : Il s'agit de décrire le dispositif d'apprentissage et les données mobilisées pour alimenter ce dispositif.	 Cette capacité s'applique aux méthodes numériques, mais aussi aux méthodes symboliques car elles mobilisent un processus d'acquisition de connaissances qui peut être numérique ou manuel. Ces recherches sont toujours en cours, sans qu'il y ait consensus sur la manière de procéder. Cette capacité peut s'exprimer avec les données de l'utilisateur. La connexion avec les différents types d'acteur est obligatoire et la collaboration avec les designers, les infographes, … est importante.

[SM] Simuler le comportement du modèle. Le simulateur peut être décliné pour différents types d'utilisateurs, mais doit permettre à l'utilisateur de « jouer » avec le modèle avec des données proposées par l'utilisateur.	 La simulation permet aux utilisateurs de vérifier s'ils seront *surpris* ou non par le comportement du dispositif dans leur situation d'usage. D'un certain point de vue, la capacité à simuler permet de faciliter le déclenchement (Concept d'un processus d'explication alors que la tâche de l'utilisateur n'est pas encore affectée par la prise en compte des résultats du dispositif. Certaines communautés partagent leurs simulateurs comme c'est le cas de SimGrid pour les systèmes distribués[14]. La plupart des systèmes symboliques peuvent être simulés par des requêtes sur la base de connaissances. Pour les systèmes numériques qui procurent un arbre de décision équivalent à leur modèle local ou global, c'est également possible puisqu'il y a un passage au symbolique.
[DP] Décrire Pédagogiquement : L'explication et sa documentation dépendent de ce qu'attend l'utilisateur. D'une certaine façon, c'est ce qui est pris en compte par la notion d'Audience, mais en généralisant à la capacité à s'adapter au registre de compréhension de l'individu.	 Une connexion avec les recherches autour de l'apprentissage humain et les sciences cognitives est nécessaire. A notre connaissance, cette capacité n'est pas encore étudiée pour elle-même. Cette capacité est en lien direct avec les concepts C4, C5 et C6.
[OE] Opérationnaliser les éléments d'Explication : Il s'agit ici de rendre visible, documentée et commentée la chaîne explicative pour faciliter l'apprentissage et la discussion des régulations. L'opérationnalisation consiste à rejouer les motifs explicatifs dans des conditions similaires, avec adaptation, ou normalisées, avec module pédagogique. Le dispositif possède alors des capacités d'individuation issues des interactions. La mise *en intelligence* est à ce prix sans doute.	 L'objectif est de *tracer* le processus d'explication à des fins d'apprentissage de façon à l'expliquer dans différentes situations et en interaction avec les utilisateurs concernés. Ces travaux ont une relation directe avec la volonté d'appropriation des techniques d'IA par la société. Ils peuvent devenir des supports pour la capacité [DP]. Bien qu'en connexion avec tous les concepts énoncés, c'est sans doute les concepts C4, C5 et C6 qui justifient le plus cette capacité. La communauté des Interactions Homme Machine, les sciences cognitives sont des partenaires tout désignés pour étudier cette capacité. Un certain nombre de travaux sur les assistants à l'usage des dispositifs techniques numériques sont mobilisables sur cette question. Les travaux faisant référence à l'individuation des dispositifs techniques numériques intégrant humains et machines seront également mobilisés : robotique et philosophie des techniques en particulier où de premiers travaux sont menés depuis quelques années.

14 *SimGrid https://simgrid.org/*

Références

[1] « Explainable Artificial Intelligence (XAI) », DARPA, Broad Agency Announcement DARPA-BAA-16-53, août 2016. Consulté le: juin 26, 2020. [En ligne]. Disponible sur: https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf.

[2] E. Bird, J. Fox-Skelly, N. Jenner, R. Larbey, E. Weitkamp, et A. Winfield, « The ethics of artificial intelligence: Issues and initiatives », Service de Recherche Parlementaire Européen (EPRS), Scientific Foresight Unit (STOA) PE 634.452, mars 2020. Consulté le: juin 26, 2020. [En ligne]. Disponible sur: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf.

[3] S. Prévost et E. Royer, *Intelligence Artificielle*, Dalloz. 2019.

[4] Mihaela van der Schaar, *Machine learning: from black boxes to white boxes*. .

[5] I. Saif et B. Ammanath, « 'Trustworthy AI' is a framework to help manage unique risk », *MIT Technology Review*, mars 25, 2020. https://www.technologyreview.com/2020/03/25/950291/trustworthy-ai-is-a-framework-to-help-manage-unique-risk/.

