IA & explicabilité

Nicolas Maudet, Grégory Bonnet, Gaël Lejeune, Dominique Longin

AfIA

Association française
pour l'Intelligence Artificielle

Bulletin N° 116

Association française pour l'Intelligence Artificielle

AfIA

AfIA
Association française
pour l'Intelligence Artificielle

PRÉSENTATION DU BULLETIN

Le Bulletin de l'AfIA vise à fournir un cadre de discussions et d'échanges au sein des communautés académique et industrielle. Ainsi, toutes les contributions, pour peu qu'elles aient un intérêt général pour l'ensemble des lecteurs, sont les bienvenues. En particulier, les annonces, les comptes rendus de conférences, les notes de lecture et les articles de débat sont très recherchés.
Le Bulletin contient également chaque trimestre un dossier plus substantiel qui porte : soit sur un thème liés à l'IA (2 numéros par an), soit sur des équipes de recherche en IA (1 fois par an), soit sur la Plate-forme Intelligence Artificielle PfIA (1 fois par an).
Le comité de rédaction se réserve le droit de ne pas publier des contributions qu'il jugerait contraire à l'esprit du bulletin ou à sa politique éditoriale. En outre, les articles signés, de même que les contributions aux débats, reflètent le point de vue de leurs auteurs et n'engagent qu'eux-mêmes.

Édito

Ce numéro de printemps du Bulletin de AfIA est est consacré à un dossier monté par Nicolas MAUDET (LIP6, Sorbonne Université) sur le thème « IA & Explicabilité ». Ce dossier met en lumière une thématique de recherche récente et cruciale pour l'IA, thématique bien représentée en France comme en témoignent les 12 contributions provenant d'autant d'équipes ou laboratoires français (voir page 4).

Dans la suite de ce Bulletin, vous retrouverez les rubriques habituelles, à savoir les comptes rendus de la Journée « Réalité Virtuelle & IA » du 9 mars 2022 (voir page 51), de la Journée Commune AfIA et réseau DEVS du 28 mars 2022 (voir page 52) et de la Journée « Perspectives et Défis de l'IA » du 7 avril 2022 (voir page 53). Vous y trouverez enfin la liste des thèses soutenues lors du trimestre écoulé, page 58. Enfin, la composition actuelle du Conseil d'Administration se trouve en quatrième de couverture de tous nos bulletins.

Encore un grand merci à tous les contributeurs et contributrices de ce numéro, sans oublier Gaël LEJEUNE pour sa relecture assidue.

Bonne lecture à tous !

Grégory BONNET
Rédacteur

AfIA
Association française
pour l'Intelligence Artificielle

SOMMAIRE

DU BULLETIN DE L'AfIA

AfIA
Association française
pour l'Intelligence Artificielle

Dossier

« IA et explicabilité »

Dossier réalisé par

Nicolas MAUDET
LIP6 / Équipe Systèmes Multi-agents
Sorbonne Université
nicolas.maudet@lip6.fr

Édito

Il existe de nombreuses situations dans lesquelles il apparaît crucial de doter un système d'IA de capacité explicative. Et si cette question est loin d'être nouvelle en intelligence artificielle, elle est redevenue centrale avec le développement récents de nombreux systèmes dits « boîtes noires » (il est utile de noter ici que l'usage de ce terme tend à confondre deux contextes très différents : celui d'un accès impossible au modèle – cas d'un système propriétaire –, et celui d'une complexité réputée trop grande du modèle, rendant difficile d'en comprendre la logique sous-jacente. Les conséquences en terme d'explications envisageables diffèrent).

L'explicabilité peut d'abord être importante dans la phase de développement d'une application, afin de permettre à un expert de débugger, réviser, faire évoluer son modèle.

En phase d'utilisation d'une application, les demandes de garanties de redevabilité des systèmes d'IA incitent aussi à développer de telles fonctionnalités explicatives. En effet, même si ce n'est pas le seul moyen d'assurer la redevabilité d'un système d'IA, la capacité à expliquer un résultat est souvent tenue pour être une condition nécessaire à son acceptabilité, en particulier dans le cas d'applications « à fort enjeu » pouvant significativement affecter les utilisateurs.

Les explications sont donc situées : elles ont un objectif qui dépend fortement du contexte d'usage et la forme comme le niveau de détail qu'elles doivent prendre pourront être très variables selon la personne à laquelle elles sont adressées. Elles peuvent aussi s'intégrer dans un cadre interactif plus riche, offrant en particulier la possibilité de les contester.

En réponse à l'appel lancé pour ce numéro du Bulletin de l'AfIA, douze contributions ont été reçues (présentations de départements, d'équipes, de projets ou de thèses en cours) et nous permettent de dresser un panorama d'une grande richesse sur les recherches menées dans le domaine de l'IA explicable en France, et d'en illustrer la variété aussi bien aussi niveau des applications que des méthodes employées.

- Les terrains applicatifs couvrent aussi bien le domaine éducatif, la planification de tournées, les maisons intelligentes, la modélisation financière, la sécurité (détection d'intrusion), la création de matériaux, que – sans surprise tant les conséquences peuvent y être critiques – une large palette d'applications médicales (pharmacologie, imagerie, etc.)
- Les méthodes étudiées vont de modèles réputés nativement interprétables (systèmes à base de règles, arbres de décision, graphes de connaissances, etc.), mais nécessitant néanmoins un traitement dédié afin d'en extraire des explications acceptables, à des approches *post-hoc* agnostiques au modèle comme les explications contre-factuelles, en passant par les approches hybrides qui cherchent dans une perspective neuro-symbolique à combiner connaissances et réseaux de neurones.

Enfin, sur la question souvent délicate de l'évaluation de la qualité des solutions proposées, le dossier illustre différentes stratégies qui sont en fait très complémentaires : garanties formelles (minimalité des explications, fidélité au modèle), expériences par simulations ou impliquant des utilisateurs (experts ou non).

Je conclus en remerciant encore une fois toutes les contributrices et tous les contributeurs de ce dossier.

AfIA
Association française
pour l'Intelligence Artificielle

Recherches en IA explicable dans l'équipe TWEAK du LIRIS : les traces d'interaction comme support de l'explication

Par

Béatrice FUCHS
LIRIS / TWEAK
Université Lyon 3

Nathalie GUIN

Marie LEFEVRE

Alain MILLE
LIRIS / TWEAK
Université Lyon 1
{prénom}.{nom}@liris.cnrs.fr

Introduction

Les travaux de l'équipe TWEAK s'inscrivent dans les disciplines de l'IA et de l'ingénierie des connaissances et explorent deux dimensions : les EIAH (Environnements Informatiques pour l'Apprentissage Humain) et le Web. Nous nous intéressons en particulier à permettre aux utilisateurs d'interagir en intelligence avec les dispositifs techniques numériques. Au-delà du développement de dispositifs intelligents, il s'agit que ces dispositifs facilitent leur appropriation par les utilisateurs. Cette appropriation suppose un apprentissage de l'utilisateur, pour qu'il puisse agir et adapter le dispositif à son contexte, à ses propres connaissances, à ses propres objectifs. Pour qu'un processus d'appropriation puisse advenir pendant l'interaction avec un dispositif technique numérique, il est nécessaire que les régulations encapsulées dans ce dispositif soient explicables lors de leur mise en œuvre par l'utilisateur.

Nous proposons de considérer les traces d'interaction comme le matériau de départ pour les processus d'explication. En effet, une trace d'interaction contient des éléments combinés issus des fonctions du dispositif et des actions de l'utilisateur. Formalisées, ces traces permettent de mener des calculs pour retrouver les schèmes de régulation, en interaction avec l'utilisateur. C'est l'utilisateur qui a l'initiative de guider la découverte de connaissance en cherchant à reformuler les interactions à un niveau d'abstraction rejoignant la manière dont il décrit sa propre activité. Depuis plus de 10 ans, l'étude du potentiel des traces modélisées est menée dans différents domaines, avec des fonctions d'assistance à l'appropriation des dispositifs et de leurs régulations [2].

L'explicabilité des dispositifs numériques est devenu un enjeu de société, à tel point que lorsque leur fonctionnement est régi par des règles issues des algorithmes d'apprentissage profond, elle soulève des questions éthiques non encore résolues et faisant l'objet d'une recherche soutenue [7]. Dans l'équipe TWEAK, nous considérons qu'un dispositif est explicable lorsqu'il réunit les conditions pour permettre à l'utilisateur de s'approprier la sémantique des régulations à l'œuvre dans ce dispositif. Cette définition impose au dispositif de pouvoir exploiter les descriptions explicites des règles disponibles dans l'environnement numérique, mais aussi de permettre à l'utilisateur d'explorer, à son initiative, le dispositif en fonctionnement. L'explicabilité ne dépend donc pas uniquement de l'explicitation, mais aussi de la capacité à la

projeter dans l'expérience utilisateur [1].

L'explicabilité en éducation

Dans le domaine éducatif, deux catégories d'utilisateurs doivent s'approprier l'environnement numérique : les apprenants et les enseignants. La question de l'appropriation des EIAH par les enseignants est essentielle pour que ces outils soient davantage utilisés dans l'enseignement. Il faut donc concevoir des systèmes que les enseignants pourront adapter afin qu'ils répondent à leurs besoins. Pour cela, l'enseignant doit pouvoir comprendre les décisions ou recommandations du système d'IA concernant ses élèves, pour avoir confiance dans le système et, d'une certaine façon, « faire corps » avec lui pour être capable d'expliquer le comportement conjugué du dispositif tel qu'il a été mis en place. Cela nécessite une représentation explicite des connaissances et des processus de décision. La question de l'explicabilité doit être prise en compte dès la conception.

Dans le cadre du projet ANR COMPER, nous travaillons à concevoir des modèles et des outils permettant de mettre en œuvre une approche par compétences pour accompagner l'apprentissage de manière personnalisée. Un référentiel de compétences défini par l'équipe pédagogique représente, pour un domaine donné, les compétences, connaissances et savoir-faire à acquérir. Les ressources pédagogiques (cours ou exercices) sont rattachées aux connaissances et savoir-faire du référentiel par les enseignants. À partir des traces d'interaction entre l'apprenant et les ressources pédagogiques, un profil de compétences de chaque apprenant est calculé. En utilisant une stratégie de personnalisation paramétrée par l'équipe pédagogique, le système recommande à l'apprenant des ressources pédagogiques adaptées à ses objectifs et à son profil de compétences.

Dans ce projet, nous cherchons à expliquer à l'apprenant et à l'enseignant, d'une part le calcul des taux de maîtrise représentés dans le profil de compétences et d'autre part les recommandations effectuées. Les explications destinées à l'apprenant visent à renforcer la confiance de l'apprenant envers le système, mais aussi à soutenir un processus d'auto-régulation de l'apprentissage chez les élèves et les étudiants [8]. En effet, l'explication, à partir des exercices réussis ou échoués mémorisés dans les traces, du taux de maîtrise d'un savoir-faire, peut amener l'apprenant à reconsidérer l'auto-évaluation qu'il en avait faite. L'explication des ressources recommandées, en justifiant pourquoi certains savoir-faire doivent être retravaillés, par rapport aux objectifs d'apprentissage de l'apprenant, l'invite à prendre du recul sur son apprentissage, en soulignant la différence entre les objectifs d'apprentissage et les notions maîtrisées, ou en identifiant des lacunes dans des prérequis. Les explications destinées à l'enseignant visent à renforcer la confiance de l'enseignant envers le système, mais aussi à lui permettre de paramétrer à la fois le calcul des profils de compétences et la stratégie de recommandation. Les explications à destination de l'enseignant sont donc plus détaillées que celles à destination des apprenants. Dans ce contexte, les capacités d'explications sont nécessaires, mais non suffisantes, il faut également que l'enseignant ait la possibilité d'agir sur les comportements et stratégies du système, ou au moins d'exprimer son avis sur un comportement qu'il faudrait améliorer. Pour cela, les traces d'interaction entre le système et ses utilisateurs (apprenants et enseignants) sont une source précieuse de connaissance [3]. Des métaconnaissances permettant au système d'analyser son propre comportement en s'observant [9] et peuvent utiliser ces traces pour détecter un dysfonctionnement ou une possibilité d'amélioration du fonctionnement.

Exploration interactive de traces

TRANSMUTE [5] est une approche de découverte interactive de connaissances à partir de traces qui met en évidence des régularités sous la forme de sous-séquences de types d'événements appelées épisodes séquentiels. Dans l'interface visuelle, les composants de la trace sont associés à des symboles graphiques choisis par l'utilisateur. Les épisodes sont localisés dans la trace pour améliorer leur compréhension dans leur contexte d'occurrence. L'utilisateur peut interagir avec les épisodes en les étiquetant afin de leur attacher une interprétation et construire un modèle du phénomène étudié. Ce modèle est mémorisé dans un système à base de traces et explicitement relié à la trace à partir de laquelle il a été construit. Il est explicable car il est possible de naviguer vers ses éléments constitutifs dans la trace d'origine.

KATIE [4] est une approche d'acquisition de connaissances qui vise à assister, en interaction avec l'utilisateur, le processus de modélisation et d'intégration des traces dans un système à base de traces, en détectant et corrigeant les erreurs résiduelles dans les données. À partir d'une trace fournie sous la forme d'un jeu de données brutes, KATIE utilise l'analyse de concepts formels [6] pour construire un modèle sous la forme d'une hiérarchie des types d'éléments constituant la trace et la proposer à l'utilisateur. Si le modèle proposé ne correspond pas aux connaissances que l'utilisateur possède sur les données, ce dernier peut introduire ses connaissances sous la forme de contraintes sur les données. Les données discordantes sous-jacentes qui expliquent ces désaccords sont extraites et montrées à l'utilisateur qui peut les rectifier. Une fois les données modifiées, KATIE réitère sur la construction du modèle jusqu'à ce qu'un consensus soit obtenu avec l'utilisateur. Ce dernier peut alors étiqueter les concepts pour mémoriser son interprétation des concepts. Finalement, le modèle de trace est créé dans le système à base de traces et la trace y est ensuite enregistrée conformément à ce modèle.

Ces approches donnent à l'utilisateur un rôle central dans la construction de connaissances et l'explicabilité y est une caractéristique recherchée. L'utilisateur peut lui-même choisir la représentation visuelle des éléments de la trace, et c'est *via* leur manipulation interactive que les règles sous-jacentes à la découverte de connaissances sont explicitées. Les modèles de connaissances obtenus après interprétation sont ainsi sémantiquement et explicitement reliés aux données dont ils sont issus.

Références

[1] P-A. Champin, B. Fuchs, N. Guin, and A. Mille. Explicabilité : vers des dispositifs numériques interagissant en intelligence avec l'utilisateur. In *Atelier Humains et IA, travailler en intelligence à EGC*, 2020.

[2] P-A. Champin, A. Mille, and Y. Prié. Vers des traces numériques comme objets informatiques de premier niveau. *Revue Intellectica*, 2013(59) :171–204, 2013.

[3] A. Cordier, M. Lefevre, P-A. Champin, A. Mille, O. Georgeon, and B. Mathern. Connaissances et raisonnement sur les traces d'interaction. *Revue STI - Série RIA*, 28 :375–396, 2014.

[4] B. Fuchs. Assister l'utilisateur à expliciter un modèle de trace avec l'analyse de concepts formels. In *Conférence IC 2017*, pages 151–162, 2017.

[5] B. Fuchs and A. Cordier. Interactive interpretation of serial episodes : experiments in musical analysis. In *Conference EKAW'2018*, LNAI 11 313, pages 131–146. Springer, 2018.

[6] B. Ganter, G. Stumme, and R. Wille. *Formal concept analysis : foundations and applications*, volume 3626. springer, 2005.

[7] A. Mille. Vers des dispositifs techniques numériques orientés éthiques ? *Revue Intellectica*, 2019(1) :119–163, 2019.

[8] L. Pierrot, C. Michel, J. Broisin, N. Guin, M. Lefevre, and R. Venant. Combiner les leviers de l'approche par compétences et de l'auto-régulation pour accompagner le travail en autonomie à l'université. Analyse du service COMPER. In *EIAH'2021*, pages 166–177, 2021.

[9] J. Pitrat. *Métaconnaissance : Futur de l'intelligence artificielle*. Hermès, 1990.

AfIA
Association française
pour l'Intelligence Artificielle

Recherches en IA explicable dans l'équipe LFI du LIP6

Par **Christophe MARSALA**
Isabelle BLOCH
Marie-Jeanne LESOT
Sabrina TOLLARI
Jean-Noël VITTAUT
LIP6 / LFI
Sorbonne Université, CNRS
{prénom}.{nom}@lip6.fr
http://lfi.lip6.fr

L'équipe *Learning Fuzzy and Intelligent systems* (LFI) du laboratoire LIP6 de Sorbonne Université développe des recherches en interprétabilité des méthodes d'intelligence artificielle dans les domaines de l'aide à la décision, la science des données et l'apprentissage automatique. Les objectifs scientifiques et applicatifs sont de concevoir et de proposer des approches à la fois explicables durant leur construction et lors de leur utilisation.

Cet article présente brièvement ces contributions développées dans le cadre de multiples collaborations, en évoquant les tâches d'apprentissage automatique, dans des approches *by design* (dès la conception) ou *post-hoc*, la caractérisation de données par résumés linguistiques, l'interprétation d'images ou la formulation d'explications dans un cadre logique.

Variables linguistiques et logique floue

La logique floue a été proposée, par Zadeh en 1965, avec l'objectif de modéliser le raisonnement humain, fournissant dès sa conception même un outil pour faciliter l'interprétation des manipulations effectuées : la représentation de degrés de vérité au-delà d'une dichotomie manichéenne vrai/faux, de transitions progressives entre ces cas extrêmes et de limites imprécises donne une plus grande souplesse et une meilleure lisibilité des traitements réalisés.

En particulier, ce formalisme offre la possibilité de modéliser la sémantique vague de termes linguistiques, comme *proche*, défini sur l'univers des distances $\mathbb{R}^+$, *jeune*, défini sur l'univers des âges, ou plus généralement *faible, moyen, élevé* pour toute valeur numérique. De plus, l'interprétation de ces termes peut être précisée par l'utilisateur selon son expertise d'interprétation des variables. Ce formalisme permet donc une intégration aisée de connaissances propres à l'utilisateur, et par là la personnalisation des outils proposés. La représentation floue des variables numériques à l'aide de termes linguistiques offre ainsi une grande intelligibilité pour un utilisateur humain non spécialiste en IA [16]

De façon générale, la logique floue et la théorie des sous-ensembles flous constituent des outils naturels pour le domaine de l'IA explicable, mis en œuvre dans de nombreux cadres par l'équipe LFI et illustrés tour à tour ci-dessous.

Interprétabilité by design

De nombreux travaux de l'équipe LFI portent sur l'amélioration de l'interprétabilité des modèles d'apprentissage à la fois durant leur construction et dans leur utilisation pour

la classification.

