Vers l'XAI sémantique – la troisième vague de l'intelligence artificielle explicable

Toward semantic XAI – the third wave of explainable artificial intelligence

Mathias Bollaert[1], Gilles Coppin[2]

[1] Equipe INUIT, Laboratoire Lab-STICC, CNRS UMR 6285 - IMT Atlantique, Thales DMS, France, mathias.bollaert@imt-atlantique.fr
[2] Equipe INUIT, Laboratoire Lab-STICC, IMT Atlantique, France, gilles.coppin@imt-atlantique.fr

RÉSUMÉ. Pour répondre aux problèmes posés par l'utilisation croissante des modèles IA dans les applications à forts enjeux socio-économiques ou de sécurité, l'intelligence artificielle explicable (XAI) a connu un essor important durant les dernières années. Initialement dévolue à la recherche de solutions techniques permettant de produire automatiquement des explications, elle s'est heurtée à plusieurs difficultés, en particulier lorsque ces solutions ont été confrontées à des utilisateurs finaux non experts. L'XAI s'est alors attachée à s'inspirer des sciences sociales pour produire des explications plus faciles à comprendre. Malgré certains résultats encourageants, cette nouvelle approche n'a pas apporté autant qu'espéré. Cet article analyse l'évolution de l'XAI à travers ces deux périodes. Il évoque des raisons possibles des difficultés rencontrées, puis propose une nouvelle approche pour améliorer la production automatisée d'explications. Cette approche, nommée explicabilité sémantique ou S-XAI, est centrée sur la cognition de l'utilisateur. Alors que les méthodes précédentes sont orientées sur les algorithmes ou sur la causalité, la S-XAI part du principe que la compréhension repose avant tout sur la capacité de ce dernier à s'approprier le sens de ce qui est expliqué.
ABSTRACT. To respond to the problems posed by the growing use of AI models in high stakes applications, explainable artificial intelligence (XAI) has experienced significant growth in recent years. Initially dedicated to the search for technical solutions making it possible to produce explanations automatically, it encountered several difficulties, in particular when these solutions were confronted with non-expert end users. The XAI then sought to draw inspiration from the social sciences to produce explanations that were easier to understand. Despite some encouraging results, this new approach has not brought as much as hoped. This article analyzes the evolution of the XAI through these two periods. He discusses possible reasons for the difficulties encountered, and then proposes a new approach to improve the automated production of explanations. This approach, called semantic explainability or S-XAI, focuses on user cognition. While previous methods are oriented towards algorithms or causality, S-XAI starts from the principle that understanding relies above all on the user's ability to appropriate the meaning of what is explained.
MOTS-CLÉS. Causalité, explicabilité, IA, intelligence artificielle, interaction, sémantique, sciences humaines et sociales, utilisateurs, XAI.
KEYWORDS. AI, artificial intelligence, end-users, causality, explainability, human and social sciences, interaction, semantics, XAI.

De nombreux auteurs ont tenté d'apporter une définition formelle partagée des concepts associés à l'XAI, notamment ceux d'explicabilité et d'interprétabilité. Pourtant aucune définition consensuelle de ces termes n'est adoptée par la communauté scientifique. Nombre d'articles revendiquent la création de modèles explicables ou interprétables [LIP 16] alors même que personne ne s'accorde sur le sens de ces mots [DOS 17]. On définira l'IA explicable en partant du sens du mot explication. S'il n'en existe pas non plus de définition consensuelle [ARR 20], il présente l'avantage d'appartenir au langage courant, et tout le monde sait, de façon plus ou moins intuitive, ce qu'il signifie. Lewis [LEW 86] propose une définition simple et suffisamment générale :

DEFINITION. – *Une explication sur un évènement consiste à fournir des informations (nommées informations explicatives) à propos de l'histoire causale de cet évènement.*

D'autres auteurs ne s'accordent pas sur ce qu'est l'objet d'une explication. Pour certains il s'agit d'évènements [SKO 14], pour d'autres de faits [BEN 88], ou encore de variables [WOO 03]. Ici, le mot « évènement » n'est utilisé que comme support général de la discussion ; les considérations sur la nature précise du sujet de l'explication n'ont pas d'incidence sur les propos de cet article. On

considérera également, dans le cadre de cet article, que les explications sont toujours de nature causale[1].

A partir de cette définition, il est possible de recenser les objectifs et les enjeux de l'XAI tout en esquivant la difficile tâche de définir l'explicabilité de façon consensuelle.

1. Les explications dans le contexte de l'XAI

Malgré une grande diversité, on peut dresser un cadre commun à toutes formes d'explications. L'explication en elle-même (en anglais *explanans*, en français *explicans*) met en jeu quatre éléments : la personne qui a besoin d'une explication (*explainee* [MIL 19] qui sera ici remplacé par *apprenant*, faute d'équivalent en français) et les motivations qui la poussent à la demander ; la personne ou le système, qui fournit l'explication (*explainer*, en français *explicateur*) et les motivations qu'il ou elle a pour le faire ; le phénomène ou évènement nécessitant une explication ou la description de ce phénomène (*explanandum*, en français *explicandum*) [HEM 48] [OVE 11] ; le contexte relatif à cette demande d'explication. *Une explication répond à une question* de l'apprenant[2], qui sera différente selon ses motivations à en faire la demande. L'explication adaptée au quadruplet, qui dépend de la question posée [BRO 92] sera nommée, dans cet article, la « meilleure explication ». *L'enjeu est de déterminer comment trouver cette « meilleure explication » parmi l'ensemble des explications possibles*.

L'apprenant n'est pas nécessairement un scientifique ; le discours explicatif doit donc aller au-delà des explications scientifiques usuellement rattachées à des modèles de déduction logique [HEM 48 *op.cit.*]. L'*attribution causale*, consistant à extraire une chaine causale et à la présenter à l'apprenant, ne vaut pas *explication causale* [MIL 19]. D'ailleurs, le principe même de l'explication en XAI ne se prête pas à une démonstration logique, car cette dernière repose sur un raisonnement déductif alors que l'explicabilité repose sur le raisonnement abductif[3] [GIL 18].

Les explications peuvent par exemple concerner le fonctionnement du modèle, la façon dont les données utilisées pour son apprentissage sont obtenues et traitées, ou les techniques utilisées pour éviter les biais de sélection. Cet article s'intéresse à la production d'explications destinées aux utilisateurs et répondant à la question suivante : « pourquoi l'IA a-t-elle fourni cette sortie en réponse à cette entrée ? », ainsi qu'à toutes les questions afférentes (par exemple « plutôt que telle autre sortie ? »).

Dans cet article, on appellera X-XAI (*eXplanations-based eXplainable Artificial Intelligence*) le sous-domaine de l'XAI visant à expliquer à l'utilisateur[4] pour quelle(s) raison(s) le modèle IA produit une sortie donnée en réponse à une entrée donnée.

---

[1] La nature causale de certaines catégories d'explications est débattue, mais ces dernières ne s'appliquent pas dans le cadre de l'XAI. Il s'agit par exemple d'explications sur la nature de la réalité, sur des axiomes mathématiques [SKO 14 *op.cit.*] ou sur des propriétés des objets de la nature.

[2] On notera qu'il est possible de fournir une explication sans qu'une question ait été posée au préalable, par exemple dans un ouvrage. Cette explication répondra néanmoins à une question supposée, ce qui n'invalide pas le principe d'explication en tant que réponse à une question.

[3] Le raisonnement abductif consiste à inférer une cause probable à partir de faits observés. C'est le type de raisonnement utilisé par un médecin lors d'un diagnostic (tels symptômes sont probablement causés par telle maladie). De la même façon, l'explicabilité consiste à expliquer un effet (l'image est identifiée comme un chat) à partir des faits observés (l'animal a des oreilles pointues et des pupilles allongées).

[4] Le terme *utilisateur* sera défini plus précisément dans la suite de l'article.

2. Cas d'usage : exemple de surveillance maritime

Pour illustrer les différents problèmes posés par la création de méthodes d'explicabilité, on utilisera un scénario fictif de surveillance maritime, tenant lieu de fil rouge[5]. Il est donné à titre d'exemple et son adaptation à d'autres domaines, voire sa généralisation sont évidemment possibles. Dans ce scénario, par temps calme et en fin de journée, un navire de pêche part, d'un pays hors UE, indiquant une fois en haute mer son activité de pêche dans des messages AIS[6]. Durant la nuit, il effectue différentes manœuvres à proximité des eaux italiennes, avant de retourner, le lendemain, à son port de départ. Pendant quelques heures, il coupe son transpondeur AIS, une opération ordinaire permettant aux marins de ne pas révéler aux concurrents leurs zones de pêche. Quelques heures plus tard, un petit navire de plaisance accoste dans un port d'une île Italienne d'où il était parti la veille.

La situation est analysée par une IA, chargée de détecter tout comportement suspect et de prédire le type de risque associé. Elle est alimentée par des données issues d'imagerie radar, de photographie aérienne et satellitaire, d'informations en provenance des services de renseignement, de messages de navigation civile, et de différentes bases de données utiles. Une interface permet à l'utilisateur de dialoguer pour comprendre les raisons de tout déclenchement d'alerte. Dans le cas présent, une alerte est déclenchée, avec prédiction de risque de trafic de clandestins, pour les raisons suivantes :

– Le capitaine du bateau de pêche a une carrière suspecte ; il a notamment déjà été condamné pour diverses infractions à la législation maritime.

– Les services de renseignement ont indiqué que ce bateau est parti d'une région surveillée ; la zone portuaire de départ n'est pas actuellement associée au trafic de clandestins, mais localisée à proximité d'un camp de réfugiés en attente d'un titre de séjour après demande d'asile. Cette zone est sous surveillance suite à des renseignements, selon lesquels une organisation criminelle s'y est installée récemment pour organiser des voyages d'immigration clandestine.

– La navigation nocturne de certains navires de pêche est une opportunité pour les passeurs, qui peuvent ainsi plus facilement échapper à la surveillance.

– L'AIS a été coupé à un endroit ne correspondant pas à une zone de pêche connue et le navire a été repéré en dehors des zones de pêche usuelles, grâce aux images satellites radar[7] (fournies par exemple par les satellites *Sentinel* du programme *Copernicus* de l'Union Européenne), avec un bon niveau de confiance ; cette trajectoire suspecte pourrait viser à contourner les radars côtiers (sémaphores, …), afin d'éviter d'être repéré.

– Le navire de plaisance parti d'Italie semble avoir falsifié ses propres messages AIS, car l'imagerie satellite ne révèle aucun navire sur la trajectoire indiquée ; la trajectoire est suspecte car elle semble artificielle ; la sortie de ce navire s'est essentiellement déroulée la nuit, ce qui est peu cohérent avec la navigation de plaisance.

– Une photographie aérienne au lever du jour est classifiée par l'IA comme un possible accostage en mer entre deux navires ; ces navires pourraient correspondre au navire de pêche et au navire de plaisance, avec un niveau de confiance moyen.

Dans cet exemple, la question initiale et centrale que l'opérateur se pose est : « pourquoi l'IA a identifié le comportement de ces navires comme révélateur d'un potentiel trafic de clandestins ? ». Cet

[5] La plausibilité de ce scénario n'a pas été validée par des experts de surveillance maritime. Il est uniquement illustratif, sans prétendre à un total réalisme.

[6] *Automatic Identification System* (AIS) : système d'identification automatique (SIA) des navires par échange automatisé de messages entre les navires et les organismes ou systèmes de surveillance de trafic d'une zone de navigation, affichant l'identité, le statut, la position et la route des navires.

[7] L'imagerie radar à synthèse d'ouverture (*Synthetic Aperture Radar, SAR* en anglais) peut fonctionner de nuit.

exemple et les réponses apportées à l'utilisateur par le système d'IA, serviront à illustrer les différentes problématiques de l'XAI.

3. Les origines de l'XAI : pourquoi les humains ont-ils besoin d'explications ?

En général, le besoin d'explication apparaît lorsque le système devient trop complexe pour que l'utilisateur puisse avoir une compréhension suffisante de son fonctionnement. Si la complexité est souvent une condition nécessaire à la demande d'explication, elle n'en est pas une condition suffisante. On utilise quotidiennement des objets dont on ne maîtrise pas le fonctionnement, comme par exemple une automobile, sans éprouver le besoin que ce dernier soit expliqué.