[6] G. R. Mayes, « Theories of Explanation ». https://www.iep.utm.edu/explanat/ (consulté le juin 04, 2020).

[7] C. G. Hempel et P. Oppenheim, « Studies in the Logic of Explanation », *Philos. Sci.*, vol. 15, nº 3, p. 135-175, 1948.

[8] P. Achinstein, *The nature of explanation*. Oxford University Press, 1985.

[9] J. H. Holland, K. J. Holyoak, R. E. Nisbett, P. R. Thagard, et S. W. Smoliar, « Induction: Processes of Inference, Learning, and Discovery », *IEEE Expert*, vol. 2, nº 3, p. 92-93, sept. 1987, doi: 10.1109/MEX.1987.4307100.

[10] Z. Horne, M. Muradoglu, et A. Cimpian, « Explanation as a cognitive process », *Trends Cogn. Sci.*, vol. 23, nº 3, p. 187-199, 2019.

[11] R. R. Hoffman et G. Klein, « Explaining Explanation, Part 1: Theoretical Foundations », *IEEE Intell. Syst.*, vol. 32, nº 3, p. 68-73, mai 2017, doi: 10.1109/MIS.2017.54.

[12] C. S. pIERCE, « La logique de la science. Première partie : comment se fixe la croyance (The Fixation of Belief, 1877) », *Rev. Philos. Fr. Létranger*, vol. Tome VI, p. 553-569, 1878.

[13] R. R. Hoffman, S. T. Mueller, et G. Klein, « Explaining Explanation, Part 2: Empirical Foundations », *IEEE Intell. Syst.*, vol. 32, nº 4, p. 78-86, 2017, doi: 10.1109/MIS.2017.3121544.

[14] G. Klein, « Explaining Explanation, Part 3: The Causal Landscape », *IEEE Intell. Syst.*, vol. 33, nº 2, p. 83-88, mars 2018, doi: 10.1109/MIS.2018.022441353.

[15] R. Hoffman, T. Miller, S. T. Mueller, G. Klein, et W. J. Clancey, « Explaining Explanation, Part 4: A Deep Dive on Deep Nets », *IEEE Intell. Syst.*, vol. 33, nº 3, p. 87-95, mai 2018, doi: 10.1109/MIS.2018.033001421.

[16] A. Nguyen, J. Yosinski, et J. Clune, « Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images », *ArXiv14121897 Cs*, avr. 2015, Consulté le: avr. 30, 2020. [En ligne]. Disponible sur: http://arxiv.org/abs/1412.1897.

[17] Robert R. Hoffman, G. Klein, et S. T. Mueller, « Explaining Explanation For "Explainable Ai" », *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 62, nº 1, p. 197-201, sept. 2018, doi: 10.1177/1541931218621047.

[18] S. T. Mueller, R. R. Hoffman, W. Clancey, et A. Emrey, « Explanation in Human-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI », p. 204, 2019.

[19] J. B. Lyons, M. A. Clark, A. R. Wagner, et M. J. Schuelke, « Certifiable Trust in Autonomous Systems: Making the Intractable Tangible », *AI Mag.*, vol. 38, nº 3, p. 37-49, oct. 2017, doi: 10.1609/aimag.v38i3.2717.

[20] P. Brezillon et J.-C. Pomerol, « Joint cognitive systems, cooperative systems and decision support systems: a cooperation in contexte », in *Proceedings of the European Conference on Cognitive Science*, Manchester, 1997, p. 129-139.