La représentation floue des termes linguistiques permet de prendre en compte la gradualité des frontières, tout en limitant la complexité des modèles construits par apprentissage. Ces approches reposent sur l'extension des algorithmes d'apprentissage pour leur permettre de prendre en compte cette représentation floue, tout en respectant leur dimension numérique, ce qui les éloigne des approches strictement symboliques. Cela passe en premier lieu par une étude des mesures impliquées dans ces algorithmes et leur généralisation pour pouvoir les appliquer à des données floues [7], tout en conservant les propriétés intrinsèques de l'algorithme d'apprentissage. Le modèle d'étude de base dans ces travaux est, dans un cadre d'apprentissage supervisé, le modèle de construction de règles de décision, par exemple l'apprentissage d'arbres de décision flous [15, 17].

Interprétabilité post-hoc

Les approches d'interprétabilité *post-hoc* visent à proposer des explications intelligibles à tout système construit par apprentissage automatique, indépendamment de cette phase d'apprentissage. Elles peuvent être agnostiques, c'est-à-dire ne pas faire d'hypothèse sur le type de classifieur utilisé, ou au contraire exploiter des informations sur celui-ci.

Dans un cadre agnostique, les travaux menés dans l'équipe LFI se placent en particulier dans le domaine de la génération d'exemples contrefactuels [11], qui expliquent une prédiction en indiquant les modifications à apporter à la donnée considérée pour changer la classe prédite. Ces travaux, initiés en collaboration avec l'équipe R&D de Marcin DETYNIECKI du groupe AXA, se prolongent dans le cadre du *Trustworthy and Responsible AI Lab* (TRAIL), laboratoire de recherche commun à Sorbonne Université et AXA créé en décembre 2021.

Dans un cadre non-agnostique considérant l'apprentissage profond pour l'analyse d'images, il s'agit d'expliquer des résultats de classification ou segmentation. Les explications *post-hoc* exhibent les zones de l'image ou les caractéristiques qui contribuent le plus à la décision. Des travaux en imagerie biologique et médicale sont menés avec le LTCI / Télécom Paris, l'ISIR / Sorbonne Université, des équipes industrielles et plusieurs hôpitaux parisiens (par exemple [9, 18, 22]). Les recherches en cours portent sur la recherche d'explications non seulement locales, mais aussi structurelles, impliquant plusieurs objets dans les images et leur organisation spatiale.

Résumés linguistiques

La génération de textes à partir de données numériques ou catégorielles, tâche aussi appelée *data-to-text*, est une approche classique d'interprétabilité, les formulations linguistiques étant considérées comme faciles à comprendre par tout utilisateur. Les résumés linguistiques ajoutent l'objectif de fournir une vue synthétique, facilitant plus encore la compréhension du contenu des données. Les résumés par protoformes, introduits initialement par Yager, s'écrivent par exemple, dans leur forme basique, QRX sont P, où X représente les données à résumer, Q un quantificateur, comme la plupart ou quelques, et R et P des termes linguistiques correspondant à des propriétés d'intérêt. Q, P et R peuvent être représentés comme des variables linguistiques dans le formalisme des sous-ensembles flous, par exemple pour représenter un résumé tel que « la plupart des vols ayant un retard important sont des vols longs ».

Outre des questions d'efficacité calculatoire posées par l'explosion combinatoire de l'exploration des résumés possibles, étudiées en collaboration avec l'IRISA-Lannion [23], de nombreuses questions d'interprétabilité sont

néanmoins à soulever [13] : la formulation linguistique peut s'avérer être un piège en ce qu'elle semble intuitive, mais peut être ambiguë. En effet, il se peut que l'utilisateur n'ait pas la même compréhension de la phrase que le sens exprimé par la mesure de qualité implémentée, pour laquelle de multiples définitions peuvent être proposées [21]. Il est aussi possible que l'interprétation considérée des termes ne soit pas en adéquation avec la structure sous-jacente des données, ce qui peut conduire à des expressions linguistiques trompeuses [14].

Dans le cas d'expressions numériques approximatives, du type *environ* x où x est un nombre, une adéquation cognitive doit également être prise en compte, afin d'éviter le risque d'interprétation erronée. Des travaux menés en collaboration avec des psychologues cogniticiens de CHART-Paris VIII ont montré qu'elle doit tenir compte de la magnitude, du dernier chiffre significatif, de la granularité et de la complexité de x [12].

Au-delà des questions d'interprétation des phrases isolées, l'interprétabilité d'un résumé porte aussi sur les phrases dans leur ensemble, notamment pour tenir compte de relations de redondance, qui nuisent à l'intelligibilité, voire de contradictions à justifier [19, 20].

Approches hybrides

Le cadre de l'IA hybride vise à modéliser et utiliser à la fois des connaissances et des données en combinant plusieurs pans de l'IA. Des approches logiques permettent de représenter et raisonner sur des connaissances. Celles-ci sont ensuite transcrites dans des modèles computationnels structurels (graphes, hypergraphes, ontologies, treillis de concepts). Ces structures sont enrichies de modèles des imprécisions inhérentes aux descriptions linguistiques des connaissances dans la théorie des ensembles flous [4, 5] (par exemple « la structure A est à droite de la structure B »). Le problème du fossé sémantique entre les concepts abstraits et les domaines concrets des données est résolu là encore dans le cadre des ensembles flous à l'aide de la notion de variable linguistique présentée au début de l'article. Ces approches, appliquées à l'interprétation d'images, permettent de conserver le lien entre les données et les connaissances, et les connaissances utilisées pour une tâche de décision (reconnaissance d'objets par exemple) fournissent des explications des décisions (par exemple, les relations spatiales utilisées pour reconnaître des objets, confrontées aux connaissances a priori sur l'organisation spatiale des objets dans la scène observée). Ces résultats peuvent alors être exprimés sous forme de descriptions linguistiques du contenu des images.

Une réflexion en cours, en particulier avec des radiologues et des philosophes des sciences [10], porte sur les questions éthiques pour lesquelles les approches hybrides de l'explicabilité pourraient apporter des éclairages.

Interfaces explicatives

Au-delà de la génération d'explications, la construction d'interfaces permettant de les visualiser et de les manipuler s'avère une composante essentielle de leur adoption par les utilisateurs : de nombreux outils d'explications s'adressent à des spécialistes d'apprentissage automatique et d'intelligence artificielle, et non à des utilisateurs sans cette expertise. Des travaux menés en collaboration avec AXA et CHART-Paris VIII sur des interfaces de présentation d'explications de type instances contrefactuelles et vecteurs d'importance locale montent l'intérêt de la contextualisation et de l'interaction, à la fois pour la compréhension objective et la satisfaction subjective [8].

Approches symboliques

Des approches purement symboliques, en particulier logiques, sont également dévelop-

pées dans l'équipe, selon deux directions. Une première approche, en collaboration avec le MICS / CentraleSupelec, le LAMSADE / Université Paris Dauphine, le CRIL / Université d'Artois et l'ULA (Merida, Vénézuela), porte sur des méthodes d'abduction, où une observation est expliquée par une formule logique en fonction d'une base de connaissances. Des exemples concrets d'opérateurs d'abduction ont été proposés dans le cadre de la morphologie mathématique, d'abord en logique propositionnelle, puis dans un cadre plus général englobant, entre autres, la logique floue [1, 2, 3].

Une deuxième direction, inspirée des travaux de Halpern et Miller, repose sur des arbres causaux. Des travaux en cours portent sur les liens entre les modèles structurels causaux et les systèmes d'argumentation abstraits (avec l'ISIR / Sorbonne Université), ainsi que sur une formalisation floue des explications par contraste dans le cadre des modèles structurels causaux [6].

Références

[1] M. Aiguier, J. Atif, I. Bloch, and R. Pino Pérez. Explanatory relations in arbitrary logics based on satisfaction systems, cutting and retraction. *International Journal of Approximate Reasoning*, 102 :1–20, 2018.

[2] M. Aiguier and I. Bloch. Logical dual concepts based on mathematical morphology in stratified institutions. *Journal of Applied Non-Classical Logics*, 29(4) :392–429, 2019.

[3] J. Atif, C. Hudelot, and I. Bloch. Explanatory reasoning for image understanding using formal concept analysis and description logics. *IEEE Transactions on Systems, Man and Cybernetics : Systems*, 44(5) :552–570, May 2014.

[4] I. Bloch. Fuzzy sets for image processing and understanding. *Fuzzy Sets and Systems*, 281 :280–291, 2015.

[5] I. Bloch. Mathematical morphology and spatial reasoning : Fuzzy and bipolar setting. *TWMS Journal of Pure and Applied Mathematics – Special Issue on Fuzzy Sets in Dealing with Imprecision and Uncertainty : Past and Future Dedicated to the memory of Lotfi A. Zadeh*, 12(1) :104–125, 2021.

[6] I. Bloch and M.-J. Lesot. Vers une formulation floue des explications par contraste. In *Rencontres Francophones sur la Logique Floue et ses Applications*, pages 191–198, 2021.

[7] B. Bouchon-Meunier and C. Marsala. Entropy and monotonicity in artificial intelligence. *International Journal of Approximate Reasoning*, 124 :111–122, 2020.

[8] C. Bove, J. Aigrain, M.-J. Lesot, C. Tijus, and M. Detyniecki. Contextualization and exploration of local feature importance as explanations to improve understanding and satisfaction of non-expert users. In *International Conference on Intelligent User Interfaces*, 2022.

[9] V. Couteaux, S. Si-Mohamed, O. Nempont, T. Lefevre, A. Popoff, G. Pizaine, N. Villain, I. Bloch, A. Cotten, and L. Boussel. Automatic knee meniscus tear detection and orientation classification with Mask-RCNN. *Diagnostic and Interventional Imaging*, 100 :235–242, 2019.

[10] V. Israël-Jost et al. L'éthique en radiologie : quand, comment ? premiers éléments. *Journal d'Imagerie Diagnostique et Interventionnelle*, 4 :238–240, 2021.

[11] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki. The dangers of post-hoc interpretability : Unjusti-

fied counterfactual explanations. In *28th International Joint Conference on Artificial Intelligence*, pages 2801–2807, 2019.

[12] S. Lefort, M.-J. Lesot, E. Zibetti, C. Tijus, and M. Detyniecki. Interpretation of approximate numerical expressions : Computational model and empirical study. *International Journal on Approximate Reasoning*, 82 :193–209, 2017.

[13] M.-J. Lesot, G. Moyse, and B. Bouchon-Meunier. Interpretability of fuzzy linguistic summaries. *Fuzzy Sets and Systems*, 292(1) :307–317, 2016.

[14] M.-J. Lesot, G. Smits, and O. Pivert. Adequacy of a user-defined vocabulary to the data structure. In *International Conference on Fuzzy Systems*, 2013.

[15] C. Marsala. Fuzzy decision trees for dynamic data. In *IEEE Symposium on Evolving and Adaptive Intelligent Systems*, pages 17–24, 2013.

[16] C. Marsala and B. Bouchon-Meunier. Fuzzy data mining and management of interpretable and subjective information. *Fuzzy Sets and Systems*, 281 :252–259, 2015.

[17] C. Marsala and D. Petturiti. Rank discrimination measures for enforcing monotonicity in decision tree induction. *Information Sciences*, 291 :143–171, 2015.

[18] G. Martin, S. El-Madafri, A. Becq, J. Szewczyk, and I. Bloch. Instruments Segmentation in X-ray Fluoroscopic Images for Endoscopic Retrograde Cholangio Pancreatography. In *Medical Informatics Europe*, 2022.

[19] G. Moyse, M.-J. Lesot, and B. Bouchon-Meunier. Oppositions in fuzzy linguistic summaries. In *International Conference on Fuzzy Systems*, 2015.

[20] A. Oudni, M.-J. Lesot, and M. Rifqi. Processing contradiction in gradual itemset extraction. In *International Conference on Fuzzy Systems*, 2013.

[21] A. Oudni, M.-J. Lesot, and M. Rifqi. Accelerating effect of attribute variations : accelerated gradual itemsets extraction. In *International Conference on Information Processing and Management of Uncertainty*, pages 395–404. Springer, 2014.

[22] A. Pirovano, L. G. Almeida, S. Ladjal, I. Bloch, and S. Berlemont. Computer-aided diagnosis tool for cervical cancer screening with weakly supervised localization and detection of abnormalities using adaptable and explainable classifier. *Medical Image Analysis*, 73 :102167, 2021.

[23] G. Smits, P. Nerzic, O. Pivert, and M.-J. Lesot. Frels : Fast and reliable estimated linguistic summaries. In *International Conference on Fuzzy Systems*, 2021.

AfIA
Association française
pour l'Intelligence Artificielle

Recherches en IA explicable à Orange

Par **Françoise FESSANT**
Emilie SIRVENT-HIEN
Moustapha ZOUINAR
Orange
{prénom}.{nom}@orange.com
https://hellofuture.orange.com

Introduction

L'Intelligence Artificielle (IA) amène chaque jour davantage de valeur dans de nombreux métiers pour Orange tels que la relation client, le marketing, le réseau et les interventions, la logistique, le SI et la sécurité et bien d'autres fonctions supports. C'est pourquoi Orange a placé la data et l'IA au cœur de son modèle d'innovation avec l'ambition de faire preuve d'exemplarité sociale et environnementale. Dans ce contexte Orange est engagé dans différents types d'actions sur le sujet de l'IA éthique et responsable, que ce soit en externe avec des contributions à différents collectifs et *think tanks* : Impact AI, Cercle InterElles, Digital Society Forum, Renaissance Numérique, etc. mais également en interne avec la création d'un conseil d'éthique de la *data* et de l'IA qui a pour mission d'accompagner la mise en œuvre par l'entreprise de principes éthiques encadrant l'utilisation des technologies de *data* et d'IA, ou encore avec la création d'une direction dédiée à la *data* et l'IA, et la mise en œuvre d'actions de formation et d'acculturation à l'IA.

L'implication des programmes de recherche

Au niveau de la recherche d'Orange, plusieurs programmes de recherche sont impliqués sur la thématique de l'IA éthique et responsable. Les travaux sont multidisciplinaires et concernent aussi bien l'étude des enjeux éthiques et juridiques autour de l'IA, la mesure et la minimisation de l'impact environnemental des solutions à base d'IA, que la construction d'algorithmes pour lutter contre les cyberattaques ou les mécanismes de protection des données personnelles.

Les travaux en IA explicable

Parmi les nombreuses questions de recherche qui sont traitées, il y a celle de la conception et de la construction de solutions à base d'IA transparente et explicable, et nous explorons différentes thématiques autour de l'outillage et la conception d'algorithmes en collaboration avec différents partenaires :

- Dans le cadre d'une prestation avec Quantmetry, nous avons travaillé à recenser et tester les outils open source pour l'IA Responsable incluant notamment les thématiques de la transparence, la traçabilité (par exemple Model cards, Ethical AI Standard, AI factsheets 360) ou l'explicabilité (par exemple Lime, Shap ou Shapash) de manière à pouvoir faire des recommandations au groupe Orange, tout en maîtrisant les limites actuelles [2].
- Dans le cadre d'une thèse CIFRE avec l'IRISA débutée fin 2020, on s'intéresse aux approches d'interprétabilité *post-hoc* locales, c'est-à-dire qui s'appliquent après l'apprentissage d'un modèle de classification, et en particulier aux explications contrefactuelles. Une explication contrefac-

tuelle se présente sous la forme d'une version modifiée de l'exemple à expliquer qui répond à la question : que faudrait-il changer pour obtenir une prédiction différente ? La plupart des méthodes d'explications contrefactuelles sont basées sur la perturbation de l'instance originale grâce à l'optimisation d'une fonction de coût. Selon les propriétés souhaitées pour l'explication, on rajoute des contraintes dans le processus d'optimisation sous la forme de termes supplémentaires dans la fonction de coût. Sur cette thématique, nous avons proposé un algorithme de génération d'explications contrefactuelles qui introduit dans la fonction de coût un terme basé sur un auto-encodeur supervisé. La supervision de l'auto-encodeur permet d'améliorer le réalisme des contrefactuels (*i.e.* fidèles à la distribution des données de la classe cible) [1].

- Un autre axe, tout aussi important que celui de la mise au point d'outils ou d'algorithmiques, concerne la manière dont il faut produire les explications à destination de personnes qui ne sont pas des experts en IA ou en science des données, dans la mesure où de telles explications peuvent être utiles, voire nécessaires. Par exemple, il peut être utile pour un agent d'une banque de comprendre pourquoi un système d'IA de détection de fraudes bancaires conclut que telle ou telle transaction bancaire caractérise une fraude. Cet axe pose ainsi des questions de méthodologie, de conception d'explications intelligibles et pertinentes pour ce type de personnes, d'interface utilisateur, d'interaction homme-machine et d'évaluations des explications générées. Dans le cadre du programme de recherche « IA responsable », des travaux sont ainsi menés sur ces différentes questions. Une stratégie sur laquelle repose ces travaux consiste à s'appuyer sur la conception centrée utilisateur, qui constitue une approche méthodologique éprouvée de conception des systèmes interactifs [3].

Nous suivons la standardisation en cours de construction sur l'IA responsable (et y participons également). Le standard IEEE P7001 travaille par exemple la transparence des systèmes autonomes avec différents niveaux de transparence pour les différentes parties prenantes concernées par le système. L'ISO s'intéresse également à la transparence (ISO/IEC AWI 12792) et à l'explicabilité de l'apprentissage machine (ISO/IEC TS 6254 AWI). Ces standards auront une influence importante sur l'applicabilité de la réglementation européenne en préparation.

Références

[1] V. Guyomard, F. Fessant, T. Bouadi, and Th. Guyet. Générer des explications contrefactuelles à l'aide d'un autoencodeur supervisé. *Revue des Nouvelles Technologies de l'Information*, Extraction et Gestion des Connaissances, RNTI-E-38 :111–122, 2022.

[2] R. Poyiadzi, X. Renard, Th. Laugel, R. Santos-Rodríguez, and M. Detyniecki. On the overlooked issue of defining explanation objectives for local-surrogate explainers. *CoRR*, abs/2106.05810, 2021.

[3] M. Zouinar. Évolutions de l'intelligence artificielle : quels enjeux pour l'activité humaine et la relation humain-machine au travail ? *Activités*, (17-1), 2020.

AfIA
Association française
pour l'Intelligence Artificielle

Recherches en IA explicable au sein du département IA de l'IRIT

Par

Pascal ZARATE
IRIT / ADRIA
Université Toulouse 1 Capitole
pascal.zarate@irit.fr
www.irit.fr/departement/intelligence-artificielle/adria/

Nathalie AUSSENAC
IRIT / MELODI
CNRS
nathalie.aussenac@irit.fr
www.irit.fr/departement/intelligence-artificielle/melodi/

The IRIT-Artificial Intelligence Department investigates the automation of reasoning and decision-making processes, based on knowledge drawn from texts and data, but also at defining natural language analysis systems, with a view to helping humans. This research addresses the following issues :

- Automated reasoning, especially under uncertainty and probabilistic reasoning ;
- Symbolic and statistical machine learning ;
- Decision support systems for an individual or a group of decision makers and automated decision processes ;
- The formalization of interaction and communication between agents, in particular the role of beliefs and the management of arguments ;
- The security of information and communication systems ;
- Models and methods for natural language processing, natural language semantics and discourse analysis ;
- Knowledge engineering and formal ontology, from knowledge extraction, its modelling and its formal representation, its linking within the semantic web and the web of data, and the study of its evolution.