Pour comprendre les raisons pour lesquelles l'explicabilité en intelligence artificielle est devenue un enjeu aussi crucial, il est nécessaire d'expliquer le fonctionnement des techniques d'IA telles que l'apprentissage profond. Etant à la fois les modèles fournissant souvent la meilleure performance [HER 22] mais aussi, simultanément, les plus difficiles à justifier et à comprendre, ils constituent la cible privilégiée des méthodes d'XAI.

3.1. *Pourquoi les modèles AI nécessitent-ils des explications ?*

L'essor de l'intelligence artificielle de ces dernières années est essentiellement dû à l'accroissement ininterrompu de la puissance des ordinateurs, ayant permis de développer des algorithmes très performants, au point souvent d'égaler les capacités des experts [TIN 17] [LIU 20], voire de dépasser les capacités humaines [HE 15]. La combinaison de l'utilisation de très gros volumes de données et de l'apprentissage profond rend cependant leur fonctionnement opaque et assimilable à des boites noires.

Ce fonctionnement, basé sur des corrélations statistiques, entraîne l'apparition potentielle de biais et une mauvaise capacité de généralisation [MAR 21]. Ceci engendre un risque important à utiliser ces systèmes [NIK 21] malgré leurs performances, et justifie de développer des techniques et outils permettant de comprendre et d'expliquer les décisions ou recommandations. C'est le cas des systèmes critiques [SAB 23], en raison des enjeux de sûreté et de sécurité [MAR 21 *op.cit.*], de la cybersécurité [HAR 21], mais aussi de tous les domaines impliquant les citoyens, et en particulier la santé [FEL 19], le domaine juridique [HEL 19] et la finance [ARR 20 *op.cit.*]. Les systèmes doivent être capables d'expliquer leurs résultats ; par exemple pour l'allocation d'un crédit immobilier, pour les assurances [ONE 16] ou pour la prévention des risques de récidive criminelle [ANG 22]. En l'absence d'explications satisfaisantes pour l'usager, les systèmes sont souvent limités à des algorithmes plus simples à interpréter mais moins performants [LUN 18]. Les concepteurs sont confrontés à la recherche d'un compromis entre la précision de l'algorithme et la capacité du système à être compris.

3.2. *Des explications : dans quel but ?*

Plusieurs auteurs ont recensé les motivations et bénéfices de la recherche d'explications [ADA 18] [LIP 16 *op.cit.*] [DOS 17 *op.cit.*]. Celles-ci servent souvent à justifier des décisions, à améliorer la fiabilité, la confidentialité, l'équité, ou la sureté de fonctionnement des modèles. L'explicabilité est utile pour améliorer leur performance ou leur robustesse, en aidant à détecter les biais, ou à déterminer le domaine d'applicabilité. Elle est aussi une piste privilégiée pour la réalisation de systèmes IA de confiance, fiables et sécurisés [MAR 21 *op.cit.*]. L'explicabilité aide également l'utilisateur à ajuster son modèle mental du fonctionnement du modèle [KUL 13] [ZHA 20] [HOF 18], ce qui améliore son acceptabilité et donc la performance globale du système humain-machine.

Les raisons de fournir des explications ainsi que le type d'explications attendues par l'utilisateur dépendent de celui-là [ARR 20 *op.cit.*]. Ainsi, les *data scientists* souhaitent améliorer l'efficience, la performance du modèle ou étendre son champ d'application ; les managers ou les utilisateurs des agences de régulation sont plus intéressés par la conformité réglementaire au niveau de l'entreprise ou de la législation ; les *utilisateurs de premier niveau*, directement en contact avec l'application, cherchent à améliorer leurs compétences ou à vérifier s'ils peuvent faire confiance au modèle ; les

utilisateurs de second niveau tels que les patients ou les citoyens, dont les données sont traitées par le modèle, souhaitent avoir des garanties sur son efficacité ainsi que son éthique ou son équité. Chacune de ces catégories sera donc intéressée par des explications spécifiques, parfois différentes.

3.3. *Les deux premières vagues de l'XAI*

Le sigle XAI a été utilisé pour la première fois en 2004 [VAN 04]. Initialement, et alors que la technologie était en plein développement, la recherche en XAI visait prioritairement à produire des explications à destination des équipes techniques. L'objectif était d'améliorer la précision et la performance des modèles, d'aider à détecter leurs biais ou, plus généralement, d'améliorer les connaissances générales en intelligence artificielle. Cette période correspond à *la première vague de l'XAI*.

Une seconde période est apparue progressivement lorsque des utilisateurs de premier et de second niveau[8] ont été confrontés à des modèles d'IA, rendant nécessaire d'expliquer de façon simple pour quelle raison l'IA produit un résultat donné pour une entrée donnée (explication du traitement [GIL 18 *op.cit.*]). Les chercheurs de la communauté XAI se sont progressivement tournés vers les sciences sociales dans le but de produire des explications plus adaptées à la cognition humaine. On appellera cette période *la seconde vague de l'XAI*. Bien que cette séparation en deux vagues soit arbitraire et que la première ne se soit en réalité jamais arrêtée, on peut considérer que la seconde vague a démarré aux alentours de l'année 2018.

4. L'essor et la première vague de l'XAI

L'XAI a gagné une attention considérable ces dernières années [ADA 18 *op.cit.*]. Un examen du vocabulaire de l'XAI révèle pourtant une absence de consensus à propos de la plupart des termes utilisés. L'interprétabilité est parfois considérée comme synonyme d'explicabilité [LOU 13] [CAB 19], et d'autres fois rapprochée de la notion de transparence[9] [RUD 19] [GUI 18]. Compréhensibilité et intelligibilité [LIP 16 *op.cit.*] ne possèdent pas de définition consensuelle. Plus généralement, l'XAI souffre d'un manque de formalisation [ARR 20 *op.cit.*]. Pour certains, une explication est une interface [GUI 18 *op.cit.*] [GUN 19]. Pour d'autres [MAR 21 *op.cit.*], l'IA est explicable si son modèle est directement compréhensible ou s'il est accompagné par une explication, alors que le terme explication n'est pas lui-même formellement défini. Pour Montavon et al. [MON 18], une explication est l'ensemble de traits ou caractéristiques ayant contribué à fournir le résultat produit. Pour Gilpin et al. [GIL 18 *op.cit.*], l'explicabilité est caractérisée par l'interprétabilité et la complétude ; pour Markus et al. [MAR 21 *op.cit.*], elle nécessite à la fois l'interprétabilité et la fidélité. Ces facteurs étant souvent mutuellement incompatibles, une explication est nécessairement un compromis, entre interprétabilité et fidélité [WEL 19], précision [ADA 18 *op.cit.*] ou performance [ARR 20 *op.cit.*], entre intelligibilité et complétude, ou entre concision et précision par exemple.

4.1. *Une définition de l'IA explicable dans le contexte restreint de l'X-XAI*

Une définition du terme IA explicable est proposée ici, *dans le seul contexte restreint de l'X-XAI* :

DEFINITION. – *Une IA explicable fournit à l'utilisateur des indices étayant l'hypothèse selon laquelle les corrélations à l'origine de la production d'un résultat à partir d'une entrée sont sous-tendues par l'existence de liens causaux.*

---

[8] Dans la suite de cet article, on désignera ces deux types d'utilisateurs comme *utilisateurs* de façon indifférenciée.

[9] Le terme *transparence* sera préféré dans cet article à celui d'*interprétabilité*, car moins ambigu.

Cette définition évite le caractère tautologique parfois rencontré dans des définitions de l'XAI, en ne faisant pas mention de la notion d'explication. Elle évite aussi le terme compréhension, pour ne pas présupposer de la capacité de chaque utilisateur à interpréter les informations fournies. La définition exclut également le fonctionnement de l'IA. Dans le scénario de surveillance maritime proposé plus haut, l'utilisateur cherche à comprendre pourquoi les navires sont suspects, et non comment l'IA parvient à cette hypothèse. Cette vision fonctionnelle est partagée par différents chercheurs [HOL 19] [MON 18 *op.cit.*]. On notera cependant que dans le cas d'algorithmes non opaques, par exemple avec un arbre de décision ou un système basé sur des règles, l'utilisateur pourrait éventuellement être intéressé aussi de savoir quelles règles ont ou n'ont pas conduit à produire le résultat.

4.2. *Une synthèse des techniques de l'X-XAI*

Bien que leur taxonomie ne soit pas exempte de défauts [SOG 22], les techniques d'XAI peuvent être catégorisées selon trois axes : *intrinsèque* vs. *post hoc*[10], *globale* vs. *locale*[11] et *Model-agnostic* vs. *Model-specific*[12]. La séparation entre méthodes intrinsèques et *post hoc* écarte une troisième catégorie, les *modèles auto-explicatifs*, dans lesquelles le mécanisme de génération d'explication est intégré dans l'architecture du modèle [ELT 20]. Cet axe écarte également les modèles IA *partiellement transparents* comme le modèle SIAM [MAR 19], qui produisent des résultats intermédiaires interprétables. Sur le deuxième axe et dans une optique de X-XAI, seules les méthodes locales sont pertinentes. Sur le troisième axe, étant donné que la notion d'explicabilité sémantique proposée dans cet article est avant tout une position épistémique, l'intérêt sera principalement orienté vers les méthodes agnostiques.

Le tableau ci-dessous, en s'inspirant de la classification d'Arrieta et al. [ARR 20 *op.cit.*], synthétise les différentes techniques de l'X-XAI appartenant à la catégorie *Outcome Explanation* [GUI 18 *op.cit.*].

Famille	Catégorie	Description	Exemple
Méthodes *post hoc* ou auto-explicatives	Méthodes d'attribution	Evaluer la contribution de chacun des traits du signal d'entrée dans le résultat obtenu	SHAP (*SHapley Additive exPlanation*) [LUN 17]
	Explication par simplification ou modification d'architecture	Transformer le modèle opaque en un modèle de la famille des méthodes intrinsèques, ou générer un tel modèle adossé au modèle opaque	LIME (*Local Interpretable Model-agnostic Explanation*) [RIB 16]
	Explication visuelle ou textuelle	Produire des représentations visuelles ou textuelles, par exemple à base de cartes de chaleur	Grad-CAM [SEL 17]
	Explication à partir d'exemples ou de contre-exemples	Extraire ou générer des exemples ou contre-exemples représentatifs	WACH [WAC 17]

[10] Un modèle intrinsèquement interprétable l'est par conception, par opposition aux méthodes post-hoc, mises en œuvre *a posteriori*.

[11] Expliquer une décision donnée (locale) vs. la totalité du fonctionnement du modèle (globale).

[12] Une méthode *model-agnostic* ne dépend pas de l'implémentation et peut être appliquée à tout type de modèle.

Méthodes intrinsèques	Régression linéaire, arbres de décision, K plus proches voisins, apprentissage à base de règles, modèles bayésiens, modèles additifs généralisés	N/A	N/A

Tableau 1. *Synthèse des techniques de l'XAI.*

Les méthodes *post hoc* et auto-explicatives sont regroupées en quatre catégories. Les *méthodes d'attribution* [SUN 17], basées sur la pertinence de traits, constituent le cœur de l'X-XAI. Les *explications par simplification ou modification d'architecture* génèrent un modèle transparent, soit pour remplacer le modèle opaque, soit pour l'y adosser afin de produire des explications. Les *méthodes d'explication visuelle ou textuelle* génèrent une image ou du texte pour expliquer le résultat produit ; par exemple, pour des tâches de classification d'image, en produisant une carte de chaleur superposable. La plupart des techniques appartiennent en réalité à plusieurs de ces catégories, même s'il reste en général possible d'identifier la principale d'entre elles. La dernière catégorie, les *explications à partir d'exemples ou de contre-exemples*, est plus clairement dissociée des précédentes.

4.3. *Limites de l'XAI*

Les explications de la première vague développées initialement par l'XAI prennent essentiellement en compte l'*explicandum*, c'est-à-dire la chose à expliquer, indépendamment de la personne ciblée, de ses motivations, du but de l'explication et du contexte. Cette première vague est conçue avant tout au bénéfice des chercheurs en IA eux-mêmes [RAS 18]. Elle est centrée sur des algorithmes, ce qui conduit Miller à expliquer que la solution à l'IA explicable n'est pas « plus d'IA » [MIL 19 *op.cit.*], ou Roscher à affirmer que l'explicabilité ne peut pas être purement algorithmique [ROS 20].