[21] S. T. Mueller et G. Klein, « Improving User's Mental Models of Intelligent Software Tools », *IEEE Intell. Syst.*, vol. 26, nº 2, p. 77-83, 2011, doi: 10.1109/MIS.2011.32.
[22] T. Miller, « Explanation in Artificial Intelligence: Insights from the Social Sciences », *ArXiv170607269 Cs*, août 2018, Consulté le: avr. 02, 2020. [En ligne]. Disponible sur: http://arxiv.org/abs/1706.07269.
[23] T. Miller, P. Howe, et L. Sonenberg, « Explainable AI: Beware of Inmates Running the Asylum », présenté à Workshop on Explainable Artificial Intelligence, 2017.
[24] T. G. Gill, « Early Expert Systems: Where Are They Now? », *MIS Q.*, vol. 19, nº 1, p. 51, mars 1995, doi: 10.2307/249711.
[25] W. Swartout et J. D. Moore, « Explainable (and Maintainable) Expert Systems », in *Proceedings of the 9th International Joint Conference on Artificial Intelligence*, Los Angeles, 1985, vol. 1.
[26] A. Newell, « The Knowledge Level », *Artif. Intell.*, nº 18, p. 87-127, 1982.
[27] C. Biemann, « Ontology Learning from Text: A Survey of Methods », *LDV-Forum*, vol. 20, nº 2, p. 75-93, 2005.
[28] « 6 areas where artificial neural networks outperform humans », *VentureBeat*, déc. 09, 2017. https://venturebeat.com/2017/12/08/6-areas-where-artificial-neural-networks-outperform-humans/ (consulté le juin 09, 2020).
[29] X. Zhang *et al.*, « AlignedReID: Surpassing Human-Level Performance in Person Re-Identification », *ArXiv171108184 Cs*, janv. 2018, Consulté le: juin 09, 2020. [En ligne]. Disponible sur: http://arxiv.org/abs/1711.08184.
[30] R. Wexler, « Opinion | When a Computer Program Keeps You in Jail », *The New York Times*, juin 13, 2017.
[31] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, et D. Pedreschi, « A Survey of Methods for Explaining Black Box Models », *ACM Comput. Surv.*, vol. 51, nº 5, p. 93:1–93:42, août 2018, doi: 10.1145/3236009.
[32] M. T. Ribeiro, S. Singh, et C. Guestrin, « "Why Should I Trust You?": Explaining the Predictions of Any Classifier », in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA, août 2016, p. 1135-1144, doi: 10.1145/2939672.2939778.
[33] A. Barredo Arrieta *et al.*, « Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI », *Inf. Fusion*, vol. 58, p. 82-115, juin 2020, doi: 10.1016/j.inffus.2019.12.012.
[34] M. Du, N. Liu, et X. Hu, « Techniques for interpretable machine learning », *Commun. ACM*, vol. 63, nº 1, p. 68–77, déc. 2019, doi: 10.1145/3359786.
[35] *Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données) (Texte présentant de l'intérêt pour l'EEE)*, vol. 119. 2016.
[36] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, et J. Dean, « Distributed Representations of Words and Phrases and their Compositionality », in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, et K. Q. Weinberger, Éd. Curran Associates, Inc., 2013, p. 3111–3119.
[37] C. Rudin, « Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead », *Nat. Mach. Intell.*, vol. 1, nº 5, p. 206-215, mai 2019, doi: 10.1038/s42256-019-0048-x.
[38] S. R. Islam, W. Eberle, et S. K. Ghafoor, « Towards Quantification of Explainability in Explainable Artificial Intelligence Methods », *ArXiv191110104 Cs Q-Fin*, nov. 2019, Consulté le: avr. 27, 2020. [En ligne]. Disponible sur: http://arxiv.org/abs/1911.10104.
[39] C. T. Wolf, « Explainability scenarios: towards scenario-based XAI design », in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, Marina del Ray, California, mars 2019, p. 252–257, doi: 10.1145/3301275.3302317.
[40] M. S. H. Aung *et al.*, « Comparing Analytical Decision Support Models Through Boolean Rule Extraction: A Case Study of Ovarian Tumour Malignancy », in *Advances in Neural*

Networks – ISNN 2007, vol. 4492, D. Liu, S. Fei, Z. Hou, H. Zhang, et C. Sun, Éd. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, p. 1177-1186.
[41] V. Arya *et al.*, « One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques », *ArXiv190903012 Cs Stat*, sept. 2019, Consulté le: oct. 19, 2020. [En ligne]. Disponible sur: http://arxiv.org/abs/1909.03012.
[42] M. R. Jean, « Emergence et SMA », Groupe de travail "Collectif IAD/SMA de AFCET/AFIA), 1997.
[43] J. Kim, « Explanation, Prediction, and Reduction in Emergentism », *Intellectica Rev. Assoc. Pour Rech. Cogn.*, vol. 25, nº 2, p. 45-57, 1997, doi: 10.3406/intel.1997.1556.
[44] G. Van de Vijer, « Emergence et explication », *Intellectica Rev. Assoc. Pour Rech. Cogn.*, vol. 25, nº 2, p. 7-23, 1997, doi: 10.3406/intel.1997.1554.
[45] S. Gay, A. Mille, O. L. Georgeon, et A. Dutech, « Autonomous construction and exploitation of a spatial memory by a self-motivated agent », *Cogn. Syst. Res.*, vol. 41, p. 1-35, 2017, doi: https://doi.org/10.1016/j.cogsys.2016.07.004.
[46] D. S. Weld et G. Bansal, « The challenge of crafting intelligible intelligence », *Commun. ACM*, vol. 62, nº 6, p. 70-79, 2019.