The AI department is composed by three teams : ADRIA, LILaC and MELODI. The interactions among the 3 teams are important, with several co-supervised PhD theses and joint projects.

Explainability is addressed by a lot of researchers using different approaches :

- Formal Explainability (cf Marques-Silva and co),
- Analogical explanations (cf Prade and Richard),
- Abstract argumentation (cf Duchatelle et al.),
- Formal Reasoning for Reinforcement learning (cf Saulières et al.),
- Explainable AI for Intrusion detection (cf Chevalier),
- Interacting a machine Learning system with an explicit reasoning system : Application on medical data (cf Mayouf et al.).

Formal explainability

Joao Marques-Silva, Martin Cooper, Xuanxiang Huang, Yacine Izza, Nicholas Asher

Since 2019, our team has been investigating formal approaches to explainability in machine learning (ML), which we refer to as Formal Explainable AI (FXAI). In contrast to most of the existing work on explainability in ML, we have proposed definitions of explana-

tions that are rigorous, that take into account the underlying ML model, and that are amenable to exact computation using automated reasoners. The team currently includes João MARQUES-SILVA (CNRS DR and ANITI Research Chair), Martin COOPER (UPS Professor and ANITI Co-Chair), Yacine IZZA (Postdoctoral researcher, ANITI and IRIT), Xuanxiang HUANG (PhD student, ANITI and IRIT), Thomas GERSPACHER (former PhD student, ANITI and IRIT), and Nicholas ASHER (CNRS DR, and ANITI Scientific Director). The initial ideas on formal explainability we presented in the following papers : [10], [11] and [9]. A recent overview of the progress in formal approaches to explainability is given in [16].

Furthermore, we have demonstrated a number of results, organized as follows :

1. Tractable explainability : We have shown that, for several well-known families of classifiers, the computation of one explanation is poly-time. This is the case of Naive Bayes Classifiers (see [14]), monotonic classifiers (see [15]), decision trees and other graph-based classifiers (see [6]), and several families of propositional languages (see [24]). The tractability of several other families of classifiers is investigated in [4].
2. Connections between fairness and explainability : some initial results were presented in [7] and more recently in [2].
3. Duality of explanations : two kinds of minimal-hitting set duality relationships were identified (see [11] and [9]).
4. Practical efficient explainability : We have shown that for decision lists and sets and for tree ensembles, the computation of one explanation has been shown to be computationally hard for decision lists and sets (see [12]), random forests (see [13]) and tree ensembles in general (see [8]). However, we also developed logic encodings that enable the efficient practical computation of explanations.
5. Assessment of model-agnostic explainers : our results demonstrate the inadequacy of well-known model-agnostic explainers in settings where the rigor of explanations is paramount (see [20]).
6. Improvements to model-agnostic explainers (see [1]).
7. Trade-offs between rigor of explanations and their size : ongoing work.

Analogical explanations

Henri Prade, Gilles Richard

The approach [21] relies on the use of analogical proportions (AP), which are statements relating four items, of the form "a is to b as c is to d". The items are represented by vectors of Boolean or categorical attribute values. a, b, c, d make a valid AP, if the attributes can be split into three subsets $\mathcal{A}, \mathcal{A}', \mathcal{A}''$ (some may be empty), in such a way that a, b, c, d are identical on $\mathcal{A}$, $a = b$ and $c = d$ on $\mathcal{A}'$, while on $\mathcal{A}''$ the same change of values takes place from a to b, and from c to d. It is pictured in the table below, where s, t, u, v, w are sub-vectors of attribute values. The change of class from x to y in pair (a, b) can be explained only by the change of values of attributes in $\mathcal{A}''$. The same change for pair (c, d) has the same effect for the classes. Thus, this provides a basis for predicting or for explaining why d is in class y. Each pair may be viewed as a potential rule expressing that in a context (described by values on $\mathcal{A} \cup \mathcal{A}'$) the change from v to w induces the flip from class x to class y. The Confidence in the rule can be evaluated on the set of examples at hand. As can be seen, the approach does not require to know how the class of d has been obtained for explaining it.

	$\mathcal{A}$ full id.	$\mathcal{A}'$ pair id.	$\mathcal{A}''$ change	class
a	s	t	v	x
b	s	t	w	y
c	s	u	v	x
d	s	u	w	?

A Query-based Explanation Model for Abstract Argumentation

Théo Duchatelle, Philippe Besnard, Sylvie Doutre, Marie-Christine Lagasquie

Abstract Argumentation [5] is a rising formalism for computing explanations [22]. An approach to explain this formalism itself is pictured in Figure 1.1.

The approach includes a formal grammar for modelling the questions the user can ask, and a process for building the answers which uses graph operations and which exploits elements of the question.

FR4RL : Formal Reasoning for Reinforcement Learning

Leo Saulieres, Martin Cooper, Florence Dupin de Saint-Cyr, Joao Marques-Silva

The PhD started in October 2021. The proposed research project is positioned at the intersection of automated reasoning (AR) and Reinforcement Learning. It aims to develop novel solutions for logic-enabled reasoning about RL-enabled ML systems. Concretely, the PhD research project is broadly organized into three main vectors :

1. First, to conduct an in-depth review of existing heuristic approaches for reasoning about RL, and to identify possible limitations of existing approaches.
2. Second, to develop a deep understanding of the work of the DeepLever team which has been working for several years on the other branches of ML including Neural Networks and statistical computational learning, on computing rigorous explanations.
3. Third, to develop formal tools for reasoning about Markov Decision Processes (MDPs), namely :
   (a) Generalize prime implicants of decision functions to the case of MDPs. One approach to investigate will be quantified functions representing strategies, similarly to what is common practice when solving quantified problems ;
   (b) Propose algorithms for computing logical formulations of MDPs behaviors ; and
   (c) Understand the practical limitations of computing compact logical formulations of MDPs, as well as the reasons behind their operation.

The first ideas are being tested on 2-person games and on examples of multi-agent path finding.

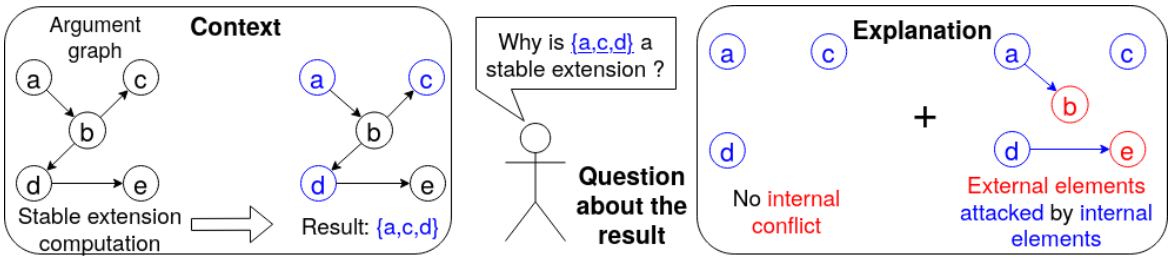


Figure 1.1 – An overview of computing explanations in abstract argumentation

Explainable AI for Intrusion Detection

Yannick Chevalier

Reflecting upon the usage of AI methods in the field of Intrusion Detection [23], Sommer and Paxson pointed out the gap between what can be offered by AI techniques for prediction and classification, and what is needed for effective intrusion detection. Among other, we shall name the need to build a system that distinguishes between anomalies and intrusions in a system, takes external descriptions of normal and intrusion behaviours into account and is able to explain its decisions to a human for further processing.

These considerations resonate with those expressed in [19] to define how a usable AI-based computer system should interact with a user, though with an emphasis on the system being an Advice Giver, to explain its decision, as much as an Advice Taker, to input external descriptions.

We built an intrusion detection system for simple networks in which the output of the learning is a set of first-order logic atoms that have to be satisfied by normal traffic [3]. This system is currently being expanded to prepare a background first-order logic theory that describes normal behaviours, and to construct abstract formulas describing the output of the learning phase.

Interacting a machine Learning system with an explicit reasoning system : Application on medical data

Mouna Sabrine Mayouf, Florence Dupin de Saint-Cyr

The PhD is about making interact a machine learning system with an explicit reasoning system for an application on medical data. This PhD started in December 2019.

A first research project has examined methodological aspects of the training procedure of neural networks in the context of a medical image classification problem. We have proposed a formalization of the data preparation. The formalism has allowed us to prove a number of useful properties of the training dataset used in the experiments, which in turn enhanced fairness of comparison and research transparency.

The second research question is concerning the conjecture that is, feeding a network with datasets of increasing magnification leverages high-level knowledge and helps the network to better classify. This hypothesis was confirmed by an experiment carried out on a dataset of breast cancer histopathological images. Results underline the importance of the order in which data is introduced to the neural network during the training phase. Extensive experiments done on the BreakHis dataset demonstrate that curriculum incremental learning reaches 98.76% accuracy for binary classification, while the best state-of-the-art approach only reaches 96.78%.

Concerning multi-class classification, curriculum incremental learning reaches 95.93% while the state-of-the-art approaches only reaches 95.49%. Also, both the computational time and the stabilization time of the learning process of the incremental curriculum learning approach are reduced (respectively by 6% and by more than 20%) as compared to a non curriculum learning approach.

We are currently working on a new way to use hierarchical constraints in order to guide the machine learning process. A first article has been accepted at the conference CAP'2021 [18] and a second article is under review for publication in an international journal [17].

Références

[1] A. Ignatiev Kuldeep S. Meel J. Marques-Silva M. Y. Vardi Aditya A. Shrotri, Nina Narodytska. Constraint-driven expla-

nations for black box ml models. In *Proc. of AAAI*, 2022.

[2] N. Asher, S. Paul, and Ch. Russell. Fair and Adequate Explanations. In *5th International Cross-Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE 2021)*, volume 12844 of *LNCS*, Vienna (virtual), Austria, August 2021.

[3] Y. Chevalier. Data exchange for anomaly detection : The case of the can bus. In *Proceedings of the Conference on Artificial Intelligence for Defence*, 2021.

[4] M. C. Cooper and J. Marques-Silva. On the tractability of explaining decisions of classifiers. In *27th Int. Conf. on Principles and Practice of Constraint Programming, CP*, volume 210, 2021.

[5] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2) :321–357, 1995.

[6] X. Huang, Y. Izza, A. Ignatiev, and J. Marques-Silva. On efficiently explaining graph-based classifiers. In *Proc. of KR*, pages 356–367, 2021.

[7] A. Ignatiev, M. C. Cooper, M. Siala, E. Hebrard, and J. Marques-Silva. Towards formal fairness in machine learning. In *Proc. of the 26th Int. Conf. on Principles and Practice of Constraint Programming (CP)*, volume 12333 of *LNCS*. Springer, 2020.

[8] A. Ignatiev, Y. Izza, P. J. Stuckey, and J. Marques-Silva. Using maxSAT for efficient explanations of tree ensembles. In *Proc. of AAAI*, 2022.

[9] A. Ignatiev, N. Narodytska, N. Asher, and J. Marques Silva. From Contrastive to Abductive Explanations and Back Again.

[10] A. Ignatiev, N. Narodytska, and J. Marques-Silva. Abduction-based explanations for machine learning models. In *Proc. of AAAI*, 2019.

[11] A. Ignatiev, N. Narodytska, and J. Marques-Silva. On relating explanations and adversarial examples. In *Proc. of NeurIPS*, 2019.

[12] A. Ignatiev and J. P. Marques Silva. SAT-based rigorous explanations for decision lists. In *Proc. of the 24th Int. Conf. on Theory and Applications of Satisfiability Testing (SAT)*, volume 12831 of *LNCS*, pages 251–269, 2021.

[13] Y. Izza and J. Marques-Silva. On explaining random forests with SAT. In *Proc. of the Thirtieth Int. Joint Conference on Artificial Intelligence, IJCAI*, pages 2584–2591. ijcai.org, 2021.

[14] J. Marques-Silva, Th. Gerspacher, M. C. Cooper, A. Ignatiev, and N. Narodytska. Explaining naive bayes and other linear classifiers with polynomial time and delay. In *Proc. of NeurIPS*, 2020.

[15] J. Marques-Silva, Th. Gerspacher, M. C. Cooper, A. Ignatiev, and N. Narodytska. Explanations for monotonic classifiers. In *Proc. of the 38th Int. Conference on Machine Learning, ICML*, volume 139, pages 7469–7479, 2021.

[16] J. Marques-Silva and A. Ignatiev. Delivering trustworthy ai through formal XAI. In *Proc. of AAAI*, pages 3806–3814, 2022.

[17] M. Sabrine Mayouf and F. Dupin De Saint Cyr Bannay. Formalizing data preparation in curriculum incremental deep learning on breakhis dataset (revised version submitted to neurocomputing). 2021.

[18] M. Sabrine Mayouf and F. Dupin De Saint Cyr Bannay. Préparation efficace des données d'apprentissage. Application à la

classification d'images pour la détection du cancer du sein. In *Conférence sur l'Apprentissage Automatique (CAp 2021)*, Saint-Étienne (virtuel), France, 2021.

[19] J. McCarthy. Situations, actions, and causal laws. Technical Report TR AIM-002, Stanford University.

[20] N. Narodytska, A. A. Shrotri, K. S. Meel, A. Ignatiev, and J. Marques-Silva. Assessing heuristic machine learning explanations with model counting. In *Theory and Applications of Satisfiability Testing (SAT)*, volume 11628 of *LNCS*, pages 267–278. Springer, 2019.

[21] H. Prade and G. Richard. Explications analogiques. In *Workshop EXPLAIN'AI'22 @ EGC conf., Blois*, 2022.

[22] A. Rago, O. Cocarascu, C. Bechlivanidis, and F. Toni. Argumentation as a framework for interactive explanations for recommendations. In *Proc. of KR*, pages 805–815, 2020.

[23] R. Sommer and V. Paxson. Outside the closed world : On using machine learning for network intrusion detection. In *2010 IEEE Symposium on Security and Privacy*, 2010.

[24] A. Ignatiev M. Cooper N. Asher X. Huan, Y. Izza and J. Marques-Silva. Tractable explanations for d-DNNF classifiers. In *Proc. of AAAI*, 2022.

AfIA
Association française
pour l'Intelligence Artificielle

Recherches en IA explicable au MICS : Modèles gaussiens, modèles génératifs et raisonnement pour l'explicabilité

Par
Wassila OUERDANE
Manuel AMOUSSOU
Martin CHARACHON
Paul-Henry COURNÈDE
Ludovic GOUDENÈGE
Céline HUDELOT
Mathieu LEROUGE
Vincent MOUSSEAU
MICS / BioMathematics, LOGIMICS, MathRisk
Université Paris Saclay
{prénom}.{nom}@centralesupelec.fr
http://www.mics.centralesupelec.fr

Introduction

Le laboratoire de Mathématiques et Informatique pour la Complexité et les Systèmes (MICS) de CentraleSupélec s'intéresse à l'analyse mathématique et informatique des systèmes et données complexes, qu'ils proviennent du vivant, de l'industrie, des sciences sociales, de l'information ou des réseaux. L'axe Intelligence Artificielle, transverse aux activités de recherche du laboratoire, héberge une variété de travaux autour de la thématique de *l'IA Explicable*. Ces travaux visent à produire des modèles et outils formels pour soutenir des décisions, recommandations, issues d'algorithmes et des mécanismes d'apprentissage automatique, de la théorie de la décision ou du choix social. De plus, le laboratoire bénéficie d'un écosystème riche de collaborations avec divers acteurs socio-économiques. Ces collaborations offrent l'opportunité d'aller, quand c'est possible, vers des propositions opérationnalisables. Nous présentons dans ce qui suit un échantillon des travaux du MICS pour la thématique de l'IA explicable en mettant en relief des méthodes différentes, mais aussi des terrains d'application variés.

Extrapolation par noyau gaussien en grande dimension

Dans le cadre des applications en finance et en dynamique moléculaire, on cherche à modéliser des phénomènes de grandes dimensions (millions de paramètres) avec quelques hyperparamètres. En effet, à l'aide d'algorithmes d'apprentissages statistiques, on est en mesure d'entraîner un modèle réduit qui va reproduire les caractéristiques essentielles. En dynamique moléculaire, cela peut correspondre à des résultats de réactions chimiques, des changements de conformation de molécules (très étudiés en biochimie) ou des transitions de phases, dans les systèmes de matières condensées, impliquées notamment dans la construction des batteries [15]. En mathématiques financières, il s'agit de reproduire les prix du marchés sous l'influence de multiples agents ou facteurs de

risques [5]. Ces modèles réduits permettent ensuite de réaliser un grand nombre de simulations numériques, mais le but est surtout de comprendre le sens physique ou financier derrière ces hyperparamètres. Chez les chimistes et physiciens, on cherche des paramètres avec des significations théoriques, comme les distances inter-atomiques ou les angles dihédraux dans une molécule. Rassemblés sous l'appellation de variables collectives, on cherche celles avec le meilleur taux d'explicabilité des dynamiques afin de trouver de nouvelles modélisations ou lois physiques. En finance, on cherche à comprendre l'influence des stratégies et des facteurs de risques sur le marché.

Classification et annotation explicable par apprentissage de relations et raisonnement

Avec les succès récents de l'apprentissage profond et les interactions toujours plus nombreuses entre êtres humains et intelligences artificielles, l'explicabilité est devenue une préoccupation majeure. En effet, il est difficile de comprendre le comportement des réseaux de neurones profonds, ce qui les rend inadaptés à une utilisation dans les systèmes critiques. Pour pallier cela, [13, 12] proposent une approche visant à classifier ou annoter des signaux tout en expliquant les résultats obtenus. Elle est basée sur l'utilisation d'un modèle transparent, dont le raisonnement est clair, et de relations floues interprétables qui permettent de représenter l'imprécision du langage naturel. Au lieu d'apprendre sur des exemples sur lesquels les relations ont été annotées, nous proposons de définir un ensemble de relations au préalable. L'évaluation de ces relations sur les exemples de la base d'entraînement est accélérée grâce à deux heuristiques [11]. Ensuite, les relations les plus pertinentes sont extraites en utilisant un nouvel algorithme de *frequent itemset mining* flou. Ces relations permettent de construire des règles pour la classification ou des contraintes pour l'annotation. Ainsi, une explication en langage naturel peut être générée. Nous présentons des expériences sur des images et des séries temporelles afin de montrer la généricité de notre approche. En particulier, son application à l'annotation d'organe explicable a été bien évaluée par un ensemble de participants qui ont jugé les explications convaincantes et cohérentes [14]

Exploitation des modèles génératifs conditionnels pour l'explication de classifieurs en imagerie médicale

En imagerie médicale, les résultats d'analyses obtenus par apprentissage profond peuvent atteindre une précision proche, voire supérieure à celle des radiologues pour certaines problématiques. Toutefois, ces modèles restent des boîtes noires et peuvent commettre de grossières erreurs lorsqu'ils sont mis en production. Dans ce travail, une attention est portée sur l'interprétabilité et l'explicabilité de ces modèles en imagerie médicale. Nous considérons qu'une explication visuelle d'un classifieur peut être produite comme la différence entre deux images générées obtenues via deux modèles génératifs conditionnels spécifiques [3, 4]. Les deux modèles génératifs sont entraînés en utilisant le modèle à expliquer et une base de données de sorte que : (i) les images générées par le premier générateur sont classées de manière similaire à l'image d'entrée, tandis que les sorties du second générateur sont classées de manière opposée ; (ii) toutes les images générées appartiennent à la distribution des images réelles ; (iii) les distances entre l'image d'entrée et les images générées correspondantes sont minimales de sorte que la différence entre les éléments générés ne révèle que des informations pertinentes pour le classificateur étudié.