Les effets de ces explications sont contrastés voire contradictoires, parfois sans réelle efficacité [KAU 20] [RAS 18 *op.cit.*], d'autres fois avec un effet sensible sur la performance globale du système cognitif humain + machine [LUN 18 *op.cit.*] [CAI 19] [LAI 19] [YAN 20].

Plusieurs difficultés sont apparues. Ainsi, le niveau d'expertise et le temps disponible pour livrer une explication ne sont pas pris en compte [GUI 18 *op.cit.*] [DOS 17 *op.cit.*]. Ou encore le fait que les explications à base de traits (*features*), soient difficilement compatibles avec l'existence de différents niveaux d'explication [MON 18 *op.cit.*].

4.3.1. *Le « dernier pas » et le problème de la corrélation*

Une question importante est de déterminer si les outils de l'explicabilité permettent de garantir la causalité des inférences réalisées par le modèle IA. En effet, le fonctionnement de l'esprit humain est de rechercher des explications causales [TAE 18] [HOF 11] [KLE 14]. Il est souvent postulé l'existence d'un lien fondamental entre explicabilité et causalité (par exemple [CHA 97] [HOL 19 *op.cit.*] [CHO 22]). Les corrélations constituent un signal positif, mais aussi significatives soient elles[13], ne sont en aucun cas une preuve de causalité. Ce problème de l'XAI peut-être reformulé :

Problème de la corrélation : est-il possible de prouver la nature causale des inférences à base de corrélations réalisées par les modèles statistiques d'IA basés sur l'apprentissage automatique ?

[13] https://www.tylervigen.com/spurious-correlations

Selon certains auteurs [ARR 20 *op.cit.*], les techniques de l'XAI permettent de vérifier et garantir que seules les variables significatives interviennent dans le processus d'inférence, ce qui fournirait ainsi une preuve de leur nature causale. Pourtant, les méthodes d'attribution telles que LIME ou SHAP n'apportent par elles-mêmes aucune information sur la nature éventuellement causale de la corrélation. Par exemple, la plupart des méthodes d'IA explicable appliquées à des systèmes de catégorisation d'images utilisent des cartes thermiques[14] basées sur la détection ou la saillance de traits [ANC 19]. Bien que présentées en tant que techniques d'explicabilité au sein de l'XAI, ces solutions ne peuvent pas être définies comme explications au sens usuel du terme. Selon [RUD 19 *op.cit.*], « *La saillance n'explique rien, si ce n'est où le réseau regarde* ».

Démontrer la nature causale d'inférences à partir de données d'observation est une tâche cognitive difficile et fortement dépendante des connaissances antérieures de l'utilisateur [PEA 09] [NIC 07]. Ce processus n'est pas automatisable [FRE 03], ce qui compromet fortement l'espoir, pour l'XAI, de résoudre le *problème de la corrélation* (*cf. supra*). En réalité, la tâche finale consistant à déterminer la nature éventuellement causale des inférences réalisées par l'IA est laissée à l'appréciation de l'utilisateur, à qui elle est déléguée. Ce recours impératif à ses connaissances ou à sa compréhension intuitive [LIP 16 *op.cit.*] pour « aller au bout de l'explication », pour « franchir le *dernier pas* » permettant d'atteindre la compréhension, est une limite de l'XAI. Il nécessite que l'utilisateur ait les compétences suffisantes pour parvenir à cette conclusion. Le but de l'X-XAI est de développer des techniques permettant d'aider à franchir ce *dernier pas*.

Dans le scénario de surveillance maritime proposé plus haut, un opérateur pourrait chercher à comprendre pourquoi la coupure de l'AIS est suspecte. Si l'explication indique que la coupure s'est faite en dehors d'une zone de pêche et qu'elle semble correspondre à une stratégie d'évitement des radars côtiers, elle dépasse une simple observation de corrélation. Elle montre bien une relation causale entre les comportements observés et la classification du navire comme suspect. Cela contribue ainsi à réduire la longueur du *dernier pas*.

La psychologie cognitive, en étudiant le fonctionnement de l'esprit humain, peut contribuer à améliorer la qualité des explications en tenant compte de contraintes et biais cognitifs. Ainsi, la difficulté naturelle des humains à comprendre les pourcentages et les probabilités [BAR 80] devrait inciter les concepteurs à présenter les données aux utilisateurs sous forme de fréquence plutôt que de probabilités [HOF 98] [LAI 19 *op.cit.*]. Les individus préfèrent les explications basées sur les causes [MIL 19 *op.cit.*], or de nombreuses techniques d'XAI sont basées sur des algorithmes probabilistes [ARR 20 *op.cit.*]. Les problèmes que les humains rencontrent avec les probabilités ne sont qu'un exemple de particularisme cognitif. Il y en a beaucoup d'autres qui nécessiteraient d'être pris en compte pour générer la *meilleure explication*. Ces enjeux sont au cœur de la troisième vague de l'XAI.

4.3.2. *Le problème de la Pensée Alien, ou l'histoire du chat et du guacamole*

L'expression « *Pensée Alien* » illustre la différence entre les mécanismes de la cognition mis en œuvre lorsqu'un individu cherche à comprendre et expliquer un phénomène, et les processus qui se déroulent dans les modèles d'intelligence artificielle. Ces processus sont assimilés à une forme de xéno-cognition, pas nécessairement supérieure, mais foncièrement différente dans ces principes de fonctionnement. Le problème de la *Pensée*[15] *Alien* peut s'énoncer de la façon suivante :

Problème de la *Pensée Alien* : les représentations internes et les traitements réalisés par les modèles IA utilisant l'apprentissage profond sont très différents des représentations internes et la formalisation du monde par les humains.

[14] Représentations graphiques de données statistiques auxquelles on superpose une couleur selon l'intensité de la valeur.

[15] Dans ce chapitre, les modèles IA basés sur l'apprentissage profond sont anthropomorphisés pour simplifier la narration.

Dans les années 90, certains travaux pouvaient laisser penser que les réseaux de neurones formels fonctionnaient selon des caractéristiques proches de celles de l'apprentissage humain, notamment dans différents problèmes de traitement du langage [RUM 86] [PLU 99] [ELM 93] [ELM 90], et plus récemment dans des problématiques de détection de numérosité [NAS 19]. Si cette similitude pouvait sembler logique en raison de l'inspiration biomimétique des réseaux, elle a été très tôt critiquée [PIN 88] et n'est plus de mise dans les travaux récents. L'explication réside dans le relâchement de contrainte résultant de l'augmentation du nombre de couches cachées, permettant aux modèles de développer des représentations internes diversifiées sans réel rapport avec l'organisation neuronale naturelle.

L'exemple du chat et du guacamole fait référence à une photographie de chat modifiée, volontairement bruitée afin d'être identifiée à tort par une IA comme du guacamole avec un taux de confiance de 99% [ATH 18], alors qu'elle est indiscernable de la photo originale pour un humain. Ce cas illustre le *problème de la Pensée Alien* et ses conséquences. Bien que dotées de performances souvent supérieures, les IA peuvent se tromper dans des tâches qui pourtant nous semblent évidentes. Un modèle réalisé en 2022 [CHE 22], visant à prédire le comportement de phénomènes physiques, fournit un autre exemple de *Pensée Alien*. Entraîné sur un pendule double[16], ce modèle a développé des représentations internes aussi efficientes que celles de nos modèles mathématiques, permettant de prédire le mouvement de l'objet, sans que les chercheurs parviennent pourtant à trouver des corrélations avec la façon humaine de modéliser un tel objet.

Les représentations supportées par les couches cachées des réseaux de neurones sont souvent incompréhensibles pour un humain. Il ne suffit donc pas d'y placer des sondes pour pouvoir expliquer les raisonnements ; il s'agit d'une autre forme de pensée. Le *problème de la Pensée Alien* accroit à la fois le caractère indispensable et la difficulté de l'XAI.

5. La seconde vague de l'XAI : apports des sciences humaines et sociales

De nombreux auteurs ont souhaité que l'XAI s'inspire des sciences humaines et sociales (SHS) pour des explications de meilleure qualité. Ils regrettent que ce soit insuffisamment le cas [ROS 20 *op.cit.*] [MOH 21 *op.cit.*] [DOS 17 *op.cit.*]. Selon eux, pour générer artificiellement de bonnes explications, il faut modéliser la façon dont les humains les produisent. Tim Miller a beaucoup contribué à cette prise de conscience dans la communauté XAI [MIL 19 *op.cit.*]. La nécessité pour l'X-XAI de fournir des explications aux utilisateurs a montré l'importance de ces critères issus des SHS [LIA 20] [SOK 20].

Certains travaux d'XAI se sont inspirés de la psychologie ou de la philosophie [KEI 06] [COL 17] pour caractériser une *meilleure explication*. Celle-là est caractérisée par le fait qu'elle ne doit dire que des choses exactes et vérifiées, ne fournir que des informations pertinentes, être claire, concise et non ambiguë [GRI 75]. Elle doit être simple, et cohérente avec les connaissances antérieures [LOM 07 *op.cit.*]. Ces données permettent de vérifier empiriquement et *a posteriori* la qualité des explications [KAU 20 *op.cit.*]. Elles ne fournissent cependant pas d'indications ni de pistes pour la création de méthodes pour des explications de meilleure qualité. Quelques critères plus tangibles recensés par Miller sont résumés dans les paragraphes suivants.

5.1. *Explications contrastées et contrefactuelles*

Une explication contrastée (*contrastive*) répond à la question « pourquoi P s'est produit plutôt que Q ? », où Q est appelé cas de contraste (*contrast case*) [LIP 90]. Elle permet de produire une explication plus appropriée, plus simple à produire et à comprendre [MIL 21]. Dans l'exemple proposé, à la question « pourquoi une suspicion d'immigration *plutôt que de trafic de stupéfiants* ? »

[16] Un pendule à l'extrémité duquel on a attaché un second pendule.

une réponse adaptée et concise repose sur trois critères : la présence d'un camp de migrants à proximité du port de départ, l'arrivée récente de l'organisation se livrant à de l'immigration clandestine, et le fait que ce port ne soit pas connu comme passerelle de trafic de stupéfiants. Par leur concision et leur pertinence, les explications contrastées satisfont les critères qualitatifs énoncés par Grice [GRI 75]. La génération d'explications contrastées [ROB 18] constitue une avancée significative, cependant elle nécessite de mettre en place une forme de dialogue afin que l'apprenant puisse expliciter le cas de contraste. D'autre part, restant exposées au problème de la corrélation, les explications contrastées générées automatiquement peuvent produire des résultats dénués de sens [RIB 16 op.cit.].

Une explication contrefactuelle répond à la question « qu'aurait-il fallu changer pour que P ne se produise pas ? ». En déterminant ce qui dans l'information d'entrée a contribué à produire P, elle aide à différencier les corrélations des causalités. Ce type de raisonnement correspond à une tendance naturelle humaine [GER 17] [FOG 64]. Cependant, utilisé avec une approche naïve, il peut conduire à de fausses déductions [HAL 05]. Il est surtout intéressant lorsqu'il concerne des évènements récents et évitables [BYR 19] [GUI 22].

Depuis les travaux fondateurs sur les explications contractuelles [WAC 17 *op.cit.*], de nombreux algorithmes ont été développés pour en produire [GUI 22 *ibid.*]. Cependant, des exemples ne peuvent pas toujours être trouvés dans le corpus d'apprentissage, et lorsqu'ils existent, ils se heurtent parfois à la protection des données personnelles. C'est pourquoi il est utile de les générer artificiellement. Cependant, il existe un risque de produire des explications dénuées de sens [LAU 19] ou pas suffisamment pertinentes pour convaincre [HER 22 *op.cit.*] ; il faut donc veiller à leur réalisme [NAI 23 *op.cit.*]. Un autre moyen d'obtenir des exemples plus réalistes est d'en générer et en montrer plusieurs, mais au prix d'une plus grande charge cognitive, ce qui n'est pas toujours acceptable [ALB 22]. Enfin, un dernier problème posé par la génération d'exemples contrefactuels est leur coût important en ressources [GUY 22 *op.cit.*].

5.2. *Causalité et graphes de causalité – expliquer les explications*

Un évènement donné n'a en général pas une cause unique, mais dépend d'une chaine causale [MIL 19 *op.cit.*] qu'on peut décrire par un graphe de causalité, chaque cause étant issue d'une ou plusieurs autres. L'utilisateur peut souhaiter obtenir des explications relatives à différentes portions de ce graphe et non aux seules causes immédiates.