Schémas d'arguments pour l'aide à la décision multicritère

La capacité de fournir à un utilisateur des explications accompagnant les recommandations est une caractéristique essentielle des outils d'aide à la décision. Nos travaux portent sur « la mise en place d'outils et d'algorithmes d'explications pour des recommandations issues de modèles multicritère » qui mettent au cœur du raisonnement les préférences et les jugements des utilisateurs. Plus précisément, l'aide multicritère à la décision vise à développer des modèles de décision explicitement basés sur la construction d'un ensemble de critères reflétant les aspects pertinents du problème de prise de décision. Dans ce cadre, un modèle très largement utilisé, que ce soit en théorie de la décision ou en apprentissage automatique, à savoir le modèle additif, est étudié. Il s'agit de produire des explications pour de la comparaison par paire (avec de l'information complète [7] ou incomplète [2]). L'approche adoptée est dite *step-wise* et vise à fournir des explications sous la forme d'une séquence d'énoncés de préférences. Chaque énoncé doit être aussi significatif, pertinent et cognitivement simple que possible pour que l'explication soit acceptée [2]. Nous formalisons les explications au travers de schémas d'arguments, qui lient des prémisses (informations fournies ou approuvées par l'utilisateur, ou déduites au cours du processus d'apprentissage des préférences, et quelques hypothèses supplémentaires sur le processus de raisonnement (des hypothèses du modèle) à une conclusion (la décision). Divers schémas sont identifiés, permettant de dériver de nouvelles connaissances, sous forme d'énoncés comparatifs, à partir d'énoncés précédemment acceptés. Ces schémas exploitent un certain nombre de propriétés du modèle additif. Pour le calcul des explications, nous faisons appel par exemple à des outils de la programmation mathématique [1].

Explication de solutions de systèmes d'optimisation : application au WSRP

Les systèmes d'optimisation, tels que ceux résolvant le *Workforce Scheduling and Routing Problem* (WSRP), sont généralement perçus par ceux qui les utilisent comme des boîtes noires produisant des solutions qui sont en principe optimales et qu'il convient donc de suivre. Cependant, il arrive parfois que ces utilisateurs s'interrogent sur la pertinence des solutions, auquel cas en l'absence d'éléments tangibles éclairant ces solutions, cela peut être compliqué pour eux de prendre leurs décisions en se basant sur celles-ci. Une approche est alors d'expliquer les solutions obtenues par ces systèmes d'optimisation. Dans nos travaux, nous proposons une approche permettant de répondre en temps réel à un certain nombre de questions (locales, restreintes et contrastives) qu'un utilisateur peut poser un utilisateur [9]. Pour le calcul des explications, nous avons recours à des algorithmes polynomiaux usant d'outils issus de la recherche locale ou à de la programmation linéaire en nombres entiers appliquée à des problèmes de petites tailles. Afin d'être intelligible pour l'utilisateur, l'explication prend la forme d'un texte concis, écrit dans un vocabulaire haut-niveau, ainsi que de graphiques pour la visualisation.

Explications et recommandations interactives

Alors que les méthodes d'élicitation incrémentale impliquent déjà un processus d'interaction assez simple par lequel le système pose des questions à l'utilisateur, de nouveaux défis se présentent lorsqu'on veut intégrer des explications. En effet, pour produire une recommandation, le système interroge l'utilisateur pour obtenir ses préférences et les adapter à un modèle. Sur la base de ces préférences, le système peut produire une recommandation. Cependant, comme la recomman-

dation elle-même peut être très large (pensez à un classement impliquant toutes les options), il est utile de permettre que des recommandations incrémentales partielles et/ou factorisées soient faites tout au long de l'interaction, sur lesquelles le système cherchera à obtenir l'accord de l'utilisateur (par exemple « sommes-nous d'accord pour dire que le produit p est meilleur que tout produit dont la couleur est rouge ? » ou « sommes-nous d'accord pour dire que le sous-ensemble d'options p_1, p_2, p_3 ne devrait pas être considéré comme le produit de choix ? »). Lorsque le système propose la recommandation, l'utilisateur peut la critiquer (les préférences peuvent être ajustées, corrigées, l'option peut ne pas être réalisable, ou ne plus être disponible, etc.) ou demander une justification, qui doit être fournie par le système. Par conséquent, le système doit traiter le *problème de révision* inhérent induit par les déclarations éventuellement incohérentes (soit entre elles, soit avec le modèle de préférences supposé de l'utilisateur) [10].

Alors que les systèmes actuels équipés de fonctions d'explication produisent généralement une justification à la toute fin du processus - en même temps que leur recommandation finale - nous pensons qu'un système à initiative mixte - [6] où l'élicitation, la recommandation et l'explication sont étroitement imbriquées, est nécessaire. Cela implique de concevoir soigneusement un protocole qui décide exactement comment et quand l'initiative doit être donnée à l'utilisateur, ou conservée par le système, et comment les différents engagements peuvent être acceptés ou contestés. Dans cette perspective, un premier pas vers la formalisation d'une telle discussion est le travail de [8], où un jeu de dialogue est proposé pour formaliser l'interaction représentant une situation d'aide à la décision, impliquant l'échange de différents types d'informations préférentielles, ainsi que d'autres locutions comme la justification.

Références

[1] M. Amoussou, K. Belahcene, Ch. Labreuche, N. Maudet, V. Mousseau, and W. Ouerdane. Explaining Robust Additive Decision Models : Generation of Mixed Preference-Swaps by Using MILP. In *From Multiple Criteria Decision Aid to Preference Learning*, Trento (virtual), Italy, 2020.

[2] Kh. Belahcene, Ch. Labreuche, N. Maudet, V. Mousseau, and W. Ouerdane. Explaining robust additive utility models by sequences of preference swaps. *Theory and Decision*, 82(2) :151–183, 2017.

[3] M. Charachon, P-H. Cournède, C. Hudelot, and R. Ardon. Visual explanation by unifying adversarial generation and feature importance attributions. In *iMIMIC/TDA4MedicalData@MICCAI*, pages 44–55, 2021.

[4] M. Charachon, C. Hudelot, P-H Cournède, C. Ruppli, and R. Ardon. Combining similarity and adversarial learning to generate visual explanation : Application to medical image classification. In *Proceedings of the 25th ICPR*, pages 7188–7195, 2020.

[5] L. Goudenège, A. Molent, and A. Zanette. Machine learning for pricing american options in high-dimensional markovian and non-markovian models. *Quantitative Finance*, 20(4) :573–591, 2020.

[6] E. Horvitz. Uncertainty, action, and interaction : In pursuit of mixed-initiative computing. *Intelligent Systems*, pages 17–20, 2000.

[7] Ch. Labreuche, N. Maudet, and W. Ouerdane. Minimal and complete explanations for critical multi-attribute decisions. In *ADT*, pages 121–134, 2011.

[8] Ch. Labreuche, N. Maudet, W. Ouerdane, and S. Parsons. A dialogue game for recommendation with adaptive preference models. In *Proceedings of the 14th AAMAS*, pages 959–967, 2015.

[9] M. Lerouge, C. Gicquel, V. Mousseau, and W. Ouerdane. Conception de méthodes d'explication des solutions émanant de systèmes d'optimisation, application à la planification d'employés mobiles. In *23ème congrès annuel de la Société Française de Recherche Opérationnelle et d'Aide à la Décision*, Villeurbanne - Lyon, France, 2022.

[10] V. Mousseau, L.C. Dias, J. Figueira, C. Gomes, and J.N. Clímaco. Resolving inconsistencies among constraints on the parameters of an MCDA model. *European Journal of Operational Research*, 147(1) :72–93, 2003.

[11] R. Pierrard, L. Cabaret, J-P. Poli, and C. Hudelot. Simd-based exact parallel fuzzy dilation operator for fast computing of fuzzy spatial relations. In *WPMVP@PPoPP*, pages 3 :1–3 :8, 2020.

[12] R. Pierrard, J-P. Poli, and C. Hudelot. Learning fuzzy relations and properties for explainable artificial intelligence. In *FUZZ-IEEE*, pages 1–8, 2018.

[13] R. Pierrard, J-P Poli, and C. Hudelot. Spatial Relation Learning for Explainable Image Classification and Annotation in Critical Applications. *Artificial Intelligence*, 292 :103434, 2021.

[14] J-P Poli, W. Ouerdane, and R. Pierrard. Generation of textual explanations in XAI : the case of semantic annotation. In *FUZZ-IEEE*, pages 1–6, 2021.

[15] H. Vroylandt, L. Goudenège, P. Monmarché, F. Pietrucci, and B.Rotenberg. Likelihood-based non-markovian models from molecular dynamics. *PNAS*, 119, 2022.

AfIA
Association française
pour l'Intelligence Artificielle

Recherches en IA explicable à l'Heudiasyc : Systèmes d'aide à la décision sincères, interprétables et redevables

Par

Khaled BELAHCÈNE
Heudiasyc / CID
Université de technologie de Compiègne
Coordinateur du projet

Sébastien DESTERCKE
Heudiasyc / CID
CNRS
Responsable de l'équipe

Sylvain LAGRUE
Heudiasyc / CID
Université de technologie de Compiègne
co-Responsable de l'équipe

Le laboratoire Heudiasyc (heuristique et diagnostic des systèmes complexes) est une unité mixte de recherche sous les tutelles de l'Université de technologie de Compiègne et de l'institut des sciences de l'information et de leurs interactions du CNRS. Il opère dans le domaine des sciences de l'information et du numérique, notamment l'informatique, l'automatique, la robotique et l'intelligence artificielle. En son sein, l'équipe CID (connaissances, incertitudes et données) regroupe les compétences du laboratoire qui sont directement convoquées par les problématiques de l'intelligence artificielle. Les recherches menées dans CID portent sur la gestion des incertitudes, l'apprentissage statistique et l'ingénierie des connaissances, en faisant appel à des formalismes très divers. L'objectif principal de l'équipe consiste à développer des formalismes, des méthodes et des systèmes pour le traitement d'informations et de connaissances en lien avec l'humain, qu'il soit prescripteur ou utilisateur du système.

La capacité à prendre en compte les tenants et les aboutissants d'une décision, et éventuellement à en rendre compte, s'avère fondamentale pour tout système de recommandation, qu'il soit mis en œuvre par des humains ou des agents artificiels. Jadis réservée aux décisions à fort impact, l'aide à la décision s'est démocratisée. Les interrogations quant à la *responsabilité sociale des algorithmes* sont nombreuses : qui décide du choix de l'algorithme ? pour quel objectif ? comment évaluer ce choix ? qui en est responsable ? etc.

Lorsqu'une procédure d'aide à la décision est mise en place, c'est à la demande d'un donneur d'ordre spécifique, afin de répondre à un besoin non moins spécifique. Cette démarche nécessite fréquemment d'arbitrer entre différents points de vue – critères, agents, mondes plus ou moins certains – et il existe de multiples procédures permettant d'effectuer cet arbitrage, issues de la théorie de la décision, ou de l'intelligence artificielle. La plupart s'appuient sur des préférences exprimées par un utilisateur – le décideur – et collectées par l'agent chargé de la recommandation, afin de les étendre à de nouvelles situations. Les approches d'élicitation, qui ont pour objet de garantir l'adéquation entre le décideur et le système mis en place, sont bien adaptées aux problématiques

industrielles, mais doivent faire face à de nouveaux défis liés à la démocratisation et l'automatisation des systèmes d'aide à la décision. En particulier, les décisions prises ont souvent un impact sur des personnes tierces au processus d'aide à la décision (par exemple, les bénéficiaires potentiels d'une procédure d'aide à l'attribution de prêt bancaire). La question de la *redevabilité* – dans quelle mesure et de quelle manière les acteurs du processus d'aide à la décision doivent-ils rendre compte de l'adéquation de la recommandation à des tiers ? [12] – est rendue particulièrement ardue lorsque cette recommandation se fonde sur un processus d'inférence à partir de données.

La problématique de l'explication articule deux thèmes structurants pour l'équipe CID : l'IA de confiance et l'IA personnalisée. Les travaux de l'équipe concernant l'explicabilité des systèmes d'IA s'organisent autour de trois axes : une série de travaux fondamentaux et appliqués sur la modélisation riche des incertitudes, la production d'explications proprement dites pour des recommandations fondées sur des mécanismes décisionnels et l'étude de l'insertion de composants explicatifs dans un processus décisionnel.

Explicabilité de la décision dans l'incertain

L'équipe est fortement impliquée sur la thématique de la modélisation des incertitudes dans des cadres théoriques généralisant ou complémentant les probabilités (théories des probabilités imprécises, des fonctions de croyance et des possibilités).Elle est pleinement investie dans la mise au point, l'étude et la mise en application de modèles interprétables, robustes et sincères pour la décision dans l'incertain. Ces travaux analytiques, visent à une représentation adéquate des situations de décision dans l'incertain, et constituent un socle préalable à la conception et à la mise en œuvre de mécanismes explicatifs.

En particulier, un des objectifs du projet doctoral de Haifei ZHANG, co-encadré par Benjamin QUOST et Marie-Helène MASSON, est de pouvoir extraire des explications concernant l'incertitude associée à une décision, par exemple en différenciant différentes sources d'incertitudes (de modèles, liées à la variabilité, au manque de connaissance, etc.)

Explication de recommandations fondées sur des principes décisionnels

L'usage de modèles intrinsèquement interprétables, ou d'approximation interprétables de modèles obscurs, permet souvent une appropriation du modèle par ceux qui le proposent ou qui le manipulent. Par ailleurs, de nombreux travaux en XAI s'attachent à fournir des outils, placés à l'extrémité terminale de la chaîne de traitement des données, permettant de rendre intelligibles les recommandations ou le fonctionnement de procédures de décisions peu intelligibles. Cependant, ces pratiques s'avèrent insuffisantes pour rendre compte non seulement de la nature des compromis jugés acceptables, mais aussi du lien entre la procédure d'agrégation employée et les données d'apprentissage. Afin de pallier ces insuffisances, [5] a montré comment, à partir d'une base de connaissance cohérente avec une posture axiomatique donnée, il est parfois possible d'utiliser les outils de l'inférence sceptique face aux incertitudes liées à la modélisation afin de justifier les décisions en employant des arguments respectant un schéma, caractéristique de cette posture, et se référant à certaines données de la base de connaissance. Formellement, cette articulation peut être réalisée de manière superficielle, au niveau des instances, en étendant des outils bien connus ([10], et [6] pour une application en choix social), au risque de ne pas pouvoir passer à l'échelle, ou bien en profondeur, à l'aide de propriétés permettant de ca-

ractériser la cohérence entre une structure de préférence et la posture axiomatique : voir [4] pour le modèle du vote par approbation et [5] pour un agrégateur additif.

Dans le cadre du projet doctoral de Hénoïk WILLOT, nous avons récemment élaboré un moteur d'explication pour des recommandations basées sur un modèle de somme pondérée ordonnée (*OWA*) favorisant une répartition équilibrée des performances relatives à différents points de vue. Nous proposons des explications, correctes vis-à-vis du modèle, fondées sur la composition par transitivité d'arguments élémentaires fondés sur la dominance de Pareto, des transferts de Pigou-Dalton, et l'information préférentielle fournie par le décideur [21].

Dans le cadre du projet doctoral de Manuel AMOUSSOU, en partenariat avec Vincent MOUSSEAU, Wassila OUERDANE (MICS, CentraleSupélec) et Nicolas MAUDET (LIP6, Sorbonne Université), nous avons récemment formalisé le concept d'explication schématique, que nous avons instancié pour proposer divers schémas explicatifs pour des recommandations fondées sur une somme pondérée d'attributs binaires [2].

Formalisation d'un processus décisionnel redevable

La problématique de l'explication ne se limite pas à la production d'*explanans*. L'exigence de redevabilité modifie en profondeur les processus d'aide à la décision [18].

Par exemple, si l'on est capable de fournir des explications pour certaines recommandations, on peut penser que seules ces recommandations justifiées seront considérées comme recevables, définissant ainsi le fragment explicable du modèle. Dans le cadre des projets doctoraux de Manuel AMOUSSOU et Jérôme GAIGNE, nous étudions des emboîtements de tels fragments, paramétrés par la complexité acceptable des schémas d'explications, respectivement pour les procédures d'agrégation à base de somme pondérée ou de tri non compensatoire.

Dans une perspective constructiviste de l'élicitation des préférences, le dialogue entre un analyste et le décideur permet à celui-ci d'atteindre un point d'équilibre, son jugement délibéré [7]. Les explications viennent alors outiller ce dialogue [13], permettant au décideur, récipiendaire de l'explication, de mieux s'approprier les conséquences de ses choix et, le cas échéant, de les remettre en question. Ceci constitue un défi pour l'élicitation incrémentale, et les outils de révision des croyances employés jusqu'à présent pour ce faire, fondés sur des ensembles maximaux consistants [15], montrent leurs limites. Nous proposons de nous appuyer sur une modélisation plus fine de l'incertitude [9, 1] pour permettre de détecter et réparer des modèles de préférences incohérents avec un certain modèle d'agrégation, voire de remettre en cause ce modèle [16].

Comment s'assurer qu'une recommandation, issue d'un processus d'aide à la décision, est adéquate ? Souvent ignorée, cette exigence est désormais mise en avant par les communautés structurées autour de la conférence FaCCT ou du séminaire SRA, ou encore le volet IA Acceptable de l'institut ANITI, et doit être prise en compte dès la conception du processus [19]. Afin de mettre en évidence et de pouvoir soumettre à un examen critique les partis-pris de modélisation, il semble crucial de structurer et de documenter le processus d'aide à la décision lui-même [8], d'accompagner la fourniture de la recommandation d'un argumentaire [11], ainsi que d'articuler ces explications au sein d'un processus interactif de haut niveau. Nous proposons de représenter la méta-connaissance concernant la structuration du problème et les choix de modélisation sous la forme de schémas d'arguments, employés dans le domaine

légal dans [20] et, plus récemment, dans le domaine de la validation de systèmes aéronautiques, dans [17]. Nous proposons de définir la redevabilité comme la capacité pour le processus d'aide à la décision d'implémenter une interaction contradictoire entre le décideur et les parties prenantes au sein d'un système multi-agents (SMA). Nous proposons de structurer cette dispute sous la forme d'un jeu de dialogue explicatif [16, 13, 3], formalisé de manière à garantir l'acceptabilité, sinon l'équité, des décisions obtenues. Au sein de ce SMA, l'agent représentant l'analyste veillera à pouvoir répondre à des questions critiques. Nous envisageons deux terrains applicatifs :

- l'automatisation de l'aide médicale à l'hôpital, en s'appuyant sur le projet transdisciplinaire L'intégration de systèmes d'IA à base d'apprentissage automatique dans les systèmes de santé, porté le laboratoire de philosophie des sciences et techniques de l'UTC (Costech) ;
- la gouvernance numérique, et en particulier les systèmes d'affectations des élèves dans les établissements d'enseignement, à New York [14] ou encore en France avec Parcour-Sup.