Prouver l'existence de relations de causalité est difficile, en raison du *problème de la corrélation* mais aussi de la difficulté à prouver un lien de causalité dans le monde réel [HAL 05 *op.cit.*] [HAL 04]. Les humains ont tendance à considérer les évènements exceptionnels [KAH 81], récents [MIL 90] ou contrôlables [GIR 91] plus importants que les autres pour déterminer une cause. Cependant, ces critères subjectifs dépendent de l'appréciation de l'utilisateur. Il n'est donc pas possible de déterminer automatiquement la bonne profondeur d'exploration. Elle dépend de l'apprenant, du contexte et de l'*explicandum*. De plus, les méthodes d'attribution sont difficilement compatibles avec l'existence d'explications multi-niveaux[17] [ROS 20 *op.cit.*].

5.3. *Sélection, interaction, adaptation*

Pour extraire une sélection pertinente dans le graphe de causalité, il faut donc implémenter un dialogue ou une conversation [HIL 90] entre l'utilisateur et le système, lui permettant d'explorer le graphe de causalité pour approfondir l'explication. Le dialogue explicite les relations de causalité et accompagne ainsi l'utilisateur dans le franchissement du *dernier pas*. Il s'agit alors « d'expliquer l'explication ». Le système interactif *Glass-Box* [SOK 20 *op.cit.*] permet à l'utilisateur de demander

[17] Cette expression « explication multi-niveaux » est souvent utilisée à propos des explications basées sur des chaines de causalité, par exemple [NAI 23 op.cit.] [FIN 21].

des explications contrefactuelles au moyen d'une interface audio en langage naturel, fournissant un exemple rare d'interaction se rapprochant d'un véritable dialogue.

Un système XAI peut également s'adapter à l'utilisateur grâce à des mécanismes de personnalisation, même s'il s'agit d'une option encore peu développée [SOK 20 *ibid.*]. De tels dispositifs faciliteraient la collaboration et l'expérience utilisateur [WEL 19 *op.cit.*]. Quelques expériences de sciences cognitives ont été réalisées afin de comparer différents types d'explications [NAI 23 *op.cit.*] [HER 22 *op.cit.*] ; ces travaux sont encore préliminaires. Produire des explications personnalisées nécessite la connaissance et la modélisation des préférences de l'apprenant ou de ses connaissances ou compétences [RIB 16 *op.cit.*] et du contexte, et requiert une compréhension de son modèle mental [SOK 20 *op.cit.*] [ZHA 20 *op.cit.*]. Il faut néanmoins éviter une personnalisation trop poussée, au détriment de la fidélité, qui peut amener l'utilisateur à se désintéresser de l'explication [KUL 13 *op.cit.*].

5.4. *La première vague passée au crible des critères de la seconde*

Au regard des critères de la seconde vague qui viennent d'être passés en revue, les explications de la première vague de l'XAI apparaissent très insuffisantes :

– Elles ne peuvent prétendre au statut d'*explication causale* véritable, en tant que dialogue, puisque dépourvues de moyens d'interaction. Les explications produites sont des produits, alors qu'elles devraient être des processus [FIN 21 *op.cit.*]. Même leur capacité d'*attribution causale* est discutable puisque cette attribution reste à l'appréciation de l'utilisateur.

– Elles n'apportent pas d'informations de nature sémantique. L'absence de prise en compte du sens est ce qui conduit un modèle IA à attribuer plus de sens à un article ou un adjectif qu'à des mots réellement porteurs de sens lors d'une tâche de classification de texte [RIB 16 *op.cit.*].

– Elles sont souvent non contrastées, ne s'attachant qu'aux résultats effectivement produits. Etant basées sur des données factuelles (le corpus d'apprentissage) elles ne peuvent pas fournir d'explications contrefactuelles robustes.

– Elles ne sont pas décomposables et donc, à la différence de l'esprit humain, incapables de produire une explication par étapes, un cheminement intellectuel séquentiel accumulant progressivement des indices expliquant la contribution au résultat de l'inférence de chaque aspect de l'information d'entrée [HEM 48 *op.cit.*].

5.5. *Illustration avec le cas d'usage*

Pour le cas d'usage, les explications produites par un système XAI de la seconde vague pourraient être les suivantes : les trafiquants recrutent souvent des capitaines ayant un passé judiciaire ; la coupure de l'AIS en dehors d'une zone de pêche n'est pas un comportement habituel (*décomposable*), et est suspecte *car* indiquant une volonté potentielle d'échapper aux contrôles ; l'accostage supposé est très suspect *car* il n'a pas de raisons de se produire entre les deux types de navires considérés (*attribution causale*) ; le pays d'origine n'est pas identifié comme source de trafic de stupéfiants (*explication contrastée*) ; une activité de pêche illégale aurait été envisagée s'il n'y avait pas suspicion d'accostage en mer avec le bateau de plaisance (*explication contrefactuelle*) ; bien que le port de départ ne soit pas identifié comme suspect, la proximité du camp implique la présence de réfugiés (*informations sémantiques*). En s'intéressant par exemple d'abord aux causes de premier niveau, puis en cherchant à comprendre pourquoi la menace de trafic de stupéfiants n'a pas été retenue, avant de s'intéresser de plus près à l'accostage, l'utilisateur interagit librement avec le système (*dialogue et explication causale*). Le système explique que cet accostage est suspect (niveau 1) parce que c'est un évènement rare entre ces deux types de navires, respectivement sous pavillons européen et hors UE (niveau 2). Pour étayer l'importance de ces éléments, il fournit des données statistiques sur le taux d'immigration clandestine à bord de navires dont le pavillon est hors UE et à destination de l'Italie, ou encore des exemples, dans le corpus d'apprentissage, de cas similaires avec accostage en mer (niveau 3 de l'*explication multi-niveaux*).

5.6. *Bilan de la seconde vague*

L'importance des sciences sociales pour l'X-XAI s'est imposée progressivement à la communauté de l'IA explicable. Cependant, la mise en œuvre concrète de leurs principes n'est pas simple : limites de l'XAI, problème général de l'identification des causes, compréhension du modèle mental de l'utilisateur, prise en compte des connaissances du domaine concerné… La capacité des outils reste limitée pour une interactivité satisfaisante. La seconde vague de l'XAI ne semble ainsi pas encore parvenir à s'affranchir des problèmes de la première vague, et doit de plus affronter de nouveaux défis.

La plupart des idées clés de cette seconde vague vise principalement le même objectif : montrer à l'utilisateur la nature causale des corrélations relevées par le modèle d'IA ou, autrement dit, résoudre le *problème de la corrélation*, objectif difficile pour des données observationnelles (*cf.* 4.3.1). La démarche interactionniste s'attaque alors au problème par un nouvel angle, en accompagnant l'utilisateur dans le *dernier pas*. La seconde vague est sans doute loin d'avoir exploré toutes les facettes des sciences humaines permettant de produire de meilleures explications. Elle a certes apporté un nouveau regard sur la production automatisée d'explications, mais un changement de perspective semble néanmoins nécessaire.

6. Vers la troisième vague

La première vague était essentiellement centrée sur « l'*explicandum* », la seconde s'est surtout intéressée à l'explicateur et à la façon de produire des explications mettant l'accent sur les relations causales. Considérant le quadruplet d'une explication (*cf. supra*, en début d'article), il reste à prendre en compte l'apprenant et le contexte.

Considérer, dans une « vision Diltheyenne » [DIL 77], une explication du point de vue de l'apprenant (*cf. supra,* page 2) plutôt que de l'explicateur peut apporter une première idée de ce que pourrait être la troisième vague. L'explicateur a pour rôle d'*expliquer* ; il est donc naturellement amené à intégrer l'explication dans une chaine causale, d'où l'importance de la causalité. L'apprenant a pour objectif de *comprendre* ; il est à la recherche du sens du phénomène observé. Pour prendre en compte l'apprenant, la *troisième vague* devrait centrer son attention sur la compréhension humaine et sur la capacité spontanée d'attribuer du sens aux phénomènes.

Cette partie présente une nouvelle conceptualisation de l'explicabilité, l'*explicabilité sémantique*[18], ou S-XAI, illustrée par des propositions d'architectures et un exemple de mise en œuvre dans le cadre du cas d'usage proposé.

6.1. *XAI sémantique*

Une explication consiste à donner du sens à la chose qu'on explique [MAL 06 *op.cit.*] [MIL 19 *op.cit.*]. De ce fait, disposer d'une représentation sémantique des connaissances aide à produire des explications [ARR 20 *op.cit.*]. Pourtant, la prise en compte d'éléments sémantiques est absente de la plupart des techniques de l'X-XAI ; les *traits* (tels que définis plus haut) ne portent pas d'information sémantique.

Trois idées maîtresses caractérisent la proposition S-XAI. La capacité d'un système à produire une explication satisfaisante dépend de sa capacité à représenter, manipuler et interpréter ou comprendre le sens des concepts liés à la chose à expliquer. Cette capacité est une fonction spécifique de l'esprit humain qui n'est pas présente dans les systèmes actuels d'IA ; elle doit donc y être insérée sciemment

[18] Le terme sémantique est utilisé dans cet article dans une acception plus large que celle usuellement employée, en tant que science de la signification, indépendamment de toute notion linguistique ou même de tout symbole (écartant ainsi également l'utilisation du terme *sémiotique*, plus général sans pour autant mieux correspondre à l'idée soutenue ici).

par des humains sous une forme ou une autre. Enfin, pour réaliser cette opération, « insérer du sens » dans l'IA, il est nécessaire d'ajouter un certain niveau de contrainte dans l'architecture du modèle, reflétant les connaissances sémantiques du domaine issues de l'expertise humaine : l'humain doit insérer une « vision du monde » dans le modèle.

Les considérations sur la causalité (cf. supra), qui constituent un élément central de la seconde vague, n'apparaissent que comme la conséquence d'un problème plus fondamental : expliquer, c'est faire comprendre. Or, la compréhension ne pouvant être déléguée au modèle, il est nécessaire que des experts humains aient intégré dans le système des éléments sémantiques matérialisant cette compréhension. Afin d'aller au-delà de l'attribution causale, les systèmes S-XAI doivent également permettre d'instaurer un dialogue, dans la lignée de certains travaux antérieurs [PAR 18] [DON 17].

Les deux premières idées maîtresses amènent à s'intéresser à la façon de représenter et manipuler du sens. La capacité de l'humain à donner du sens, à raisonner et à expliquer ses raisonnements est sous-tendue par l'existence de *représentations mentales (*ou *représentations internes)* représentant des concepts réels ou abstraits. Concevoir des systèmes manipulant des représentations sémantiques pour produire des explications est une idée déjà présente dans l'IA symbolique [MIC 83]. L'approche purement symbolique[19] a depuis montré ses limites ; la théorie des modèles mentaux développée par Johnson-Laird [JOH 83] a étayé l'idée que la représentation du sens chez les humains ne peut pas être assimilée à un système purement symbolique. En basant la production d'explications sur des principes proches de ceux de la cognition humaine, celles-ci seront probablement mieux compatibles avec les modèles mentaux et donc plus simples à comprendre.

Pour permettre de raisonner efficacement sur le monde, les représentations mentales doivent manifester des caractéristiques sémantiques telles que la productivité et la systématicité [FOD 88]. Il est généralement admis que les représentations mentales d'entités similaires doivent être elles-mêmes similaires pour permettre l'émergence de ces caractéristiques. Certains modèles d'apprentissage profond présentent effectivement de telles propriétés. Par exemple, dans des réseaux de neurones, la proximité sémantique entre deux concepts peut être exprimée par la distance entre les vecteurs correspondant à la représentation de ces concepts [MIK 13b]. Cette particularité permet même de développer un raisonnement analogique basé sur une arithmétique sémantique (« Rome est à l'Italie ce que Paris est à la France ») [MIK 13a]. Certains modèles d'apprentissage profond semblent donc fonctionner selon des principes analogues à la cognition humaine. Cependant, il faut également s'assurer que ces représentations permettent de différencier ce qui est vrai de ce qui ne l'est pas. Les problèmes récurrents des IA génératives avec les hallucinations [JI 23] montrent que ce point n'est pas résolu.