Références

[1] L. Adam and S. Destercke. Possibilistic preference elicitation by minimax regret. In *Proc. of UAI*, 2021.

[2] M. Amoussou, Kh. Belahcène, N. Maudet, V. Mousseau, and W. Ouerdane. Step-wise explanations for the additive model. Submitted to JIAF, 2022.

[3] N. Asher, S. Paul, and Ch. Russell. Fair and adequate explanations. In *Proc. of CD-MAKE*, 2021.

[4] Kh. Belahcène, Y. Chevaleyre, Ch. Labreuche, N. Maudet, V. Mousseau, and W. Ouerdane. Accountable approval sorting. In *Proc. of IJCAI*, pages 70–76, 2018.

[5] Kh. Belahcène, Ch. Labreuche, N. Maudet, V. Mousseau, and Wassila Ouerdane. Comparing options with argument schemes powered by cancellation. In *Proc. of IJCAI*, pages 1537–1543, 2019.

[6] A. Boixel and U. Endriss. Automated justification of collective decisions via constraint solving. In *Proc. of AAMAS*, pages 168–176, 2020.

[7] O. Cailloux and Y. Meinard. A formal framework for deliberated judgment. *Theory and Decision*, 88(2) :269–295, 2020.

[8] J. Cobbe, M. Seng Ah Lee, and J. Singh. Reviewable automated decision-making : A framework for accountable algorithmic systems. In *Proc. of FAccT '21*, pages 598–609, 2021.

[9] S. Destercke. A generic framework to include belief functions in preference handling and multi-criteria decision. *Int. J. Approx. Reason.*, 98, 2018.

[10] U. Junker. QUICKXPLAIN : preferred explanations and relaxations for over-constrained problems. In Deborah L. McGuinness and George Ferguson, editors, *Proc. of AAAI*, pages 167–172, 2004.

[11] A.-H. Karimi, B. Schölkopf, and I. Valera. Algorithmic recourse : from counterfactual explanations to interventions. In *Proc. of FAccT '21*, 2021.

[12] J. A. Kroll, J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu. Accountable algorithms. *University of Pennsylvania Law Review*, 165(3) :633–705, February 2017.

[13] Ch. Labreuche, N. Maudet, W. Ouerdane, and S. Parsons. A dialogue game for recommendation with adaptive preference models. In *Proc. of AAMAS*, pages 959–967, 2015.

[14] A. Marian. Algorithms in the wild : A case for transparency, accountability and error mitigation in public policy decision-making. In A. Tsoukias, editor, *the 2nd Social Responsibility of Algorithms workshop*, 2019.

[15] V. Mousseau, J. R. Figueira, L. C. Dias, C. Gomes da Silva, and J. C. N. Clímaco. Resolving inconsistencies among constraints on the parameters of an MCDA model. *Eur. J. Oper. Res.*, 147(1), 2003.

[16] W. Ouerdane, N. Maudet, and A. Tsoukiàs. Argument schemes and critical questions for decision aiding process. In *Proc. of COMMA*, 2008.

[17] Th. Polacsek. Diagramme de justification. un outil pour la validation, la certification et l'accréditation. *Ingénierie des Systèmes d Inf.*, 22(2), 2017.

[18] A. Tsoukiàs. On the concept of decision aiding process : an operational perspective. *Ann. Oper. Res.*, 154(1) :3–27, 2007.

[19] K. Vaccaro, K. Karahalios, D. K. Mulligan, D. Kluttz, and T. Hirsch. Contestability in algorithmic systems. In Eric Gilbert and Karrie Karahalios, editors, *Proc. of CSCW*, 2019.

[20] D. Walton, Ch. Reed, and F. Macagno. *Argumentation Schemes*. Cambridge University Press, 2008.

[21] H. Willot, Kh. Belahcene, and S. Destercke. Explications de recommandations fondées sur des principes d'équité à l'aide de transferts. Submitted to JIAF, 2022.

AfIA
Association française
pour l'Intelligence Artificielle

Le projet EXPlainable artificial intelligence : a KnowlEdge CompilaTion FoundATION (EXPEKCTATION)

Par

Gilles AUDEMARD

Steve BELLART

Lounès BOUNIA

Frédéric KORICHE

Jean-Marie LAGNIEZ

Pierre MARQUIS

Nicolas SZCZEPANSKI
CRIL/ EXPEKCTATION
Univ. Artois, CNRS, CRIL UMR 8188, Lens
{nom}@cril.fr
https://www.cril.univ-artois.fr/expekctation/

Contexte et objectifs

EXPEKCTATION (ANR-19-CHIA-0005-01) est un projet de recherche réalisé au CRIL et financé par l'ANR dans le cadre d'une chaire de recherche et d'enseignement en IA du programme national pour l'IA. Le projet a débuté en septembre 2020 pour quatre ans.

EXPEKCTATION est l'abréviation de *EXPlainable artificial intelligence : a KnowlEdge CompilaTion FoundATION*. Le projet EXPEKCTATION concerne le développement d'approches d'IA explicable pour un apprentissage automatique interprétable et robuste, en utilisant des méthodes de raisonnement automatisé à base de contraintes, en particulier la compilation de connaissances.

Nous recherchons des techniques de prétraitement capables d'associer à un prédicteur boîte noire une boîte blanche, pouvant être utilisée pour fournir diverses formes d'explication et répondre à des requêtes de vérification sur la boîte noire afin d'aboutir à des systèmes d'IA en qui l'utilisateur peut avoir confiance. La boîte blanche correspondante pouvant être prétraitée afin de faciliter la génération d'explications des prédictions, indépendamment des entrées associées, la compilation de connaissances apparaît comme une approche très prometteuse à cet égard.

Parmi les questions de recherche abordées, nous souhaitons en particulier déterminer les modèles d'apprentissage et les langages de représentation des boîtes blanches associées qui admettent des algorithmes « efficaces » pour dériver des explications et prendre en charge des requêtes de vérification. Nous étudions la complexité du calcul de divers types d'explications. Nous comptons aussi développer et évaluer des algorithmes pour ces tâches. Nous souhaitons enfin étudier comment produire des explications qui soient les plus intelligibles possible, en prenant en compte des critères intrinsèques aux explications (taille, nombre, structure, etc.) mais aussi des critères extrinsèques à celles-ci (le contexte de l'explication, l'utilisateur).

Travaux déjà réalisés

Les travaux réalisés jusqu'ici ont concerné plusieurs questions de recherche et ont permis

d'avancer vers des solutions possibles :

- Formaliser le niveau de confiance de familles de classeurs par un ensemble de requêtes XAI traitables [3] : nous avons identifié un ensemble de requêtes d'explication et un ensemble de requêtes de vérification de systèmes de classement. Ouvrir la boîte noire correspondant à la famille de classeurs considérée, c'est précisément disposer d'algorithmes efficaces pour répondre à ces requêtes. L'ensemble des requêtes traitables obtenues peut être vu comme une caractérisation multicritère de l'interprétabilité de la famille de classeurs étudiée.
- Calculer ce niveau de confiance pour diverses familles de classeurs [1] : nous avons déterminé parmi un vaste ensemble de requêtes XAI celles qui sont traitables pour diverses familles de classeurs booléens, incluant les arbres de décision, les forêts aléatoires, les formules DNF, les listes de décision, les perceptrons booléens multicouches et les réseaux de neurones binaires. Les arbres de décision connus pour leur capacité à expliquer simplement le classement opéré se démarquent des autres familles de classeurs étudiés par le fait que toutes les requêtes XAI considérées sont traitables pour les arbres de décision, alors qu'aucune d'entre elles n'est traitable pour les autres familles analysées.
- Étendre la logique booléenne quantifiée en ajoutant des opérateurs à la logique booléenne pour quantifier universellement les littéraux [5] : de nombreuses requêtes XAI correspondent, en effet, à un processus de sélection d'instances avec des propriétés particulières, où la formule booléenne caractérisant ces instances constitue la réponse à la requête XAI. La quantification universelle des littéraux peut être utilisée pour sélectionner de telles instances et pour formuler et généraliser certaines notions récemment introduites dans le domaine de l'IA explicable.
- Définir une notion d'explication abductive adaptée aux forêts aléatoires [2] : calculer une explication abductive irredondante d'un classement réalisé, aussi appelée raison suffisante de ce classement, est calculatoirement difficile pour les forêts aléatoires et l'est encore plus quand on se focalise sur les explications de taille minimale. Pour pallier cette difficulté, nous avons introduit une nouvelle notion d'explication abductive adaptée aux forêts aléatoires, que nous avons appelé raison majoritaire. Nous avons présenté un algorithme en temps polynomial pour calculer de telles raisons. Nous avons également montré empiriquement la possibilité de dériver des raisons majoritaires de tailles plus petites que celles des raisons suffisantes que l'on arrive à dériver.
- Définir un cadre pour la rectification de classeurs qui prédisent de façon erronée [4] : nous avons abordé la question de la prise en compte de connaissances expertes dans un classeur booléen *multilabel* et défini un ensemble de postulats à respecter pour rectifier un classeur booléen quand les classements qu'il propose entrent en conflit avec ceux issus de connaissances utilisateur, jugés plus fiables. Nous avons en particulier montré que l'opération de rectification de classeurs booléens diffère de celle de révision ou encore de celle de mise à jour, bien connues dans le domaine du changement de croyances.

Références

[1] G. Audemard, S. Bellart, L. Bounia, F. Koriche, J.-M. Lagniez, and P. Marquis. On the computational intelligibility of boolean classifiers. In *Proc. of KR'21*, pages 74–86, 2021.

[2] G. Audemard, S. Bellart, L. Bounia, F. Ko-

riche, J.-M. Lagniez, and P. Marquis. Trading complexity for sparsity in random forest explanations. In *Proc. of AAAI'22*, 2022.

[3] G. Audemard, F. Koriche, and P. Marquis. On tractable XAI queries based on compiled representations. In *Proc. of KR'20*, pages 838–849, 2020.

[4] S. Coste-Marquis and P. Marquis. On belief change for multi-label classifier encodings. In *Proc. of IJCAI'21*, pages 1829–1836. ijcai.org, 2021.

[5] A. Darwiche and P. Marquis. On quantifying literals in boolean logic and its applications to explainable AI. *J. Artif. Intell. Res.*, 72 :285–328, 2021.

AfIA
Association française
pour l'Intelligence Artificielle

Les projets Hybrid Approaches for Interpretable Artifical Intelligence et Framework for Automatic Interpretability in Machine Learning

Par **Elisa FROMONT**
IRISA/ INRIA LACODAM team
Université de Rennes
elisa.fromont@irisa.fr

Luis GALÁRRAGA
IRISA/ INRIA LACODAM team
INRIA
luis.galarraga@inria.fr
https://team.inria.fr/lacodam/

The HyAIAI project

Recent progress in Machine Learning (ML) and especially Deep Learning has made ML pervasive in a wide range of applications. However, current approaches rely on complex numerical models : their decisions, as accurate as they may be, cannot be easily explained to the layman that may depend on these decisions (ex : get a loan or not). In the Inria Défi HyAIAI (Hybrid Approaches for Interpretable Artifical Intelligence), we tackle the problem of making "Interpretable ML" through the study and design of hybrid approaches that combine state of the art numeric models with explainable symbolic models. More precisely, our goal is to be able to integrate high level (domain) constraints in ML models, to give model designers information on ill-performing parts of the model, and to give the layman/practitioner understandable explanations on the results of the ML model.

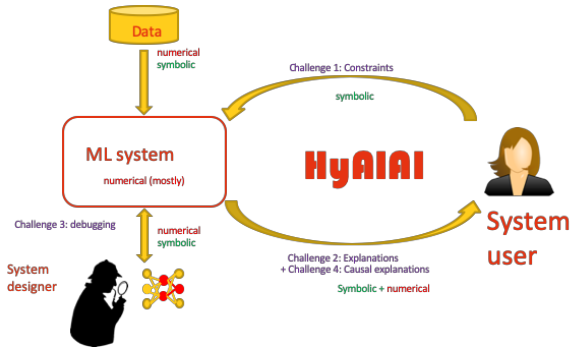
We explore 4 different challenges :

- Challenge 1 ("System must understand human requirements") focuses on methods to constrain black box models with user preferences or requirements and to integrate background knowledge into black boxes.
- Challenge 2 ("Human must understand system responses") focuses on developing methods for interpretable AI, whether they are symbolic (e.g. rule-based) ; numerical (e.g. explainable-by-design black boxes) or both. The majority of the work and publications within the HyAIAI project concerns this challenge.
- Challenge 3 ("Human must understand the inner working of the system") is about trying to find patterns in the activation maps of the neurons of a neural network, that could be characteristics of the neural network's mispredictions. The current conclusion is that it is possible to find activation patterns for correct classifications but it is unclear that such patterns exist for mispredictions.
- Challenge 4 ("Causality, yet another dimension of explainability"). Causality is at the heart of explanations, as one often seek causal answers to questions regarding why or how something actually occurred. However, new techniques should be developed to step away from the standard explanations which are based on the correlations between some data inputs and the model outputs.

The members of the project are currently collectively working (together with Julien Aligon in Toulouse) to set up a challenge on explainable AI including, if possible, problems which would involve causality explanations. These are the requirements for the datasets

we seek for this challenge :

- Data should be public or publishable without privacy issues ;
- Data should be original and non trivial (enough predicted variables, enough samples) ;
- Data should be associated with a prediction problem that could be modeled with complex ML methods (e.g. neural networks). Ideally, data should be labeled, and the ML method should be supervised. Explanations will be issued according to the predictions of this model ;
- One (black box) predictions model should be made available for the targeted task (to support post-hoc explanation methods) ;
- The (possibly causal) explanation ground truth should be available together with the dataset and the prediction problem.



The FAble project

FAbLe (Framework for Automatic Interpretability in Machine Learning) is an ANR-funded (JCJC) project that aims to develop techniques for the automatic selection of the most suitable explanations for machine learning models. This suitability comprises two intertwined dimensions : technical and human.

On the technical side we work on the techniques to characterize an explanation use case consisting of an instance (e.g., a loan request profile) and a target model (e.g., a classifier used to accept or refuse loan requests), and decide whether this use case can be explained unambiguously using a feature attribution ranking – often computed via a linear approximation. To this end, we have developed APE (Adapted Posthoc Explanations), a method that builds upon the notions of distribution uni-modality and linear separability to characterize the decision frontier of the classifier. When the decision frontier is too complex, i.e., it is composed of clusters of instances that are hardly linearly separable with a single model, APE proposes alternatives such as rule-based explanations. In all cases, the explanations are augmented with counter-factual instances. These are complementary explanations that show the minimal changes in the input required to alter the black-box answer. They can, for instance, illustrate the changes required in a loan request profile in order to pass from a "reject" to an "accept".

On the human side, we envision to study the comprehensibility of the different explanation paradigms, namely feature-attribution rankings, rule-based explanations, and exemplars (e.g., counterfactuals). To this end, we have planned to conduct a series of user studies that will confront human users to such kind of explanations (including combinations of the aforementioned paradigms). We will then measure the effects of the explanations on comprehensibility, but also on the trust that users have on AI models (see Fig. 1.2).

All this work will be part of the PhD of Julien DELAUNAY, supervised by Christine LARGOUËT and Luis GALÁRRAGA.

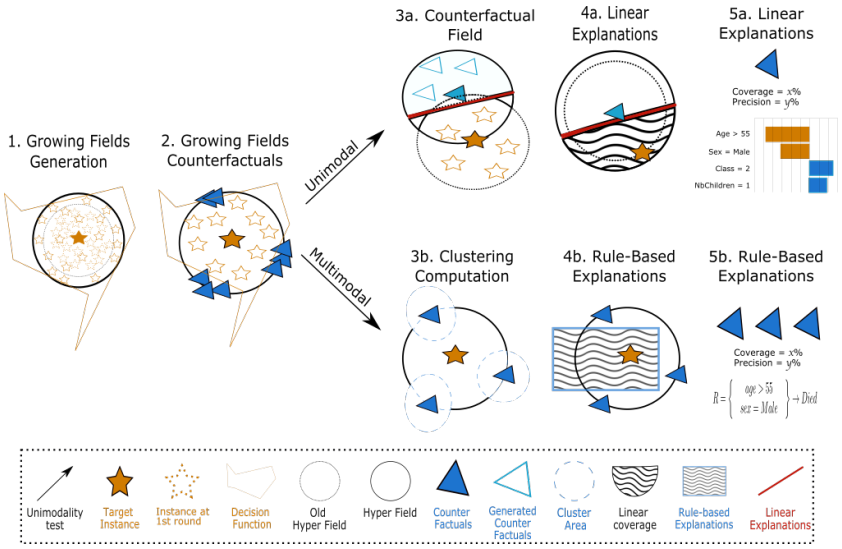


Figure 1.2 – Overview of the Adapted Posthoc Explanation method

Étudier les mécanismes des effets indésirables des médicaments avec l'IA explicable : expériences avec la fouille de graphes de connaissances

Par

Pierre MONNIN
Orange
pierre.monnin@orange.com
https://pmonnin.github.io

Adrien COULET
Centre de Recherche des Cordeliers, Inria Paris/HeKa
adrien.coulet@inria.fr
https://team.inria.fr/heka/team-members/coulet/

Introduction

Les événements indésirables médicamenteux (EIM) sont mis en évidence et statistiquement caractérisés grâce à des essais cliniques randomisés, puis à la pharmacovigilance après mise sur le marché. Cependant, les mécanismes moléculaires responsables de ces effets restent inconnus dans la plupart des cas. C'est notamment le cas pour les toxicités hépatiques ou cutanées qui sont pourtant particulièrement surveillées durant la phase d'évaluation des médicaments. En marge des essais cliniques, de nombreuses autres connaissances à propos des médicaments et des molécules les constituant sont disponibles. En particulier, des graphes de connaissances en accès libre décrivent leurs propriétés, interactions et implications dans les réseaux biologiques (*i.e.*, les *pathways*). D'autre part, des classifications expertes ont été établies manuellement et permettent de disposer de listes de médicaments connus pour causer, ou non, différents types

d'EIM.

L'IA explicable pour expliquer les EIM

Dans un récent article [1], nous avons fouillé des graphes de connaissances biomédicaux afin d'identifier les propriétés biomoléculaires des médicaments qui permettent de reproduire automatiquement les classifications expertes, distinguant les médicaments causant ou non un type spécifique d'EIM. Dans la perspective de produire des artefacts capables d'expliquer ces classifications, nous avons mis en œuvre des méthodes de classification nativement explicables : les arbres de décision et les règles de classification. En effet, ces méthodes produisent des modèles compréhensibles par les humains qui expliquent la classification de nouveaux exemples [4]. De plus, nous avons évalué l'hypothèse selon laquelle les propriétés associées aux médicaments pourraient aussi fournir aux experts des éléments explicatifs pour les mécanismes moléculaires sous-jacents aux EIM.

Nous avons testé notre approche à partir de deux classifications expertes distinguant les médicaments causant ou non des toxicités hépatiques (ou DILI, pour *Drug-Induced Liver Injuries*) et cutanés (ou SCAR, pour *Severe Cutaneous Adverse Reactions*). Pour extraire des propriétés associées à ces médicaments, nous les alignons avec un graphe de connaissances que nous fouillons ensuite. Il s'agit ici de PGx-LOD, un graphe de connaissances biomédicales que nous avons précédemment créé [2] et qui intègre, connecte et complète des graphes de connaissances publics (par exemple CTD, DisGeNET, DrugBank, etc.). Dans le processus de fouille, les nœuds représentant les médicaments sont appelés « graines » et nous extrayons leurs voisins, les chemins et les patrons de chemins dont ils sont racines. Par exemple, sur la figure 1.3, v_4 est un voisin de n_2, $\xrightarrow{p_1} v_2 \xrightarrow{p_2} v_3$ est un chemin dont n_1 est racine, et $\xrightarrow{p_1} C_1 \xrightarrow{p_2} C_3 \xrightarrow{p_3} v_6$ est un patron de chemins dont n_1 et n_2 sont racines. L'intérêt des patrons de chemins réside dans leur capacité, par généralisation, à être communs à davantage de graines que les chemins. La figure 1.3 illustre comment les nœuds de deux chemins distincts peuvent être généralisés et remplacés par les classes que ceux-ci instancient pour créer un patron de chemins unique. Nous avons particulièrement étudié le passage à l'échelle de ce processus de fouille en créant l'algorithme `kgpm` [3].

Sur la base des propriétés extraites, nous avons entraîné deux classifieurs à distinguer les médicaments causant ou non chacun des deux types d'EIM étudiés. Nous isolons les propriétés qui sont à la fois discriminantes dans la reproduction des classifications expertes et interprétables par des experts (*i.e.*, des termes de la *Gene Ontology*, des molécules cibles de médicaments ou des noms de réseaux biologiques). Le caractère explicatif des propriétés isolées a ensuite été manuellement évalué par 3 experts en pharmacologie.

Évaluation quantitative et qualitative

Les propriétés extraites du graphe permettent de reproduire avec un bon niveau de fidélité les classifications expertes des médicaments causant ou non les EIM (*accuracy* de 0,74 et 0,81 pour DILI et SCAR respectivement). Les experts ont évalué unanimement que 73 % (pour DILI) et 38 % (pour SCAR) des propriétés discriminantes sont possiblement explicatives des EIM associés. Ils sont partiellement d'accord (2/3) pour dire que 90 % (pour DILI) et 77 % (pour SCAR) de ces propriétés sont possiblement explicatives.

Par exemple, le patron de chemin $\xrightarrow{\text{interactsWith}}$ **Enzyme** $\xrightarrow{\text{cellularComponent}}$ **Endoplasmic reticulum** a été unanimement évalué comme potentiellement explicatif pour DILI. En effet, le réticulum endoplasmique est

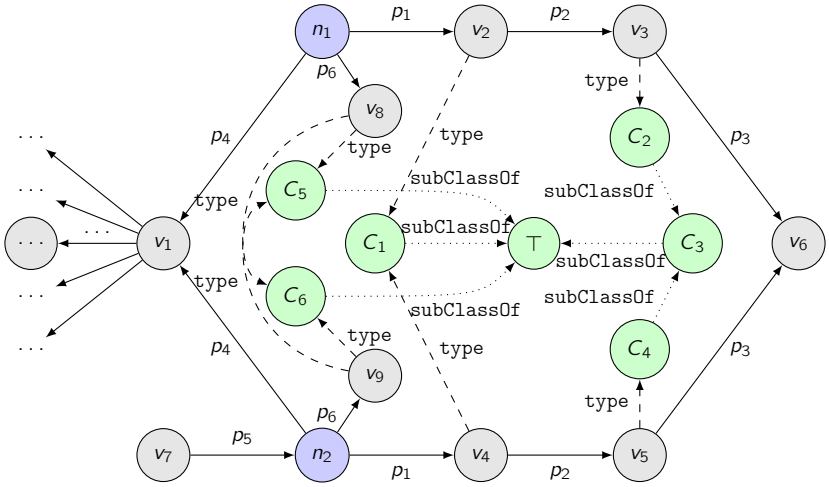


Figure 1.3 – Exemple de graphe de connaissances. n_1 et n_2 sont les « graines » pour l'extraction de propriétés, les v_i sont des individus voisins, et les C_i sont des classes d'ontologie utilisées pour la construction de patrons de chemins. Par exemple, n_1 et n_2 sont caractérisés par le voisin $\boldsymbol{v_1}$, le chemin $\xrightarrow{p_4} \boldsymbol{v_1}$, et les patrons de chemins $\xrightarrow{p_6} \top$ et $\xrightarrow{p_1} \boldsymbol{C_1} \xrightarrow{p_2} \boldsymbol{C_3} \xrightarrow{p_3} \boldsymbol{v_6}$. Des paramètres s'appliquent pour limiter le nombre de voisins, chemins, et patrons de chemins extraits (par exemple support, spécificité, longueur) [3].

connu, en particulier dans les tissus hépatiques, pour contenir des enzymes de la famille des cytochromes P450 qui sont fortement impliqués dans le métabolisme des médicaments.

Conclusion et perspectives

Notre étude suggère que le graphe de connaissances utilisé fournit des propriétés suffisamment diverses pour permettre à des modèles simples et explicables de distinguer les médicaments causant ou non des EIM. En plus de permettre de reproduire les classifications expertes, les propriétés discriminantes semblent porter le sens d'une réalité biologique qui en font des candidats à considérer pour étudier de manière plus approfondie les mécanismes biologiques sous-jacents aux EIM.

Dans notre travail, nous nous sommes concentrés sur des classifieurs nativement explicatifs, mais relativement naïfs (arbres de décision et règles de classification). Il semble intéressant de reproduire nos expériences avec des réseaux de neurones profonds adaptés aux graphes comme les *Graph Convolutional Networks* et les *Graph Neural Networks* pour mesurer la perte de performance potentielle associée à notre choix de modèle. Une telle comparaison quantitative devrait toutefois s'accompagner d'une comparaison qualitative prenant en compte la nécessite d'étapes supplémentaires pour expliquer ces modèles profonds. Par exemple, l'utilisation de cartes de saillance (*saliency maps*) permet d'obtenir des informations sur la couche ou les neurones activés par certaines instances. Ces informations sont particulièrement intéressantes pour les spécialistes en apprentissage machine, mais nécessitent un travail d'interprétation supplémentaire avant d'être compréhensibles par des experts du domaine d'application, parfois non familiers avec les réseaux de neurones. Les valeurs de Shapley

sont plus généralement interprétables et pour cette raison, leur adaptation aux méthodes de fouille de graphes mériterait d'être considérée dans la poursuite de ce travail.

Références

[1] E. Bresso et al. Investigating ADR mechanisms with explainable AI : a feasibility study with knowledge graph mining. *BMC Medical Informatics and Decision Making*, 21(1) :171, 2021.

[2] P. Monnin et al. PGxO and PGxLOD : a reconciliation of pharmacogenomic knowledge of various provenances, enabling further comparison. *BMC Bioinformatics*, 20-S(4) :139 :1–139 :16, 2019.

[3] P. Monnin et al. Tackling scalability issues in mining path patterns from knowledge graphs : a preliminary study. In *Proc. of the 1st Int. Conf. "Algebras, graphs and ordered sets"*, volume 2925 of *CEUR Workshop Proceedings*, pages 123–137. CEUR-WS.org, 2020.

[4] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5) :206–215, 2019.

AfIA
Association française
pour l'Intelligence Artificielle

Enjeux éthiques de l'usage de l'intelligence artificielle en médecine : apports d'un regard épistémologique

Par

Éric PARDOUX

Mogens LÆRKE
UMR 5317 « Institut d'Histoire des Représentations et des Idées dans les Modernités »
UAR 3129/UMIFRE 11 Maison Française d'Oxford
{prénom}.{nom}@ens-lyon.fr

Thomas GUYET
INRIA – Centre de Lyon
thomas.guyet@inria.fr

Introduction

Nous présentons ici le projet de recherche ED-AIM (*Ethical Design for Artificial Intelligence Models*)[1] qui aborde depuis la philosophie les enjeux de l'intégration de l'intelligence artificielle (IA) dans le domaine de la santé. Notre approche vise à mêler l'épistémologie de l'IA (l'étude de la façon dont elle se fait) et l'éthique de ses usages. L'un des enjeux principaux que nous portons est de traiter les questions éthiques soulevées par l'IA en santé au travers de la mise en place d'un cadre de travail transdisciplinaire international visant à faire dialoguer informaticiens, philosophes et praticiens hospitaliers. Il s'agit donc non seulement de considérer les enjeux éthiques et épistémiques de l'IA en elle-même, mais également par rapport au contexte socio-technique dans lequel elle doit s'intégrer.

Éthique et AI : une recherche interdisciplinaire

Le présent projet de recherche est financé par la Mission pour les initiatives transverses et interdisciplinaires (MITI) du CNRS. Par sa nature internationale, le projet possède une visée comparative entre les cadres socio-techniques français et britannique.

Ce projet s'accompagne de collaborations étroites avec Lionel TARASSENKO (Institute of Biomedical Engineering, Université d'Oxford) et Angeliki KERASIDOU (Ethox Centre & Wellcome Centre for Ethics and Humanities, Université d'Oxford).

Forts de ce cadre interdisciplinaire, nous avons la volonté d'organiser au cours des années à venir plusieurs événements, à Oxford et en France, sur les thématiques liées à l'épistémologie de l'IA et à l'éthique de ses usages et développements en santé.

L'IA en santé : quelles attentes ?

L'IA, notamment avec l'essor de l'apprentissage automatique, suscite un intérêt croissant dans le domaine des soins de santé. Elle est censée améliorer la précision de la médecine et accroître la personnalisation des soins. Les applications de l'IA dans le domaine des soins de santé sont légion : de l'assistance automatisée des robots chirurgiens aux algorithmes d'aide à la décision en passant par la gestion du niveau d'insuline pour les diabétiques, elle s'intègre profondément dans le tissu médical. Parmi cette pléthore d'applications, nous avons choisi de consacrer nos recherches plus spécifiquement aux modèles d'IA employés pour l'aide

1. Projet CNRS Prime 80 financé sur 2021–2024.

à la décision médicale, que ce soit pour le diagnostic, le choix de thérapeutiques ou la stratification des risques de patients. Traditionnellement, ces tâches reposaient en grande partie sur la prise de décision des médecins et personnels hospitaliers, aidés par des cadres de référence et des systèmes de score clinique. L'émergence de ce que l'on appelle la médecine fondée sur les preuves (*Evidence-Based Medicine* ou EBM) a favorisé le développement de tels outils cliniques, par une pratique accrue des essais randomisés par exemple. Parallèlement, le développement des techniques biomédicales est supposé transformer la pratique médicale en rendant la médecine de plus en plus prédictive, personnalisée, préventive et participative.

La délégation à l'IA du traitement des données générées par les patients est donc avant tout envisagée pour parvenir à tirer profit de ces avancées [1]. En effet, la quantité de données biomédicales – aussi bien dans leur ensemble que celles disponibles pour chaque patient – devient si importante qu'une vue d'ensemble qui soit exhaustive est difficilement envisageable par l'humain seul. Néanmoins, cette introduction de l'IA en santé soulève des problèmes philosophiques.

L'explicabilité, un enjeu seulement technique et épistémologique ?

L'arrivée de l'IA en santé vient potentiellement bouleverser le paradigme à l'œuvre dans la médecine. La donnée acquiert une importance centrale et des raisonnements basés sur des mécanismes causaux risquent de laisser place à une causalité avant tout basée sur l'inférence statistique [2]. Il convient d'étudier les implications (positives comme négatives) que ces transformations provoquent dans les systèmes de santé. Du point de vue éthique tout d'abord, de nombreux biais sont liés aux données : qu'il s'agisse de leur adéquation au problème abordé, de leur nature potentiellement discriminatoire ou encore de leur capacité ou non à transcrire fidèlement des phénomènes médicaux. De plus, du fait même de la façon dont les systèmes d'IA basés au moins partiellement sur l'apprentissage machine sont construits, une question éthique d'une nature semblant nouvelle se pose : l'explicabilité. Si cette dernière est depuis longtemps présente en informatique sous l'avatar de l'interprétabilité des modèles pour leur débogage, elle acquiert une nouvelle dimension dans le cadre de l'IA en santé, ses contours devenant flous [3].

Afin de se saisir pleinement des enjeux liés à l'IA, toutes les parties prenantes doivent avoir accès à un niveau minimal de compréhension du fonctionnement de l'IA et des modèles d'aide à la décision qu'elle peut générer. Cela a mené à ériger l'explicabilité comme cinquième grand principe éthique à associer au développement d'IA, en complément des quatre autres grands principes empruntés à l'éthique biomédicale : bienfaisance, non-malfaisance, justice et respect de l'autonomie du patient [4]. Une telle posture se traduit également dans les recommandations émises au niveau européen [2] [3].

Cette exigence d'explicabilité renvoie directement à la nature des modèles d'IA employés dans le champ médical. L'IA peut se représenter comme une boîte noire qu'il s'agirait d'ouvrir afin de pouvoir l'utiliser en connaissance de cause. La capacité à expliquer le fonctionnement des décisions obtenues grâce à l'IA peut sembler centrale à l'encadrement de celle-ci. Néanmoins, des enjeux liés aux pratiques médicales se posent également suite à l'introduction de l'IA. La question de la confiance dans les systèmes de soin est majeure lors de l'introduction de l'IA [5].

Pour cette raison, ce n'est pas seulement

2. Voir par exemple la proposition de règles harmonisées du 21/04/2021
3. Voir également les Lignes directrices en matière d'éthique pour une IA digne de confiance.

le modèle basé sur l'IA qu'il s'agit de considérer, mais l'intégralité du milieu socio-technique dans lequel il s'intègre. Cela implique de suivre toute la vie du dispositif, de sa conception initiale jusqu'au suivi de son intégration dans le milieu hospitalier. Aboutir à un dispositif éthique implique de considérer les enjeux éthiques liés à chaque étape de son développement et pas uniquement de ses usages. On ne saurait en effet tolérer un dispositif éthique dans ses usages, mais dont la production ne l'aurait pas été.

Vers une IA digne de confiance éthique by design ?

Il s'agit donc de défendre une conception de l'IA pour la santé qui intègre un aspect systémique. Chaque partie prenante doit être considérée dans le processus de développement.

Dans cet objectif, la philosophie de la médecine nous offre un regard englobant sur la pratique médicale qui permet de compléter les attentes qui peuvent être formulées au sein d'un système de santé idéal. La relation entre patient et soignant va ainsi au-delà des grands principes de l'éthique biomédicale déjà formulés. Des valeurs comme la confiance ou encore l'empathie sont à l'œuvre dans les relations de soin. Il est donc important de questionner la façon dont l'IA peut être incorporée à ce tissu sans le dénaturer, voire même en favorisant l'émergence et le renforcement de ces valeurs dans le soin. L'objectif fixé est celui d'une IA digne de confiance, à toutes les échelles – à la fois dans la relation du patient au soignant, mais aussi vis-à-vis de ses concepteurs ou encore des institutions.

Concrètement, ce travail va passer par une revue systématique de la littérature grise portant sur l'IA médicale, afin de questionner la pertinence des exigences éthiques transmises aux développeurs. Ensuite, il s'agira de poser un cadre éthique en évolution de celui pré-existant, soit via une reformulation, soit via une refonte profonde de certains aspects. La question de la reconfiguration des savoirs et pratiques médicales engendrée par l'IA sera ici centrale. La clarification de ce cadre éthique devrait permettre d'élaborer un ensemble de règles ou *a minima* de recommandations spécifiquement adaptées au contexte médical, à destination tout autant des informaticiens que des autres parties prenantes dans le développement d'IA (experts médicaux, institutions, etc.).

L'objectif est ainsi de parvenir à dessiner les contours d'un cadre de développement qui puisse garantir que l'IA soit éthique et digne de confiance *by design*. L'usage se pensant dès la conception, les considérations éthiques doivent pouvoir accompagner le processus de conception pour garantir un design final du système favorisant les usages éthiques.

Références

[1] A. Rajkomar *et al.* Machine Learning in Medicine In *New England Journal of Medicine*, pages 1347–1358. Massachusetts Medical Society, 2019.

[2] J.P. Dupuy La nouvelle science des données. In *Esprit*, pages 89–98. CAIRN, 2019.

[3] J.J. Wadden. Defining the undefinable: the black box problem in healthcare artificial intelligence. In *Journal of Medical Ethics*. BMJ, 2021.

[4] L. Floridi *et al.* AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. In *Minds and Machines*, pages 689–707. Springer, 2018.

[5] C. Kerasidou *et al.* Before and beyond trust: reliance in medical AI. In *Journal of Medical Ethics*. BMJ, 2021.

AfIA
Association française
pour l'Intelligence Artificielle

L'explicabilité des maisons intelligentes

Par

Étienne HOUZÉ
SEIDO
EDF R&D
www.edf.fr

Jean-Louis DESSALLES

Ada DIACONESCU
LTCI
Télécom Paris
{prénom}.{nom}@telecom-paris.fr
www.telecom-paris.fr

Introduction

L'Intelligence Artificielle Explicable (XAI) est un sujet qui a pris une ampleur grandissante au cours de ces dernières années, en réponse à un besoin pour plus de transparence et de confiance envers des algorithmes qui prennent une place prépondérante dans le quotidien. Notamment, la réglementation européenne du RGPD définit un « droit à l'explication » des utilisateurs affectés par des décisions d'algorithmes. La recherche actuelle s'oriente vers les modèles d'IA dits « boîte noire » dont le fonctionnement opaque pose de graves problèmes de compréhension. Cependant, d'autres domaines pourraient bénéficier d'explicabilité. C'est le cas des technologies d'IA embarquée, des appareils connectés que l'on retrouve par exemple dans les maisons intelligentes.

Ces dernières se présentent comme des systèmes cyber-physiques complexes, dans lesquels de nombreux équipements (radiateurs, thermomètres, capteurs de présence, de lumière, contrôleurs, etc.) interagissent avec un environnement physique unique. Alors que chaque composant pourrait individuellement faire l'objet d'une approche XAI pour expliquer ses décisions, l'explication d'un phénomène faisant intervenir plusieurs composants est pour l'instant non traitée. Notre but est donc de pallier ce manque par une approche respectant les caractéristiques du système de contrôle existant.