L'approche sémantique n'est en aucun cas contrainte à ne s'appliquer qu'à des modèles d'apprentissage profond ; la modélisation des connaissances sous forme de représentations structurées peut s'effectuer au moyen d'une IA symbolique déployée dans un système hybride [GAR 19]. Ce constat conduit à proposer une première forme de S-XAI basée sur une approche hybride. La *S-XAI hybride* utilise IA hybride et apprentissage profond, bénéficiant ainsi de leurs avantages combinés. La contrainte imposée (troisième idée maîtresse) provient ici de l'utilisation d'IA symbolique. IA symbolique et apprentissage profond sont complémentaires ; la première est performante pour la représentation des connaissances exprimables sous la forme de faits et de règles, les seconds sont mieux adaptés pour d'autres tâches, notamment analyser et traiter des données analogiques, visuelles ou sonores.

En confiant aux seuls concepteurs experts la responsabilité de structurer les connaissances sémantiques, la S-XAI hybride n'exploite pas la capacité des modèles d'apprentissage profond à

[19] Par exemple les langages formels [CHO 14], le fonctionnalisme computationnel [PUT 73] ou le computationnalisme [FOD 97].

développer des représentations internes sémantiques. Une autre approche consiste donc à guider ou orienter les représentations internes tout en laissant suffisamment de liberté au modèle pour bénéficier de sa puissance représentationnelle, plutôt que d'essayer *a posteriori* de donner du sens à un système qui n'a pas été conçu pour cela ; on la définit comme « S-XAI intrinsèque »[20]. La connaissance et le savoir-faire des experts sont ici utilisés pour orienter, valider ou contraindre les représentations internes en veillant à ce qu'elles soient directement interprétables, car basées sur les concepts associés au problème à résoudre. Ces représentations internes sont nommées *traits sémantiques*, dans une acception sans doute proche de ce que [RUD 19 *op.cit.*] entend par *traits significatifs* (*meaningful features*). Concrètement, les experts peuvent collaborer avec les chercheurs en IA, par exemple pour analyser l'espace vectoriel de représentation et vérifier la cohérence avec la conceptualisation du domaine ; ils peuvent aussi proposer des représentations sémantiques spécifiques pour exprimer des concepts non atomiques pouvant difficilement s'exprimer sous la forme d'un seul vecteur. Les traits sémantiques ne sont pas nécessairement associés à des concepts simples tels que des objets, ou des caractéristiques d'objets. Ils peuvent également représenter les étapes intermédiaires d'un raisonnement, des concepts composites ou hiérarchisés par niveau d'abstraction. L'extraction de traits sémantiques permet d'élaborer des représentations internes correspondant aux concepts manipulés par les experts du domaine. En combinaison avec des mécanismes interactifs de dialogue, cette architecture serait ainsi en mesure de produire des explications similaires à celles fournies par les experts.

La S-XAI intrinsèque introduit une part de transparence dans le modèle au moyen d'une architecture modulaire. Des modèles opaques traitent les informations d'entrée pour produire les traits sémantiques. Ces derniers sont ensuite utilisés de deux façons. Premièrement, ils sont envoyés vers un module de classification sémantique, pour réaliser la tâche dévolue au système[21]. Deuxièmement, ils sont utilisés pour produire des explications. Etant basée sur les états internes du modèle, à la différence des méthodes *post hoc*, la S-XAI intrinsèque limite significativement le risque que les explications ne reflètent pas correctement son fonctionnement [RUD 19 *op.cit.*]. Pour améliorer encore leur fiabilité, il est possible de combiner la S-XAI intrinsèque avec l'XAI auto-explicative [GUI 22 op.cit.], en ajoutant un module générant les explications. En introduisant des traits sémantiquement interprétables entre des modules opaques, les modèles de S-XAI intrinsèque, *semi-transparents*, réalisent un nouveau compromis entre puissance et capacité à être expliqués. Il existe déjà dans la littérature des méthodes intrinsèques pour des modèles opaques [ADA 18 *op.cit.*], dont certaines mentionnent explicitement l'idée d'introduire des notions de sémantique [DON 17 *op.cit.*] [MAR 19 *op.cit.*].

Les deux approches, *S-XAI hybride* et *S-XAI intrinsèque*, ne sont pas exclusives et leur utilisation conjointe est sans doute une piste prometteuse. Dans cette conjonction, l'IA symbolique modélise les connaissances formalisables, tandis que les modèles opaques réalisent des tâches plus complexes ou moins formalisables et traitent les signaux analogiques. La Figure 1 illustre une architecture possible d'un modèle combinant S-XAI, hybride et intrinsèque, et XAI auto-explicative. Dans cette illustration, la représentation des traits sémantiques sous la forme de nœuds ne présuppose pas de leur implémentation réelle. Pour permettre des représentations internes structurées exhibant les caractéristiques des représentations internes de la cognition humaine, ils pourraient correspondre à des structures plus élaborées. Il s'agit uniquement de premières pistes.

[20] Le terme "intrinsèque » a un sens précis en XAI, qui a été introduit dans la taxonomie (§4.2).

[21] La tâche de classification sémantique est retenue par cohérence avec le cas d'usage, mais la S-XAI peut être appliquée à tous types de tâches, par exemple à un système d'aide à la décision.

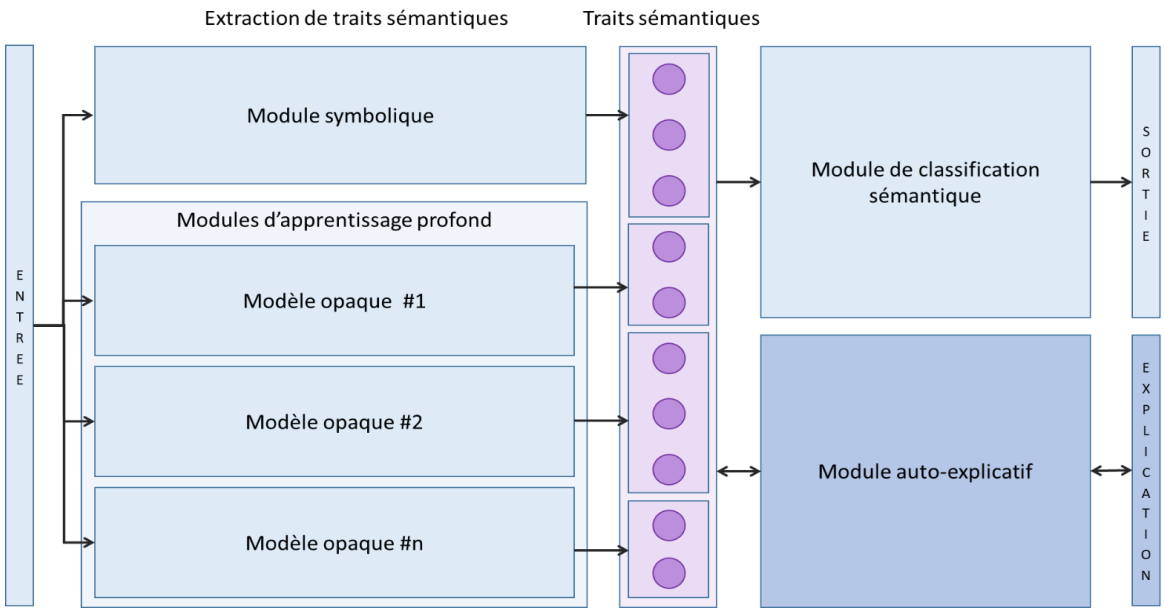

Figure 1. *Architecture générique d'un modèle S-XAI combinant les approches hybride, intrinsèque, et XAI auto-explicative. Un module d'IA symbolique ainsi qu'un ou plusieurs modèles opaques extraient, à partir de l'information d'entrée, différents traits sémantiques. Ceux-là sont utilisés comme nouvelles entrées, d'une part d'un classificateur fournissant la sortie, et d'autre part d'un générateur d'explications (dans le cas d'un modèle S-XAI auto-explicatif).*

Sur le plan technique, des algorithmes de *Machine Learning* tels que les modèles attentionnels développés par certains auteurs [XU 15] [VAS 17] [XIA 17] [LU 16], eux-mêmes inspirés du fonctionnement des aires cérébrales [REN 00] [COR 02], pourraient être utilisées pour simplifier les premières étapes de l'identification de traits sémantiques, de même que les modèles de type LLM pour des données textuelles ; des auto-encoders pourraient permettre l'extraction automatique de traits sémantiques.

6.2. *Application au cas d'usage maritime*

Les traits sémantiques sont élaborés grâce à la collaboration d'experts des différents domaines (IA, maritime, judiciaire, construction navale…). Entrent ainsi en ligne de compte le profil judiciaire de l'équipage, les types de zones maritimes, les occurrences et conjonctions suspectes de comportements, d'évènements ou de coordonnées (*e.g.* coupure de l'AIS en dehors d'une zone de pêche[22], accostages en mer entre navires respectivement sous et hors pavillon UE), la reconnaissance sémantique des différentes parties d'un navire permettant afin de les reconnaitre lorsque cela n'est pas possible par simple identification (double identification, mauvaises conditions, camouflage, ...), etc. Des modules d'IA symbolique traitent les données peu bruitées : éphémérides, données AIS, données textuelles issues de bases de données ; des modèles opaques traitant les signaux analogiques ou plus bruitées : photographies, images satellites, communications radio, cartes météorologiques, images de radars côtiers, etc. Les traits sémantiques alimentent le module de classification sémantique chargé de détecter les comportements suspects et d'identifier le type de risque. Ce module est entrainé à partir de classifications réalisées par des experts du domaine sur des cas réels.

Un tel système serait capable de produire des explications sémantiques telles que l'exemple fourni en 4.3.1 (stratégie d'évitement des radars côtiers, etc.), expliquant pour quelle raison la coupure de l'AIS est suspecte, à la différence des modèles d'XAI classiques qui se limiteraient à relever une corrélation entre coupure de l'AIS et comportement suspect.

[22] Ces conjonctions constituent des exemples de traits sémantiques composites.

Grâce à un meilleur contrôle des représentations internes[23], l'architecture informatique évite à la fois le risque inhérent aux techniques *post hoc*, consistant à produire une explication sans rapport avec l'inférence réellement réalisée, mais aussi de produire une explication dénuée de réalisme fonctionnel[24]. En cas de doute sur celle-ci, il est possible de fournir l'accès aux traits sémantiques afin de permettre aux utilisateurs expérimentés d'en faire une analyse *a posteriori*. Le module d'explication propose une interface en langage naturel pour répondre aux questions de l'utilisateur, proposer des exemples ou contre-exemples, approfondir les explications ou remonter dans le graphe de causalité.

6.3. *S-XAI : une solution aux défis de l'XAI*

En associant une architecture basée sur les traits sémantiques et une interface de dialogue, les modèles S-XAI réduisent le risque de fausses corrélations et aident à réduire la longueur du *dernier pas*. À mi-chemin entre une IA symbolique, où cette charge est entièrement portée par les concepteurs, et une IA opaque non contrainte apportant plus de possibilités créatives, mais au détriment de son explicabilité, la S-XAI propose une voie médiane. Par sa cohérence interne au regard du domaine et des connaissances des experts, le modèle simplifie notablement la production d'explications et facilite la compréhension de l'utilisateur par une meilleure congruence avec ses propres modèles mentaux.

Puisque la représentation des connaissances de l'utilisateur dépend de son niveau de compétence, utilisant des portions cognitives (*cognitive chunks*) différentes [NEA 03], la prise en compte du niveau d'expertise pourrait être réalisée en développant un modèle composé de plusieurs sous-systèmes parallèles, chacun d'entre eux utilisant des traits sémantiques différents. L'intégration de contraintes temporelles pourrait bénéficier du même type d'architecture, avec des modèles comportant un nombre distinct de traits sémantiques selon la séquencialité globale impartie (cf. Figure 2).

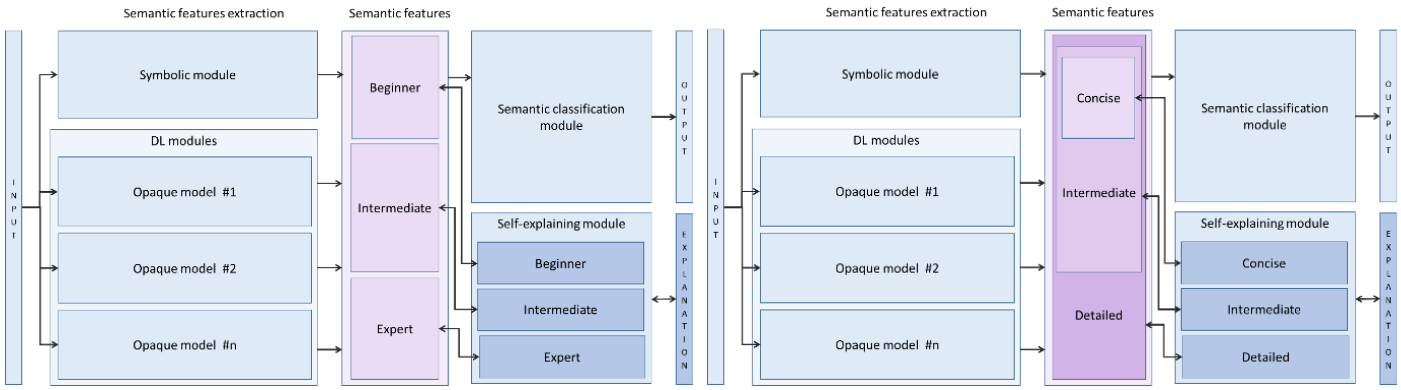

Figure 2. *Système S-XAI prenant en compte le niveau d'expertise (à gauche) ou le temps imparti (à droite). À gauche, en utilisant des traits sémantiques spécifiques au niveau de compétences des utilisateurs, il est possible de mieux se conformer à leurs modèles mentaux. À droite, en sélectionnant le nombre de traits sémantiques et en les hiérarchisant selon leur importance, le module auto-explicatif est capable de générer des explications plus ou moins concises selon le temps imparti à la tâche.*

En organisant les traits sémantiques sous la forme d'un graphe de causalité, il devient possible de réaliser un système S-XAI proposant des mécanismes d'explication multi-niveaux. Une base de connaissances pourrait compléter le système pour enrichir les explications par des connaissances générales du domaine (Figure 3). Le module auto-explicatif produit alors une explication à partir des

[23] Les représentations internes (traits sémantiques) sont explicitement choisies par les concepteurs (*cf. supra*).

[24] L'idée développée est que, grâce au contrôle des représentations internes (traits sémantiques), le réalisme fonctionnel devient intrinsèque à l'architecture du réseau.

traits de premier niveau (1), puis des explications supplémentaires correspondant aux niveaux suivants dans le graphe de causalité (2, 3) en fonction de l'interaction avec l'utilisateur.

L'approche S-XAI doit permettre de développer des mécanismes plus fiables pour la génération d'explications contrastées ou contrefactuelles, en évitant l'écueil du manque de réalisme ou de rupture de cohérence avec le domaine. Par ses analogies avec la cognition humaine, elle pourrait améliorer le modèle mental de l'utilisateur [SOK 20 *op.cit.*] ; les résultats et erreurs du modèle seront plus faciles à appréhender [KAS 88], et la relation de confiance entre l'humain et la machine pourrait être significativement améliorée ([KUL 13 *op.cit.*] [ZHA 20 *op.cit.*] [HOF 18 *op.cit.*]), contribuant ainsi à valoriser la performance globale.

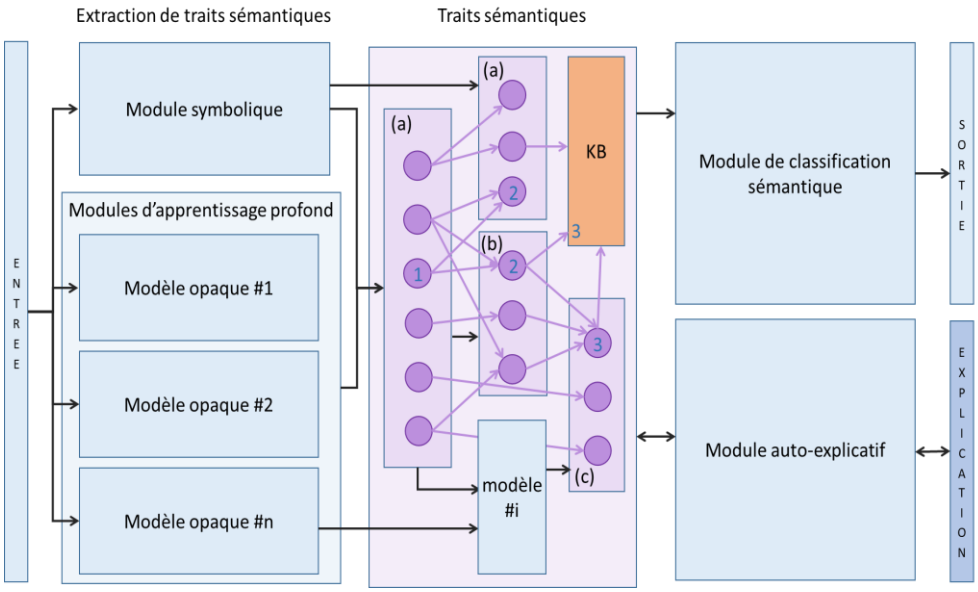

Figure 3. *Système S-XAI proposant un mécanisme d'explications multi-niveaux. Les traits sémantiques sont hiérarchisés en fonction du graphe de causalité établi par des experts. Chaque flèche violette indique une cause potentielle. Le graphe de causalité du cas traité est déterminé par les traits sémantiques activés, à partir des données d'entrée (a), d'autres traits sémantiques (b), ou de modèles supplémentaires (c). Une base de connaissances (KB) peut fournir des explications relatives aux connaissances générales du domaine. La représentation des traits sémantiques sous la forme de nœuds simples est uniquement illustrative ; les traits pourraient être implémentés au moyen d'un modèle IA dédié, par exemple un réseau sémantique.*

L'approche S-XAI et les modèles proposés ne résolvent cependant pas certaines difficultés de l'XAI. La personnalisation, qui nécessite de simuler le modèle mental spécifique de l'utilisateur, constitue un problème non résolu et probablement pas selon certains auteurs dans un futur proche [SOK 20 *op.cit.*]. L'approche dialectique de l'explication est un autre problème. Les explications ne seront en effet pas les mêmes selon qu'il s'agisse d'expliquer ou de convaincre. L'objectif de l'explication devrait même pouvoir évoluer au cours de l'interaction, visant d'abord à expliquer, puis à convaincre si un opérateur perd trop de temps alors qu'il a d'autres tâches à réaliser.

7. Conclusion

Cet article a présenté les principales solutions de l'X-XAI pour expliquer aux utilisateurs les résultats produits par les modèles IA, en proposant une séparation en deux vagues distinctes. Les techniques de la première vague de l'XAI, centrées sur l'*explicandum*, ne prennent pas réellement en compte les aspects humains d'une explication. Étant le plus souvent basées sur la construction de modèles adjacents au modèle IA ou s'y substituant, elles ne peuvent garantir que l'explication proposée soit conforme au traitement réellement effectué par le modèle original. La seconde vague a répondu à certaines de ces difficultés en s'inspirant des SHS. Elle a mis l'accent sur les relations causales, grâce aux explications contrastées ou contrefactuelles et en promouvant une meilleure

interactivité. Cependant, elle n'a pas répondu à tous les défis posés par l'X-XAI. Etant centrée sur l'explicateur et la production d'explications, elle n'a pas exploité tous les apports potentiels des sciences de la cognition humaine dans la compréhension des processus cognitifs. Cette seconde vague n'a pas non plus permis d'avancée significative pour la prise en compte des éléments contextuels.

En déplaçant l'attention de l'explicateur vers l'apprenant et en s'intéressant aux mécanismes de la compréhension, une nouvelle approche de l'X-XAI est proposée. Partant du constat que la faculté de compréhension est une caractéristique de l'esprit humain et que les modèles AI ne sont pas en capacité de déterminer l'origine éventuellement causale de leurs inférences, l'idée proposée dans cet article repose sur le postulat suivant : afin de garantir la production d'explications cohérentes avec le domaine d'application considéré, il est nécessaire de contraindre les modèles pour que la dimension sémantique et causale apportée par l'expertise humaine se reflète dans leur architecture, leur fonctionnement ou leurs algorithmes. Cette nouvelle approche, nommée S-XAI ou explicabilité sémantique, pourrait constituer une troisième vague de l'XAI. La proposition est illustrée par une architecture générique, combinant IA symbolique, apprentissage profond et XAI auto-explicative. Des architectures plus spécifiques sont aussi proposées pour générer des explications tenant compte du temps imparti, du niveau de compétence de l'utilisateur, ainsi que des explications multi-niveaux permettant de naviguer dans le graphe de causalité.

La S-XAI, à mi-chemin entre les systèmes très contraints de l'IA symbolique et les modèles opaques, pourrait ainsi bénéficier de leurs avantages respectifs tout en produisant des explications compréhensibles par l'utilisateur. Conjointement avec l'approche sémantique, les IA auto-explicatives pourraient contribuer à limiter les risques de produire des explications ne reflétant pas les inférences réellement réalisées par le modèle. Afin d'accompagner l'utilisateur dans sa démarche d'appropriation du sens et lui permettre de franchir le *dernier pas* de la compréhension, les systèmes S-XAI doivent également implémenter une forme de dialogue.

Cette approche en devenir nécessite de faciliter la collaboration interdisciplinaire en XAI : spécialistes en interaction humain-machine pour produire une expérience utilisateur fluide et efficace, spécialistes du domaine d'application considéré pour structurer les connaissances sous forme de traits sémantiques, spécialistes de sciences cognitives pour améliorer les connaissances sur la nature de l'expertise humaine et étudier l'efficacité des modèles, chercheurs en IA pour concevoir des architectures basées sur les modèles les plus récents, spécialistes en linguistique, sémantique et logique pour mettre au point des processus de contrôle afin d'éviter les erreurs de raisonnement ou hallucinations typiques des IA génératives.

Les perspectives sont nombreuses. La S-XAI pourrait contribuer à résoudre les principales difficultés auxquelles se sont heurtées les deux premières vagues, pour la production d'exemples et de contre-exemples, ainsi que pour l'exploration et à la sélection interactive des graphes de causalité, permettant à l'XAI d'évoluer de l'attribution causale vers l'explication causale. Cette approche pourrait aider à fournir des explications plus adaptées au besoin de l'utilisateur. Des travaux existants déjà cités explorent des idées similaires, avec des résultats encourageants qui confirment la validité de la démarche. La S-XAI pourrait améliorer la relation de confiance entre l'humain et la machine, en raison d'une plus grande facilité pour l'utilisateur à construire un modèle mental fiable du système. Elle pourrait aussi améliorer la fiabilité des corpus d'apprentissage constitués à partir de données synthétiques, en limitant les biais grâce à la contrainte interne apportée par les traits sémantiques. Pour parvenir aux meilleurs résultats, l'IA explicable doit être construite comme une interaction entre l'humain et la machine. Elle doit proposer à l'utilisateur des outils pour faire des recherches, pour explorer les graphes de causalité, pour poser et vérifier des hypothèses ou pour étudier des scénarios contrefactuels.

La S-XAI n'apporte cependant pas de solution immédiate à certaines difficultés de l'XAI, telles que la personnalisation ou la prise en compte des aspects dialectiques, nécessitant d'adapter les

explications selon le profil de l'utilisateur, la tâche à réaliser ou le contexte. Cette prise en compte constituera vraisemblablement, avec la personnalisation, l'un des enjeux futurs de l'explicabilité.

Bibliographie

[ADA 18] ADADI A., BERRADA M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, 6, 52138-52160, 2018.

[ALB 22] ALBINI E., LONG J., DERVOVIC,D., MAGAZZENI D. Counterfactual shapley additive explanations. 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT'22). Seoul (CO) 21-24 June 2022. FAccT'22 Proceedings, pp.1054-1070, 2022.

[ANC 19] ANCONA M., CEOLINI E., ÖZTIRELI C., GROSS M. Gradient-based attribution methods. In W.Samek, G.Montavon, A.Vedaldi, L.K.Hansen, K.R.Müller (eds.) *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, New-York (USA, NY): Springer Cham, pp.169-191, 2019.

[ANG 22] ANGWIN J., LARSON J., MATTU S., KIRCHNER L. Machine Bias. *Ethics of data and analytics*. Auerbach Publications, pp.254-264, 2022.

[ARR 20] ARRIETA A.B., DÍAZ-RODRÍGUEZ N., DEL SER J., BENNETOT A., TABIK S., BARBADO A., GARCIA S., GIL-LOPEZ S., MOLINA D., BENJAMINS R., CHATILA R., HERRERA F. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58, pp.82-115, 2020.