Les enjeux de l'explication des Smart Homes

La maison intelligente consiste en un ensemble d'équipements connectés qui, ensemble, permettent la réalisation d'objectifs de haut niveau affectant la maison et la vie de son occupant, tel que le confort (température, lumière), la sobriété énergétique, la sécurité, le suivi de la santé d'un occupant âgé [5]. Cependant, malgré les nombreuses promesses portées par le domaine, l'adoption des technologies existantes reste pour le moment cantonnée à une utilisation anecdotique, loin de la démocratisation annoncée. Comment expliquer ce décalage ?

Au-delà des aspects de coût qui sont bien évidemment cités, le manque de transparence des technologies de maison intelligente, ainsi que les craintes face à une perte de contrôle de son logement ou un non-respect de la vie privée, ont été identifiées comme étant des arguments importants [7]. Or, ces derniers points sont parmi les axes identifiés comme principaux pour l'XAI [6]. Il semble donc logique que des solutions XAI soient développées afin de

contribuer à l'amélioration des maisons intelligentes et à leur adoption par le grand public. Cependant, les maisons intelligentes sont des cas complexes pour le développement d'une IA explicative.

Smart Home et XAI : difficultés

Les maisons intelligentes sont des exemples de systèmes auto-adaptatifs : en relation étroite avec un environnement changeant, elles sont capables de fournir un haut niveau de service. Par exemple, une caractéristique commune des équipements récents est le « plug-and-play » : un nouvel équipement est automatiquement intégré au système existant, offrant de nouvelles possibilités de contrôles et d'interactions. Cependant, ces caractéristiques posent un problème au niveau de l'explicabilité : il faut en effet qu'un système explicatif pour la maison intelligente les préserve, mais aussi en tire parti : ainsi, l'ajout d'un nouvel équipement doit pouvoir être considéré dans un raisonnement pouvant expliquer une situation étrange à l'utilisateur.

Le sujet des explications à fournir pose également un problème majeur : il en ressort que les situations les plus à même de nécessiter une explication par le système sont les situations inhabituelles, voire étranges aux yeux de l'utilisateur : un malfonctionnement d'un des systèmes de régulation, un comportement vu pour la première fois, etc.. Or, ces situations sont par définition rares, il est dès lors probable que peu de données préexistantes permettent de proposer avec grande précision des hypothèses correctes. Il faut donc un système explicatif qui soit axé autour de ces événements inhabituels et présente un certain « droit à l'erreur ».

Les maisons intelligentes et autres systèmes cyber-physiques complexes présentent souvent des capacités d'auto-adaptation afin de faciliter leur maintenance et permettre leur bon fonctionnement. De tels systèmes peuvent présenter des organisations variées : centralisées, hybrides ou décentralisées [4]. Ces différentes organisations rendent difficiles la réalisation d'un système permettant de couvrir toutes les architectures. De plus, il est possible que des équipements gardent secrets certaines de leurs mesures ou fonctionnement, dans un souci de propriété intellectuelle ou de sécurité.

Solutions existantes

Peu de solutions existent dans l'état de l'art actuel, que ce soit au sein de la communauté XAI ou système auto-adaptatifs, pour répondre au problème de l'explicabilité de tels systèmes. Une étude précédente [1] propose par exemple de s'inspirer du paradigme MAPE (*Monitor - Analyze - Plan - Execute*), déjà largement utilisé dans les systèmes adaptatifs [4], afin d'intégrer l'explication des décisions. Cependant, la méthode MAB-EX ainsi obtenue considère le système de contrôle comme un ensemble monolithique, ce qui ne correspond pas à l'organisation souvent décentralisée ou hybride des maisons intelligentes.

Nous proposons donc d'utiliser une approche distincte de la question de l'explicabilité, en visant à fournir une approche de haut niveau permettant d'intégrer les caractéristiques d'auto-adaptation et de généricité du système de contrôle de la maison intelligente. Pour ce faire, nous nous basons sur des travaux précédents qui avaient identifié trois étapes clés et suffisantes à la tenue d'un dialogue argumentatif : la *détection de conflits*, l'*abduction* et la *négation* [2]. Nous proposons d'appliquer ces principes dans un but d'explication : en considérant le phénomène questionné comme un conflit, la trace du raisonnement argumentatif visant à le résoudre est présenté à l'utilisateur comme explication. Afin de s'adapter aux contraintes du système, la localité des connaissances est préservée : chacune des trois principales étapes du raisonnement est effectuée par

un composant local directement attaché à un équipement. Un coordinateur central se charge de mettre en relation les composants nécessaires et d'interroger le composant expert dans le conflit à expliquer présentement. Le résultat est la génération d'un raisonnement itératif qui explore les causes possibles successivement [3].

Références

[1] M. Blumreiter, J. Greenyer, F. J. Chiyah Garcia, V. Klös, M. Schwammberger, Ch. Sommer, A. Vogelsang, and A. Wortmann. Towards self-explainable cyber-physical systems. In *ACM/IEEE 22nd Int. Conf. on Model Driven Engineering Languages and Systems Companion (MODELS-C)*, pages 543–548. IEEE, 2019.

[2] J.-L. Dessalles. A Cognitive Approach to Relevant Argument Generation. In Matteo Baldoni, Cristina Baroglio, and Floris Bex, editors, *Principles and Practice of Multi-Agent Systems, LNAI 9935*, pages 3–15. Springer, 2016.

[3] E. Houzé, A. Diaconescu, J.-L. Dessalles, D. Menga, and M. Schumann. A Decentralized Approach to Explanatory Artificial Intelligence for Autonomic Systems. In *ACSOS Conf. Proc., Companion*, 2020.

[4] Ch. Krupitzer, F. Maximilian Roth, S. VanSyckel, G. Schiele, and Ch. Becker. A survey on engineering approaches for self-adaptive systems. *Pervasive and Mobile Computing*, 17 :184–206, 2015.

[5] D. Marikyan, S. Papagiannidis, and E. Alamanos. A systematic review of the smart home literature : A user perspective. *Technological Forecasting and Social Change*, 138 :139–154, 2019.

[6] S. T Mueller, R. R. Hoffman, W. Clancey, A. Emrey, and G. Klein. Explanation in human-AI systems : A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. Technical report, Defense Technical Information Center, 2019.

[7] G. Zimmermann, T. Ableitner, and Ch. Strobbe. User Needs and Wishes in Smart Homes : What Can Artificial Intelligence Contribute ? In *14th Int. Symposium on Pervasive Systems, Algorithms and Networks & 11th Int. Conf. on Frontier of Computer Science and Technology & Third Int. Symposium of Creative Computing (ISPAN-FCST-ISCC)*, pages 449–453. IEEE, 2017.

AfIA
Association française
pour l'Intelligence Artificielle

ExpressIF®, une IA symbolique et explicable

Par
Nadia BEN ABDALLAH
Laurence BOUDET
Edwin FRIEDMANN
Aurore LOMET
Jean-Philippe POLI
CEA List
Université Paris Saclay
{prénom}.{nom}@cea.fr
https://expressif.cea.fr

Introduction

Au sein du CEA List (Université Paris Saclay), dans le laboratoire Intelligence Artificielle et Apprentissage Automatique, l'équipe ExpressIF développe le logiciel éponyme.

Notre but est de proposer une intelligence artificielle symbolique pour différentes raisons : une alternative aux approches basées sur les données, la possibilité d'apprendre à partir de peu de données, et le caractère interprétable des modèles et explicable des décisions.

Cela fait 10 ans que nous développons et transférons ExpressIF® à des industriels, tout en y agrégeant les méthodes de l'état de l'art et celles issues de nos propres travaux de recherche. Nous travaillons dans le cadre de projets collaboratifs ou de thèses avec des équipes de recherche comme le MICS de CentraleSupélec ou le LIP6 de Sorbone Université.

Nos travaux s'orientent vers trois principaux aspects de l'intelligence artificielle : l'expressivité, l'apprentissage automatique et l'explicabilité des décisions.

Expressivité

ExpressIF® est un logiciel basé sur les connaissances. Nous formalisons ces connaissances avec un mélange de logique binaire et de logique floue. La logique floue a été un choix dès le départ afin d'obtenir des règles ou des contraintes proches du langage naturel, et par conséquent des modèles plus facilement interprétables. À ce jour, ExpressIF® dispose d'un moteur d'inférence floue et d'un solveur de contraintes floues.

Nous avons implémenté notre propre logiciel dans un but d'extensibilité et de performances [9]. Ces caractéristiques nous permettent d'ajouter de nouvelles relations qui viennent enrichir le vocabulaire utilisé dans les règles et les contraintes.

Nous nous sommes intéressés par exemple aux relations temporelles, avec des moyens d'exprimer la précédence, les variations, la simultanéité, tout en se basant sur des historiques flous (c'est-à-dire qu'il est possible de pondérer les moments du passé qui vont compter lors de la prise de décision) [11].

Naturellement, nous disposons également d'opérateurs spatiaux issus des travaux d'Isabelle BLOCH, que nous avons adapté pour fonctionner aussi bien sur des images traditionnelles que sur des données géographiques 2D [3], puis 2.5D [7].

En combinant ces aspects temporels et spatiaux, nous proposons également des relations spatio-temporelles qui permettent de caracté-

riser les trajectoires d'objets [10].

Chacun de nos projets contribue à l'expansion du vocabulaire qui peut être utilisé par ExpressIF®, renforçant ainsi la facilité de représentation des connaissances, mais également l'interprétabilité des bases de règles et améliorant les explications générées automatiquement.

Apprentissage automatique

Nous nous intéressons depuis quelques années à l'extraction automatique de contraintes ou de règles depuis des données. Même si l'état de l'art est conséquent dans ce domaine, nous bénéficions de nos partenariats pour obtenir des données réelles qui nécessitent des adaptations ou des approches totalement différentes. L'objectif est que notre outil soit accessible à tout industriel disposant de données qu'il souhaite valoriser. D'un point de vue recherche, nous nous intéressons surtout au rapport performance/interprétabilité.

Pour cela, nous nous sommes intéressés, par exemple, à la construction de *features* interprétables. C'est en effet une grosse différence de paradigme entre les approches connexionnistes qui apprennent leurs représentations des données et les modèles à base de règles ou de contraintes qui n'ont pas cette capacité. Nous avons appliqué ces travaux à la recherche fondamentale pour laquelle une compréhension des résultats est nécessaire. Notre étude a montré que les *features* que nous construisons automatiquement ont du sens pour les physiciens [4].

Puisque nous disposons d'une véritable bibliothèque de relations, nous nous sommes intéressés également à l'apprentissage de relations. Nous l'avons appliqué à l'annotation sémantique d'images en apprenant des contraintes spatiales entre régions d'une image. En particulier, nos travaux montrent qu'à partir de 10 images viscérales annotées seulement, ExpressIF® est capable d'annoter les suivantes [8].

La pluralité de nos applications nous permet de multiplier les travaux dans ce domaine et d'évaluer les méthodes existantes sur des données réelles : création de matériaux, reconnaissance de composés chimiques [5], données spécifiques [6], etc..

Nous réalisons que les méthodes que nous développons permettent d'extraire, à partir des données, des connaissances qui sont à la fois bénéfiques pour les experts humains et utilisables par ExpressIF®.

Explicabilité des décisions

Nous nous intéressons à l'explication des décisions prises par ExpressIF®. Pour cela, nous exploitons la trace très complète du moteur d'inférence et du solveur [1, 2]. Nous bénéficions également de l'expressivité de notre système, c'est-à-dire de l'ensemble des relations que l'on met à disposition de nos partenaires.

Nos travaux nous ont mené à nous intéresser à des domaines différents tels que la *Natural Language Generation*, les sciences cognitives et la psychologie pour proposer des explications en langage naturel de plus en plus efficaces [12].

En résumé

Nous sommes une équipe de 5 chercheurs permanents depuis 2022 qui travaille sur l'applicabilité des systèmes d'inférence floue aux problématiques industrielles. Nous accueillons chaque année des doctorants, des post-doctorants, des stagiaires et des apprentis. Nous fournissons à nos partenaires un accès au logiciel ExpressIF® que nous avons développé dans ce but. Notre recherche scientifique nous permet de nous intéresser à de nombreux sujets que nous valorisons par des transferts industriels, en particulier lorsque l'interprétabilité des modèles et l'explicabilité des décisions est nécessaire. Nous proposons ainsi une approche

originale de l'XAI avec comme but principal son utilisabilité par des non-experts.

Références

[1] I. Baaj and J.-Ph. Poli. Natural language generation of explanations of fuzzy inference decisions. In *IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6, 2019.

[2] I. Baaj, J.-Ph. Poli, W. Ouerdane, and N. Maudet. Min-max inference for possibilistic rule-based system. In *IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6, 2021.

[3] L. Boudet, J.-Ph. Poli, L.-P. Bergé, and M. Rodriguez. Situational assessment of wildfires : a fuzzy spatial approach. In *IEEE 32nd Int. Conf. on Tools with Artificial Intelligence (ICTAI)*, pages 1180–1185, 2020.

[4] N. Cherrier, J.-Ph. Poli, M. Defurne, and F. Sabatié. Embedded feature construction in fuzzy decision tree induction for high energy physics classification. In *IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC)*, pages 615–622, 2020.

[5] E. Friedmann, J.-Ph. Poli, O. Hotel, and Ch. Mer-Calfati. Fuzzy classifiers for chemical compound recognition from saw sensors signals. In *IEEE 32nd Int. Conf. on Tools with Artificial Intelligence (ICTAI)*, pages 917–922, 2020.

[6] A. Grivet Sébert and J.-Ph. Poli. Material classification from imprecise chemical composition : Probabilistic vs possibilistic approach. In *IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8, 2018.

[7] C. Iphar, L. Boudet, and J.-Ph. Poli. Topography-based fuzzy assessment of runoff area with 3D spatial relations. In *IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6, 2021.

[8] R. Pierrard, J.-Ph. Poli, and C. Hudelot. Spatial relation learning for explainable image classification and annotation in critical applications. *Artificial Intelligence*, 292 :103434, 2021.

[9] J.-Ph. Poli and L. Boudet. A fuzzy expert system architecture for data and event stream processing. *Fuzzy Sets and Systems*, 343 :20–34, 2018. Special Issue : Fuzzy Logic and Applications, Selected Papers from the French Fuzzy Set Conference LFA 2015.

[10] J.-Ph. Poli, L. Boudet, and J.-M. Le Yaouanc. Online spatio-temporal fuzzy relations. In *IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8, 2018.

[11] J.-Ph. Poli, L. Boudet, and D. Mercier. Online temporal reasoning for event and data streams processing. In *IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE)*, pages 2257–2264, 2016.

[12] J.-Ph. Poli, W. Ouerdane, and R. Pierrard. Generation of textual explanations in XAI : the case of semantic annotation. In *IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6, 2021.

AfIA

Association française
pour l'Intelligence Artificielle

Comptes rendus de journées, événements et conférences

AfIA
Association française
pour l'Intelligence Artificielle

3e Journée Réalité Virtuelle et Intelligence Artificielle

Par **Domitile LOURDEAUX**
Heudiasyc
Université de technologie de Compiègne
domitile.lourdeaux@hds.utc.fr

Indira THOUVENIN
Heudiasyc
Université de technologie de Compiègne
indira.thouvenin@utc.fr

Introduction

Le GDR Informatique Graphique et Réalité Virtuelle (IG-RV) et l'Association Française d'Intelligence Artificielle (AFIA) au travers de son Collège Interaction avec l'Humain (IAH), ont organisé une journée commune sur le thème « Environnements virtuels : adaptation du système à l'humain ou de l'humain au système ? » le 9 mars 2022 à l'Université de technologie de Compiègne (UTC), avec le soutien du laboratoire Heudiasyc.

Cette journée était une journée scientifique autour de sujets à l'intersection des deux domaines de recherche que sont l'Intelligence Artificielle (IA) et la Réalité Virtuelle (RV), mettant en évidence des liens possibles entre les deux disciplines. La journée a eu lieu en mode hybride avec 28 participants en présentiel et jusqu'à 27 autres participants en distanciel.

L'objectif de cette journée était de réunir des actrices et des acteurs du domaine, afin d'aborder des questions scientifiques, technologiques, ou des questions portant sur les facteurs humains et les usages. Les participants et participantes étaient issus aussi bien du monde académique que du monde industriel, permettant de confronter différentes approches et différents domaines d'application. Le format de cette journée a offert un contexte opportun pour mettre en commun les expérience- et réflexions sur les approches actuelles, sur les challenges et les perspectives de recherche, au travers de présentations invitées, de démonstrations et de contributions directes.

Programme

9h30. Accueil et café
10h00. « Ouverture », par Domitile LOURDEAUX (AFIA), Indira THOUVENIN (GDR IG-RV) et Philippe BONNIFAIT (directeur de l'URM 7253)
10h15. « Pourquoi cette journée ? Environnements virtuels : adaptation du système à l'humain ou de l'humain au système ? », par Domitile LOURDEAUX (Heudiasyc, Compiègne) et Indira THOUVENIN (Heudiasyc, Compiègne)
10h35. « Les agents virtuels dans les environnements immersifs d'apprentissage », par David PANZOLI (IRIT, Toulouse)
11h05. Pause-café
11h15. « La réalité virtuelle en tant qu'outil de formation adapté aux pratiques de l'enseignant : scénarisation des travaux pratiques, enseignement du geste et adaptation au comportement », par Ludovic HAMON (LIUM, Laval)
11h45. « Facteurs humains et réalité virtuelle : détecter les effets secondaires avec capteurs physiologiques et machine learning », par Alexis SOUCHET (Heudiasyc, Compiègne)
12h30. Buffet
14h00. Visite du CAVE et démonstrations

- Projet KIVA (formation au geste technique pour la fonderie d'aluminium) par Yohan BOUVET et Indira THOUVENIN (Heudiasyc)
- Projet ORCHESTRAA (formation au commandement d'opérations aériennes OTAN) par Romain LELONG (Reviatech)

- Projet VICTEAMS-Stress (formation de leaders médicaux à la gestion d'un afflux massif de blessés) par Luca PELISSERO-WITOSLAWSKI (Heudiasyc)

15h00. « Environnements virtuels et modèles de décision pour l'interaction », par Marc MACÉ (IRISA, Rennes)

15h30. « Toucher social pour l'interaction humain-agent incarné en environnement virtuel », par Fabien BOUCAUD (ISIR, Paris)

15h50. « Retours adaptatif en réalité augmentée basés sur l'état du conducteur de véhicule hautement automatisé lors de la reprise de contrôle », par Baptiste WOJTKOWSKI (Heudiasyc, Compiègne)

16h15. « Table ronde (Environnement virtuels : adaptation du système à l'humain ou de l'humain au système ?) », par Philippe CHOPIN (Thalès), Romain LELONG (Reviatech), Domitile LOURDEAUX (Heudiasyc), David PANZOLI (IRIT), Alexis SOUCHET (Heudiasyc) et Indira THOUVENIN (Heudiasyc)

17h00. Clôture de la journée

Bilan scientifique

Les présentations étaient riches et variées et ont abordé différentes facettes de l'adaptation des environnements virtuels à l'humain (feedbacks adaptatifs, modélisation de l'utilisateur, scénarisation adaptative, etc.). Les orateurs ont essayé de répondre à la question posée par le thème de la journée et leurs points de vue était très pertinents. Les auditeurs et auditrices, tant en présentiel qu'en distanciel, ont soulevé des débats passionnants qui nous ont permis d'avancer sur nos questions de recherche. Les participants et participantes étaient très satisfaits de cette journée.