[ATH 18] ATHALYE A., CARLINI N., WAGNER D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *35th International Conference on Machine Learning (ICML 2018)*, Stockholm (SE) 10-15 July 2018. *Proceedings of Machine Learning Research*, 80, pp.274-283, 2018.

[BAR 80] BAR-HILLEL, M. The base-rate fallacy in probability judgments. *Acta Psychologica*, 44, *3*, pp.211-233, 1980.

[BEN 88] BENNETT J. Events and their Names. Oxford (UK): Oxford University Press. 1988.

[BRO 92] BROMBERGER S. On what we know we don't know: Explanation, theory, linguistics, and how questions shape them. Chicago (USA, MI): University of Chicago Press. 1992.

[BYR 19] BYRNE R.M. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. 38th International Joint Conference on Artificial IntelligenceIn (IJCAI 19), Macao (MO) 10-16 August 2019. Proceedings, pp.6276-6282, 2019.

[CAB 19] CABITZA F., CAMPAGNER A., CIUCCI D. New Frontiers in Explainable AI: Understanding the GI to Interpret the GO. *3rd International Cross-Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE)*, Canterbury (UK) 26-29 August 2019. Proceedings, pp.27-47, 2019.

[CAI 19] CAI C.J., JONGEJAN J., HOLBROOK J. The effects of example-based explanations in a machine learning interface. *International Conference on Intelligent User Interfaces, Proceedings (IUI)*, Part F147615. New-York (USA, NY): Proceedings by *Association for Computing Machinery*, pp.258-262, 2019.

[CHA 97] CHAJEWSKA U., HALPERN J.Y. Defining explanation in probabilistic systems. 30th Conference on Uncertainty in Artificial Intelligence, Providence (USA, RI): Morgan Kaufmann Publishers Inc., Proceedings, pp.62-71, 1997.

[CHE 22] CHEN B., HUANG K., RAGHUPATHI S., CHANDRATREYA I., DU Q., LIPSON H. Automated discovery of fundamental variables hidden in experimental data. *Nature Computational Science*, 2, *7*, pp.433-442, 2022.

[CHO 14] CHOMSKY N. *Aspects of the Theory of Syntax*. Cambridge (USA, MA); MIT press, 2014.

[CHO 22] CHOU Y. L., MOREIRA C., BRUZA P., OUYANG C., JORGE J. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion*, 81, pp.59-83, 2022.

[COL 17] COLOMBO M., BUCHER L., SPRENGER J. Determinants of judgments of explanatory power: Credibility, generality, and statistical relevance. *Frontiers in Psychology*, 8, article 1430, 2017.

[COR 02] CORBETTA M., SHULMAN G.L. Control of goaldirected and stimulus-driven attention in the brain. *Nature reviews neuroscience*, 3, *3*, pp.201–215, 2002.

[DIL 77] DILTHEY, W. Ideas concerning a descriptive and analytic psychology (1894). *Descriptive psychology and historical understanding*. Dordrecht: Springer Netherlands, pp.21-120, 1977.

[DON 17] DONG Y., SU H., ZHU J., ZHANG B. Improving interpretability of deep neural networks with semantic information. IEEE conference on computer vision and pattern recognition (2017 IEEE/CVF CVPR). Honolulu (USA, HI) 21-26 July 2017. Washington (USA, DC): Proceedings by IEEE Computer Society, pp.4306-4314, 2017.

[DOS 17] DOSHI-VELEZ F., KIM B. Towards a rigorous science of interpretable machine learning. Preprint *arXiv:1702.08608*, 2017.

[ELM 90] ELMAN J.L. Finding structure in time. *Cognitive science*, 14, *2*, pp.179-211, 1990.

[ELM 93] ELMAN, J.L. Learning and development in neural networks: The importance of starting small. *Cognition*, 48, *1*, pp.71-99, 1993.

[ELT 20] ELTON D.C. Self-explaining AI as an alternative to interpretable AI. *13th International Conference on Artificial General Intelligence (AGI 2020)*, St. Petersburg (RU) 16-19 September 2020. Proceedings of Lecture Notes in Computer Science, 12177, Heidelberg (DE): Springer International Publishing, pp.95-106. 2020.

[FEL 19] FELLOUS J.M., SAPIRO G., ROSSI A., MAYBERG H., FERRANTE M. Explainable artificial intelligence for neuroscience: behavioral neurostimulation. *Frontiers in neuroscience*, 13, 1346, 2019.

[FIN 21] FINZEL B., TAFLER D.E., SCHEELE S., SCHMID U. Explanation as a process: user-centric construction of multi-level and multi-modal explanations. In KI 2021 - Advances in Artificial Intelligence: 44th German Conference on AI, Virtual Event (DE) 27 September - 1 October 2021. Berlin (DE): Proceedings by Springer International Publishing, pp.80-94, 2021.

[FOD 88] FODOR J.A., PYLYSHYN Z.W. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28, *1-2*, pp.3-71, 1988.

[FOD 97] FODOR J.A. The representational theory of mind. *American Behavioral Scientist*, 40, *6*, pp.829-841, 1997.

[FOG 64] FOGEL, R.W. Railroads and American Economic Growth: Essays in Economic History. Baltimore: John Hopkins Press, pp.207-237, 1964.

[FRE 03] FREEDMAN D.A. Structural equation models: A critical review. Technical Report n°651 of the University of California Berkeley Statistics Department. Berkeley (USA, CA): University of California Berkeley Press, 2003.

[GAR 19] GARCEZ A., GORI M., LAMB L.C., SERAFINI L., SPRANGER M., TRAN, S.N. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *Journal of Applied Logics*, 6, *4*, pp.611-632, 2019.

[GER 17] GERSTENBERG T., PETERSON M.F., GOODMAN N.D., LAGNADO D.A., TENENBAUM J.B. Eye-tracking causality. *Psychological Science*, 28, *12*, pp.1731-1744, 2017.

[GIL 18] GILPIN L.H., BAU D., YUAN B.Z., BAJWA A., SPECTER M., KAGAL L. Explaining explanations: An overview of interpretability of machine learning. *5th IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. Turin (IT): 1–4 October 2018, *IEEE proceedings*, pp.80-89, 2018.

[GIR 91] GIROTTO V., LEGRENZI P., RIZZO A. Event controllability in counterfactual thinking. *Acta Psychologica*, 78, *1-3*, pp.111-133, 1991.

[GRI 75] GRICE H.P. Logic and conversation. In P. Cole, J.L. Morgan (eds.) *Syntax and Semantics, Vol. 3, Syntax & semantics 3: Speech arts.* New York (USA, NY): Academic Press, pp.41-58, 1975.

[GUI 18] GUIDOTTI R., MONREALE A., RUGGIERI S., TURINI F., GIANNOTTI F., PEDRESCHI D. A survey of methods for explaining black box models. *ACM computing surveys* (CSUR), 51, *5*, pp.1-42, 2018.

[GUI 22] GUIDOTTI R. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, On line Springer Nature, pp.1-55, 2022.

[GUN 19] GUNNING D., AHA D. DARPA's explainable artificial intelligence (XAI) program. *AI magazine*, 40, *2*, pp.44-58, 2019.

[GUY 22] GUYOMARD V., FESSANT F., GUYET T., BOUADI T., TERMIER A. VCNet: A self-explaining model for realistic counterfactual generation. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Cham: Springer International Publishing, pp.437-453, 2022.

[HAL 04] HALL N. Two concepts of causation. In J. Collins, N. Hall, L.A. Paul (eds.) *Causation and Counterfactuals*, Cambridge (USA, MA): MIT Press, pp.225-276, 2004.

[HAL 05] HALPERN J.Y., PEARL J. Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science*, 56, *4*, pp.843-887, 2005.

[HAR 21] HARIHARAN S., VELICHETI A., ANAGHA A.S., THOMAS C., BALAKRISHNAN N. Explainable artificial intelligence in cybersecurity: A brief review. *4th International Conference on Security and Privacy (ISEA-ISAP)*. Dhanbad (IN) 27-30 Octobre 2021. *IEEE Proceedings*, pp.31-42, 2021.

[HE 15] HE K., ZHANG X., REN S., SUN J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *IEEE International Conference on Computer Vision (ICCV)*. Santiago (CL) 7-13 December 2015. IEEE Proceedings, pp.1026-1034, 2015.

[HEL 19] HELBING D. Societal, economic, ethical and legal challenges of the digital revolution: from big data to deep learning, artificial intelligence, and manipulative technologies. In D.HELBING (ed.) *Towards Digital Enlightenment: Essays on the Dark and Light Sides of the Digital Revolution*. Berlin (DE): Springer, Cham, pp.47-72, 2019.

[HEM 48] HEMPEL C.G., OPPENHEIM P. Studies in the Logic of Explanation. *Philosophy of science*, 15, *2*, pp.135-175, 1948.

[HER 22] HERM L.V., HEINRICH K., WANNER J., JANIESCH C. Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability. *International Journal of Information Management*, 102538. 2022.

[HIL 90] HILTON D.J. Conversational processes and causal explanation. *Psychological Bulletin*, 107, *1*, pp.65-81, 1990.

[HOF 18] HOFFMAN R.R., MUELLER S.T., KLEIN G., LITMAN J. Metrics for explainable AI: Challenges and prospects. Preprint *arXiv*, 1812.04608, 2018.

[HOF 98] HOFFRAGE U., GIGERENZER G. Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, 73, *5*, pp.538-540, 1998.

[HOF 11] HOFFMAN R., KLEIN G., MILLER J. Naturalistic investigations and models of reasoning about complex indeterminate causation. *Information Knowledge Systems Management*, 10, *1-4*, pp.397-425, 2011.

[HOL 19] HOLZINGER A., LANGS G., DENK H., ZATLOUKAL K., MÜLLER H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9, *4*, e1312, 2019.

[JI 23] JI Z., LEE N., FRIESKE R., YU T., SU D., XU Y., ISHII E., BANG Y.J., MADOTTO A., FUNG P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55, *12*, pp.1-38, 2023.

[JOH 83] JOHNSON-LAIRD P.N. *Mental models: Towards a cognitive science of language, inference, and consciousness* Cambridge (UK): Cambridge University Press, 1983.

[KAH 81] KAHNEMAN D., TVERSKY A. *The simulation heuristic*. Stanford (USA, CA): Stanford University, 1981.

[KAS 88] KASS R., FININ T. The need for user models in generating expert system explanations. *International Journal of Expert Systems*, 1, *4*, 1988.

[KAU 20] KAUR H., NORI H., JENKINS S., CARUANA R., WALLACH H., WORTMAN V.J. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. 2020 CHI conference on human factors in computing systems. Honolulu (USA, HI) 25-30 April 2020. Proceedings New-York (USA, NY): Association for Computing Machinery, pp.1-14, 2020.

[KEI 06] KEIL F.C. Explanation and understanding. *Annual review of psychology*, 57, pp.227-254, 2006.

[KLE 14] KLEIN G., RASMUSSEN L., LIN M.H., HOFFMAN R.R., CASE J. Influencing preferences for different types of causal explanation of complex events. *Human factors*, 56, *8*, pp.1380-1400, 2014.

[KUL 13] KULESZA T., STUMPF S., BURNETT M., YANG S., KWAN I., WONG W.K. Too much, too little, or just right? Ways explanations impact end users' mental models. *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. San Jose (USA, CA) 15-19 September 2013. IEEE Proceedings, pp.3-10, 2013.

[LAI 19] LAI V., TAN C. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. *Conference on fairness, accountability, and transparency (ACM FAT)*. Atlanta (USA, GA) 29-31 January 2019. Proceedings, pp.29-38, 2019.

[LAU 19] LAUGEL T., LESOT M.J., MARSALA C., RENARD X., DETYNIECKI M. The dangers of post-hoc interpretability: unjustified counterfactual explanations. *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp.2801-2807, 2019.

[LEW 86] LEWIS D. Causal explanation. *Philosophical Papers*, 2, pp.214-240, 1986.

[LIA 20] LIAO Q.V., GRUEN D., MILLER S. Questioning the AI: informing design practices for explainable AI user experiences. *2020 CHI Conference on Human Factors in Computing Systems (CHI'20),* Honolulu (USA, HI) 25-30 April 2020. New York (USA, NY): Proceedings by Association for Computing Machinery. pp.1-15, 2020.

[LIP 16] LIPTON Z.C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16, *3*, pp.31-57, 2016.

[LIP 90] LIPTON P. Contrastive explanation. *Royal Institute of Philosophy Supplements*, 27, pp.247-266, 1990.

[LIU 20] LIU Y., JAIN A., ENG C., WAY D.H., LEE K., BUI P., COZ D. A deep learning system for differential diagnosis of skin diseases. *Nature medicine*, 26, *6*, pp.900-908, 2020.

[LOM 07] LOMBROZO T. Simplicity and probability in causal explanation. *Cognitive psychology*, 55, *3*, pp.232-257, 2007.

[LOU 13] LOU Y., CARUANA R., GEHRKE J., HOOKER G. Accurate intelligible models with pairwise interactions. *19th ACM International Conference on Knowledge Discovery and Data Mining (KDD'13)*. Chicago (USA, IL) 11-14 August 2013. New-York (USA, NY): Proceedings by the Association for Computing Machinery, pp.623–631, 2013.

[LU 16] LU J., YANG J., BATRA D., PARIKH D. Hierarchical question-image co-attention for visual question answering. *30th Conference on Neural Information Processing Systems (NIPS 2016).* Barcelona (SP) 5-10 December 2016. NeurIPS Proceedings pp.289-297, 2016.

[LUN 17] LUNDBERG S.M., LEE, S.I. A unified approach to interpreting model predictions. *31st Conference on Neural Information Processing Systems* (NIPS'17), Long Beach (USA, CA) 4-9 December 2017. *Advances in neural information processing systems*, 30, 2017.

[LUN 18] LUNDBERG S.M., NAIR B., VAVILALA M.S., HORIBE M., EISSES M.J., ADAMS T., LEE S.I. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2, *10*, pp.749-760. 2018.

[MAL 06] MALLE B.F. *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Cambridge (MA, USA): MIT Press. 2006.

[MAR 19] MARCOS D., LOBRY S., TUIA D. Semantically Interpretable Activation Maps: what-where-how explanations within CNNs. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, Seoul (KR) 27-28 octobre 2019. IEEE proceedings, pp.4207-4215, 2019.

[MAR 21] MARKUS A.F., KORS J.A., RIJNBEEK P.R. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113, 103655, 2021.

[MIC 83] MICHALSKI R.S. A theory and methodology of inductive learning. In R.S.Michalski, T.J.Carbonell, T.M.Mitchell (eds.) Machine Learning: An Artificial Intelligence Approach. Palo Alto (USA, CA): TIOGA Publishing Co., pp.83-134, 1983.

[MIK 13a] MIKOLOV T., YIH W.T., ZWEIG G. Linguistic regularities in continuous space word representations. 2013 conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2013). Atlanta (USA, GE) 9-15 June 2013. Cambridge (USA. MA): Proceedings by MIT Press for the Association for Computational Linguistics (ACL), pp.746-751, 2013.

[MIK 13b] MIKOLOV T., CHEN K., CORRADO G., DEAN J. Efficient estimation of word representations in vector space. Preprint *arXiv:1301.3781*. 2013.

[MIL 90] MILLER D.T., GUNASEGARAM S. Temporal order and the perceived mutability of events: Implications for blame assignment. *Journal of personality and social psychology*, 59, *6*, pp.1111–1118, 1990.

[MIL 19] MILLER T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267, pp.1-38, 2019.

[MIL 21] MILLER T. Contrastive explanation: A structural-model approach. *The Knowledge Engineering Review*, 36, article e14, 2021.

[MOH 21] MOHSENI S., ZAREI N., RAGAN E.D. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11, *3-4*, pp.1-45, 2021.

[MON 18] MONTAVON G., SAMEK W., MÜLLER K.R. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing: A Review Journal*, 73, pp.1-15, 2018.

[NAI 23] NAISEH M., AL-THANI D., JIANG N., ALI R. How the different explanation classes impact trust calibration: The case of clinical decision support systems. *International Journal of Human-Computer Studies*, 169, 102941, 2023.

[NAS 19] NASR K., VISWANATHAN P., NIEDER A. Number detectors spontaneously emerge in a deep neural network designed for visual object recognition. *Science advances*, 5, *5*, article eaav7903, 2019.

[NEA 03] NEATH I., SURPRENANT A.M. *Human Memory: An Introduction to Research, Data, and Theory*. Harrisburg (USA, PA): Cengage Learning, 2003.

[NIC 07] NICHOLS A. Causal inference with observational data. *The Stata Journal*, 7, *4*, pp.507-541, 2007.

[NIK 21] NIKOLSKAIA K.Y., NAUMOV V.B. The relationship between cybersecurity and artificial intelligence. *2021 International Conference on Quality Management, Transport and Information Security, Information Technologies (IT&QM&IS).* Yaroslavl (RU): 06-10 September 2021. *Proceedings IEEE,* pp.94-97, 2021.

[ONE 16] O'NEIL C. Weapons of Math Destruction: How Big Data increases Inequality and Threatens Democracy. Portland (USA, OR): Broadway Books, 2016.

[OVE 11] OVERTON J.A. Scientific Explanation and Computation. *ExaCt,* 7 (July), pp.41-50, 2011.

[PAR 18] PARK D.H., HENDRICKS L.A., AKATA Z., ROHRBACH A., SCHIELE B., DARRELL T., ROHRBACH M. Multimodal explanations: Justifying decisions and pointing to the evidence. *IEEE conference on computer vision and pattern recognition* (2018 IEEE/CVF CVPR). Salt Lake City (USA, UT) 18-22 June 2018. Washington (USA, DC): Proceedings by IEEE Computer Society, pp.8779-8788, 2018.

[PEA 09] PEARL J. *Causality*. Cambridge (USA, MA): Cambridge University Press, 2009.

[PIN 88] PINKER S., PRINCE A. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition,* 28, *1-2*, pp.73-193, 1988.

[PLU 99] PLUNKETT K., JUOLA P. A connectionist model of English past tense and plural morphology. *Cognitive Science*, 23, *4*, pp.463-490, 1999.

[PUT 73] PUTNAM, H. Psychological Predicates: The nature of mental states. In W.H.Capitan, D.D.Merill (eds.) *Art, Mind, and Religion,* Pittsburgh (USA, MI): University of Pittsburgh Press, pp.37-48, 1973.

[RAS 18] RAS G., VAN GERVEN M., HASELAGER P. Explanation methods in deep learning: Users, values, concerns and challenges. In H.J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, M. van Gerven (eds.) *Explainable and interpretable models in computer vision and machine learning*. Heidelberg (DE): Springer, pp.19-36, 2018.

[REN 00] RENSINK R.A. The dynamic representation of scenes. *Visual cognition*, 7, *1-3*, pp.17–42, 2000.

[RIB 16] RIBEIRO M.T., SINGH S., GUESTRIN C. "Why should I trust you?": Explaining the predictions of any classifier. *KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco (USA, CA) 13-17 August 2016. New-York (USA, NY): Proceedings by the Association for Computing Machinery, pp.1135–1144, 2016.

[ROB 18] ROBEER M.J. Contrastive explanation for machine learning. Master Thesis in Business Informatics. Department of Information and Computing Sciences. Utrech (NL): Utrech University, 2018.

[ROS 20] ROSCHER R., BOHN B., DUARTE M.F., GARCKE J. Explainable machine learning for scientific insights and discoveries. *IEEE Access,* 8, pp.42200-42216, 2020.

[RUD 19] RUDIN C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence,* 1, *5*, pp.206-215, 2019.

[RUM 86] RUMELHART D.E., MCCLELLAND, J.L. On learning the past tenses of English verbs. In J.MCCLELLAND, D.RUMETHART, and the PDP Research Group (eds.) *Parallel distributed processing*, Cambridge (USA, MA): MIT Press., 2, *24*, pp.535-551, 1986.

[SAB 23] SAßMANNSHAUSEN T., BURGGRÄF P., HASSENZAHL M., WAGNER J. Human trust in otherware – a systematic literature review bringing all antecedents together. *Ergonomics*, 66, *7*, pp.976-998, 2023.

[SEL 17] SELVARAJU R.R., COGSWELL M., DAS A., VEDANTAM R., PARIKH D., BATRA D. Grad-cam: Visual explanations from deep networks via gradient-based localization. 2017 IEEE international conference on computer vision. Venice (IT) 22-29 October 2017. Proceedings, pp.618-626, 2017.

[SKO 14] SKOW B. Are there non-causal explanations (of particular events)? *The British Journal for the Philosophy of Science*, 65, *3*, pp.445–467, 2014.

[SOG 22] SØGAARD A. Shortcomings of Interpretability Taxonomies for Deep Neural Networks. *Advances in Interpretable Machine Learning and Artificial Intelligence (AIMLAI'22)*, Atlanta (USA, GA) 17-21 October 2022. Aachen (DE): RWTH Aachen University CEUR Workshop Proceedings, 3318, 2022.

[SOK 20] SOKOL K., FLACH P. One explanation does not fit all: The promise of interactive explanations for machine learning transparency. *KI-Künstliche Intelligenz,* 34, *2*, pp.235-250, 2020.

[SUN 17] SUNDARARAJAN M., TALY A., YAN Q. Axiomatic attribution for deep networks. *International conference on machine learning*, PMLR, pp.3319-3328, 2017.

[TAE 18] TAEHYUN H., SANGWON L., SANGYEON K. Designing explainability of an artificial intelligence system. Technology, Mind, and Society ACM Conference, Washington (USA, DC) 5–7 April 2018. New-York (USA, NY): Proceedings by American Psychological Association, 08A, p.13, 2018.

[TIN 17] TING D.S.W., CHEUNG C.Y.L., LIM G., TAN G.S.W., QUANG N.D., GAN A., WONG T.Y. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *Jama*, 318, *22*, pp.2211-2223, 2017.

[VAN 04] VAN LENT M., FISHER W., MANCUSO M. An explainable artificial intelligence system for small-unit tactical behavior. *9th National Conference on Artificial Intelligence, & 6th Conference on Innovative Applications of Artificial Intelligence*, San Jose (USA, CA) 25-29 july 2004, Proceedings pp.900-907.

[VAS 17] VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A.N., KAISE L., POLOSUKHIN I. Attention is all you need. 31st Annual Conference on Neural Information Processing Systems (NIPS 2017), Long Beach (USA, CA): 4-9 December 2017. In I.GUYON U.VON LUXBURG, S.BENGIO, H.WALLACH, R.FERGUS, S.VISHWANATHAN, R.GARNETT (eds.) *Advances in neural information processing systems* 30. New-York (USA, NY): Curran Associates, Inc, 2017.

[WAC 17] WACHTER S., MITTELSTADT B.D., RUSSELL C. Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harvard Journal of Law and Technology*, 31, *2*, pp.842-887, 2017.

[WEL 19] WELD D.S., BANSAL G. The challenge of crafting intelligible intelligence. *Communications of the ACM*, 62, *6*, pp.70-79, 2019.

[WOO 03] WOODWARD J., HITCHCOCK C. Explanatory generalizations, part I: A counterfactual account. *Noûs*, 37, *1*, pp.1-24, 2003.

[XIA 17] XIAO T., XU Y., YANG K., ZHANG J., PENG Y., ZHANG, Z. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. *IEEE conference on computer vision and pattern recognition*. Boston (USA, MA) 8–10 June 2015. Washington (USA, DC): IEEE Proceedings, pp.842-850, 2015.

[XU 15] XU K., BA J., KIROS R., CHO K., COURVILLE A., SALAKHUDINOV R., ZEMEL R., BENGIO Y. Show, attend and tell: Neural image caption generation with visual attention. *32nd International Conference on Machine Learning* (ICML 2015). Lille (FR) 6-11 July 2015. In F.R. BACH, D.M. BLEI (eds.) Proceedings of ICML'15 - MLR Workshop and Conference Proceedings, *37*, pp.2048-2057, 2015.

[YAN 20] YANG F., HUANG Z., SCHOLTZ J., ARENDT D.L. How do visual explanations foster end users' appropriate trust in machine learning? *25th International Conference on Intelligent User Interfaces*. Cagliari (IT) 18-20 March 2020. Proceedings, pp.189-201, 2020.

[ZHA 20] ZHANG Y., LIAO Q.V., BELLAMY R.K. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *2020 Conference on Fairness, Accountability, and Transparency (FAT'20)*, Barcelona (SP) 27-30 January 2020. New York (USA, NY): Proceedings by the Association for Computing Machinery, pp.295-305, 2020.