La journée a réuni des académiques, des industriels, des étudiants ainsi que différentes disciplines (informatique, neuroscience, ergonomie, psychologie, automatique).

Le format en hybride a permis à un plus grand nombre de personnes de pouvoir profiter de cette journée, même si ce format ne permet pas les échanges moins formels et ne favorise pas l'émulsion créative de futures et riches collaborations autour de nouvelles questions de recherche.

Journée Commune AfIA & Réseau DEVS

Par

Fabien MICHEL
LIRMM / SMILE
Université de Montpellier
fmichel@lirmm.fr

Paul-Antoine BISGAMBIGLIA
UMR CNRS SPE / UMS Stella Mare
Université de Corse
bisgambiglia@univ-corse.fr

Introduction

Le lundi 28 mars 2022, l'AfIA et le réseau DEVS (RED) – soutenu par le département MIAT de l'INRAe et l'UMR CNRS SPE – ont conjointement organisé une journée dédiée à la simulation pour l'intelligence artificielle et l'intelligence artificielle pour la simulation. Cette journée commune a pris place dans le programme de l'atelier du réseau RED : les journées francophones de la modélisation et de la simulation.

L'objectif de la journée était de mettre en avant les convergences entre les domaines de la simulation et de l'intelligence artificielle et ainsi voir comment les travaux de chaque domaine peuvent s'imbriquer. Le programme a été construit pour montrer ces liens sur des exemples concrets d'applications.

Il y a eu plus de 70 participants à distance ou présents à l'Institut des Études Scientifiques de Cargèse. Les supports et vidéos seront bientôt disponibles sur le site de l'AfIA. Chaque intervention a été ponctuée de remarques et questions, les organisateurs souhaitent remercier tous les participants pour la qualité des interventions et des discussions.

Programme

9h30. « Ouverture », par Emmanuel ADAM (AfIA), Fabien MICHEL (LIRMM) et Pierre-Antoine BISGAMBIGLIA (Réseau DEVS)
9h45. « Simulation des systèmes multi-robots : outils et enjeux », par Olivier SIMONIN (CITI)
10h45. « Apprentissage automatique pour l'amélioration de la qualité de modèles », par Grégory BEURIER (CIRAD)
11h45. Pause-café
12h00. « Vers des équipes humains-IAs : Écosystèmes d'intelligence pour cas d'utilisation à fort enjeu », par Clodéric MARS (AI Redefined)
14h30. « Simulations and risk, from simulations in simulation to agent-based model and reinforcement learning », par Arthur CHARPENTIER (Université du Québec)
15h30. « SCAMP : A Stigmergic Approach to Modeling Intelligent Behavior », par H. Van Dyke PARUNAK (Parallax Advanced Research)
16h30. Pause-café
17h00. « Facing complexity with self-organization », par Carlos GERSHENSON (Université Nationale Autonome du Mexique)
18h00. Clôture de la journée

8e Journée Perspectives et Défis de l'IA

Par

Fayçal HAMDI
CEDRIC
CNAM Paris

Engelbert MEPHU NGUIFO
LIMOS
Université Clermont Auvergne

Davy MONTICOLO
ERPI
Université de Lorraine

Fatiha SAÏS
LISN
Université Paris Saclay

Introduction

Dix ans après la parution dans le Bulletin de l'AfIA, numéro 78 d'octobre 2012, du dossier consacré à la créativité et à l'innovation, coordonné par Eunika MERCIER-LAURENT et François PACHET, l'AfIA a choisi le thème de la créativité pour sa journée Perspectives et Défis de l'Intelligence Artificielle (PDIA 2022) du 7 avril 2022.

En effet, nous assistons aujourd'hui à une multiplication des usages des technologies d'intelligence artificielle pour la résolution de nombreux et difficiles problèmes, parmi lesquelles ceux en lien avec la créativité humaine. Par exemple, le développement récent des techniques d'apprentissage profond ont permis la génération automatique d'œuvres d'art dont certaines rivalisent celles d'artistes de renommée.

PDIA 2022 a réuni des scientifiques et créateurs ayant abordé le problème difficile de l'usage des algorithmes d'IA pour la créativité qui par nature mobilisent des capacités hu-

maines et cognitives qui est de premier abord difficile à transcrire dans un programme.

La journée est construite autour d'exposés accessibles et de retours d'expériences favorisant une grande interaction. Cette journée a permis à des chercheurs académiques et industriels, d'avoir des échanges sur les progrès effectués durant la dernière décennie autour de cette thématique. 52 chercheurs étaient inscrits à cette journée dont 18 en présentiel.

Programme

9h15. « Présentation de l'AfIA », par Benoît LE BLANC (AfIA)
9h20. « Ouverture de la journée », par Daviy MONTICOLO (AfIA)
9h30. « Repenser l'interaction avec les technologies d'apprentissage », par Baptiste CARAMIAUX (ISIR, Sorbonne Université, HCI Sorbonne Group)
10h30. « Intelligence artificielle pour assister l'idéation et la conception amont », par Alex GABRIEL (ERPI, Université de Lorraine)
11h30. Pause-café
11h45. « Machines à écrire : créer des programmes qui créent pour apprendre à se servir de l'IA », par Anne-Gwenn BOSSER (STICC, ENIB, Université de Brest Bretagne Loire)
12h45. Pause déjeuner
11h45. « Angelia – une intelligence artificielle pour la musique électronique », par Jean-Claude HEUDIN (Chercheur en IA, écrivain et compositeur)
11h45. « Musique et « IA » pour « Instruments Artificiels » », par Jérôme NIKA (IRCAM)
16h00. Pause-café
16h15. « Quelques réflexions sur la création musicale assistée par l'IA », par François PACHET (Spotify Creator Technology Research Lab)
17h15. Clôture de la journée

Résumé des interventions

Repenser l'interaction avec les technologies d'apprentissage, par Baptiste CARAMIAUX, chercheur CNRS au laboratoire ISIR, Sorbonne Paris Université, membre du HCI Sorbonne Group

« Les algorithmes d'apprentissage machine sont présents dans un grand nombre d'applications et de services qu'on utilise au quotidien. Ces technologies sont, par design, conçues de manière dissociée de leurs utilisateurs, ce qui entraîne une normalisation de leurs utilisations et un contrôle centralisé de leurs capacités. Créer des technologies d'apprentissage plus près des personnes et de leur contexte d'utilisation ouvre le champ à des interactions plus adaptées, appropriables et inclusives. Dans cet exposé, je présenterai le contexte et la communauté de recherche qui travaille sur ces thématiques à l'intersection entre IHM et IA. Ensuite, je mettrai l'accent sur mes travaux dans le domaine artistique. Je montrerai des exemples de recherche où l'approche artistique est parfois vu comme outil de réflexion sur les technologies en tant qu'acteurs culturels, et parfois vu comme outil d'inspiration pour la conception d'interactions riches et expressives. »

Intelligence artificielle pour assister l'idéation et la conception amont, par Alex GABRIEL, chercheur postdoctoral au laboratoire ERPI, Université de Lorraine

« Les différents domaines de l'intelligence artificielle permettent de réaliser un nombre toujours plus important d'innovations technologiques pour faciliter l'activité humaine. Pour autant, avant qu'une solution innovante soit mise sur le marché, celle-ci aura subi de multiples modifications et évolutions depuis l'idée originale. Les organisations mettent en œuvre diverses pratiques pour promouvoir et favoriser la production d'idées, notamment au travers

de processus créatifs. À l'instar d'autre secteur, les processus créatifs et d'innovation possèdent également des outils numériques supports. Le secteur est d'ailleurs en forte croissance et il existe une offre importante d'outils. Ces Innovation/Idea Management System font l'objet de recherche depuis plus d'une vingtaine d'années. Cette présentation fera le point sur les fonctionnalités de ces outils et l'application d'ontologies et de traitement automatique du langage dans ce contexte de gestion de la créativité. »

Machines à écrire : créer des programmes qui créent pour apprendre à se servir de l'IA, par Anne-Gwenn BOSSER, maîtresse de conférences au laboratoire STICC, ENIB, Université de Brest Bretagne Loire

« Les ateliers de programmation collaboratifs sont des événements populaires. La nuit de l'informatique, par exemple, rassemble tous les ans de nombreux établissements d'enseignement supérieur francophones, et l'AfIA y participe au travers d'un défi qu'elle propose. Au niveau international, un événement comme la Global Game Jam peut rassembler des dizaines de milliers de participantes et de participants le temps d'un week-end. Lors de tels événements, la créativité, la collaboration et le partage de connaissances sont mis à l'honneur dans un contexte ludique. Le collège Cécilia de l'AfIA proposera ainsi aux participantes et aux participants de PFIA 2022 une « Jam de création de textes poétiques ou drôles (ou les deux) ». Nous présenterons cet événement et ses nombreuses inspirations comme l'Oulipo ou la machine à écrire de Jean Baudot. Nous montrerons au travers d'un état de l'art que l'exercice se prête bien à l'utilisation d'une variété de techniques d'IA, en faisant un prétexte riche à l'apprentissage de nouvelles techniques. »

Angelia – une intelligence artificielle pour la musique électronique, par Jean-Claude HEUDIN, chercheur en IA, écrivain et compositeur

« Les récents progrès en Intelligence Artificielle ont permis des avancées spectaculaires dans de nombreux domaines. Moins médiatisées pour la musique, les applications de l'IA n'en sont pas moins importantes. Elles suscitent dès lors de nombreuses interrogations : l'IA peut-elle égaler les meilleurs compositeurs ? Va-t-elle un jour remplacer les artistes ? Est-ce le futur de la musique ? Pour répondre à ces questions, nous retraçons brièvement l'histoire de l'IA en musique, puis avec Angelia, une IA dédiée à la musique électronique, nous mettrons en évidence ses enjeux et perspectives. »

Musique et « IA » pour « Instruments Artificiels », par Jérôme NIKA, chercheur à l'IRCAM

« Une machine sera-t-elle bientôt capable de remplacer l'humain dans la création musicale ? Pour toute une partie des artisans de l'intelligence artificielle appliquée à la musique, artistes comme scientifiques, il est difficile de répondre à cette question récurrente. . . car ce n'est pas celle qui se pose. Si on « apprend » la musique à des ordinateurs dotés d'une « mémoire » musicale inspirée de la cognition humaine, l'enjeu réside précisément dans le fait de partir de ces modèles pour explorer la production d'une musique nouvelle plutôt que la reproduction d'une musique crédible. La présentation des pratiques musicales permises par ces instruments d'une nouvelle génération, au service de la créativité humaine, sera illustrée par des extraits de productions récentes. »

Quelques réflexions sur la création musicale assistée par l'IA, par François PACHET, directeur du Spotify Creator Technology Research Lab

« Peut-on concilier l'intelligence artificielle (IA), qui implique technicité et rigueur scientifique, avec la création musicale qui fait appel à la sensibilité et la créativité ? François Pachet est compositeur et directeur du Spotify Creator Technology Research Lab où il conçoit la prochaine génération d'outils pour les musiciens basés sur l'intelligence artificielle. Au travers de son expérience, à la fois de musicien et de scientifique, François Pachet aborde toutes les potentialités que l'IA peut offrir aux artistes sans dénaturer pour autant le plaisir de la création. Avec son label Flow Records, il a récemment produit et publié un album intitulé Hello World, premier opus musical composé avec une intelligence artificielle, fruit de la collaboration avec de nombreux artistes comme Stromae, AI, Benoit Carré, alias SKYGGE, Médéric Collignon (le jazzman aux trois Victoires de la musique)... »

AfIA
Association française
pour l'Intelligence Artificielle

Thèses et HDR du trimestre

Si vous êtes au courant de la programmation de soutenances de thèses ou HDR en Intelligence Artificielle cette année, vous pouvez nous les signaler en écrivant à redaction@afia.asso.fr.

Thèses de Doctorat

Ygor GALLINA

« Indexation de bout-en-bout dans les bibliothèques numériques scientifiques »

Supervision : *Béatrice DAILLE*
Florian BOUDIN

Le 28/03/2022, à l'Université de Nantes

Léon Paul SCHAUB

« Dimensions mémorielles de l'interaction écrite humain-machine : une approche cognitive par les modèles mnémoniques pour la détection et la correction des incohérences du système dans les dialogues orientes-tâche »

Supervision : *Patrick PAROUBEK*
Gil FRANCOPOULO
Samuel RUMEUR

Le 22/03/2022, à l'Université Paris-Saclay

Balthazar DONON

« Deep statistical solvers & power systems applications »

Supervision : *Isabelle GUYON*
Marc SCHOENAUER
Remy CLEMENT

Le 16/03/2022, à l'Université Paris-Saclay

Julien GÉRARD

« Drone recognition with deep learning »

Supervision : *Joanna TOMASIK*
Christèle MORISSEAU
Arpad RIMMEL

Le 16/02/2022, à l'Université Paris-Saclay

Jean Yves FRANCESCHI

« Apprentissage de représentations et modèles génératifs profonds dans les systèmes dynamiques »

Supervision : *Patrick GALLINARI*
Sylvain LAMPRIER

Le 14/02/2022, à Sorbonne Université

Marvin LASSERRE

« Apprentissages dans les réseaux bayésiens à base de copules non-paramétriques »

Supervision : *Christophe GONZALES*

Le 11/03/2022, à Sorbonne Université

Remy PORTELAS

« Automatic curriculum learning for developmental machine learners »

Supervision : *Pierre Yves OUDEYER*
Katja HOFMANN

Le 11/02/2022, à l'Université de Bordeaux

Roman BRESSON

« Neural learning and validation of hierarchical multi-criteria decision aiding models with interacting criteria »

Supervision : *Johanne COHEN*
Christophe LABREUCHE

Le 02/02/2022, à l'Université Paris-Saclay

Habilitations à Diriger les Recherches

Nous n'avons malheureusement pas eu connaissance ce trimestre d'HDR dans le domaine de l'IA.

N'hésitez pas à nous envoyer les informations concernant celles dont vous avez entendu parler. (redaction@afia.asso.fr).

AfIA
Association française
pour l'Intelligence Artificielle

À PROPOS DE
L'AfIA

L'objet de l'AfIA, Association Loi 1901 sans but lucratif, est de promouvoir et de favoriser le développement de l'Intelligence Artificielle (IA) sous ses différentes formes, de regrouper et de faire croître la communauté française en IA et, à la hauteur des forces de ses membres, d'en assurer la visibilité.

L'AfIA anime la communauté par l'organisation de grands rendez-vous. Se tient ainsi chaque été une semaine de l'IA, la « Plate-forme IA » (PfIA 2020 à Angers, PfIA 2021 à Bordeaux, PfIA 2022 à Saint-Étienne) au sein de laquelle se tiennent la Conférence Nationale d'Intelligence Artificielle (CNIA), les Rencontres des Jeunes Chercheurs en IA (RJCIA) et la Conférence sur les Applications Pratiques de l'IA (APIA) ainsi que des conférences thématiques hébergées qui évoluent d'une année à l'autre, sans récurrence obligée.

Ainsi, PfIA 2022 héberge du 27 juin au 1er juillet 2022 à Saint-Étienne, outre la 25e CNIA, les 20es RJCIA et la 8e APIA : les 33es IC, les 17es JFPC, les 17es JFPDA, les 30es JFSMA et les 16es JIAF, 4 journées thématiques hébergées (EIAH & IA, IoT & IA, Résilience & IA, Santé & IA), et plusieurs tutoriels hébergés.

Forte du soutien de ses 304 adhérents à jour de leur cotisation en 2021, l'AfIA assure :

- le maintien d'un site Web dédié à l'IA reproduisant également les Brèves de l'IA ;
- une *journée industrielle* « Forum Industriel en IA » (FIIA 2021) ;
- une *journée recherche* « Perspectives et Défis en IA » (PDIA 2021) ;
- une *journée enseignement* « IA pour l'enseignement » (EFIA 2022) ;
- la remise annuelle d'un prix de thèse en IA ;
- le soutien à 8 collèges ayant leur propre activité :
  - collège Industriel (depuis janvier 2016) ;
  - collège Apprentissage Artificiel (depuis janvier 2020) ;
  - collège Interaction avec l'Humain (depuis juillet 2020) ;
  - collège Représentation et Raisonnement (depuis avril 2017) ;
  - collège Science de l'Ingénierie des Connaissances (depuis avril 2016) ;
  - collège Systèmes Multi-Agents et Agents Autonomes (depuis octobre 2016) ;
  - collège Technologies du Langage Humain (depuis juillet 2019) ;
  - collège Création d'Événements Collaboratifs, Inclusifs et Ludiques en IA (depuis octobre 2021) ;
- la parution trimestrielle des Bulletins de l'AfIA ;
- un lien entre ses membres et sympathisants sur les réseaux sociaux LinkedIn, Facebook et Twitter ;
- le parrainage scientifique, mais aussi éventuellement financier, d'événements en IA ;
- la diffusion mensuelle de Brèves sur les actualités de l'IA en France (abonnement ou envoi à la liste) ;
- la réponse aux consultations officielles ou officieuses (Ministères, Missions, Organismes) ;
- la réponse aux questions de la presse, écrite ou orale, également sur internet ;
- la divulgation d'offres de collaborations, de formations, d'emploi, de thèses et de stages.

L'AfIA organise aussi des journées communes avec d'autres associations. Pour 2021 : Jeux & IA avec le GDR IA; Santé & IA avec AIM; Défense & IA avec ONERA; Classification & IA avec SFC.

Enfin, l'AfIA encourage la participation de ses membres aux grands événements de l'IA, dont PfIA. Ainsi, les membres de l'AfIA, pour leur inscription à PfIA, bénéficient d'une réduction équivalente à deux fois le coût de leur adhésion, leur permettant d'assister à PfIA 2022 sur 5 jours au tarif de 114 € TTC !

Rejoignez-nous vous aussi et adhérez à l'AfIA pour contribuer au développement de l'IA en France. L'adhésion peut être individuelle ou au titre de personne morale. Merci également de susciter de telles adhésions en diffusant ce document autour de vous !

AfIA
Association française
pour l'Intelligence Artificielle

Pour contacter l'AfIA

Président
Benoit LE BLANC
École Nationale Supérieure de Cognitique
Bordeaux-INP
109 avenue Roul, 33400 Talence
Tél. : +33 (0) 5 57 00 67 00
president@afia.asso.fr

Serveur WEB
http://www.afia.asso.fr

Adhésions, liens avec les adhérents
Isabelle SESÉ
tresorier@afia.asso.fr

Calendrier de parution du Bulletin de l'AfIA

	Hiver	Printemps	Été	Automne
Réception des contributions	15/12	15/03	15/06	15/09
Sortie	31/01	30/04	31/07	31/10