Réponse à appel à projet doctoral SCAI

Explanation Games

Présentation du sujet

Contexte

Le besoin d'explication dans les systèmes à base d'intelligence artificielle (IA) est devenu incontournable pour leur adoption auprès du grand public. Que l'on s'intéresse à de systèmes à base de réseaux neuronaux, de méthodes d'agrégations issues de la théorie de la décision, voire de systèmes à base de règles, toute décision prise par un système d'IA doit pouvoir être comprise. Le RGPD introduit la notion de redevabilité (*accountability*) qui désigne l'obligation pour les entreprises de mettre en œuvre des mécanismes et des procédures internes permettant de démontrer le respect des règles relatives à la protection des données. Enfin les problèmes d'équité restent essentiels.

Aussi, ces dernières années ont vu l'explosion du champ de recherche de l'intelligence artificielle explicable (*eXplainable Artificial Inteligence* – XAI). Celui-ci est l'un des plus actifs et des plus cruciaux des champs de recherche actuels liés à l'IA. Les formes que peuvent prendre les explications sont variées : elles peuvent être agnostiques au modèle (comme par exemple les explications contre-factuelles) ou dépendante au modèle (ce que permettent plus facilement les modèles symboliques ou hybrides). Néanmoins dans quasiment toutes ces approches, les explications sont envisagées de manière individuelle, alors que le caractère situé et social de l'échange explicatif a été souligné depuis longtemps **[Mil17]**, entraînant par là la nécessité de porter l'analyse dans un cadre dialectique.

Dans le cadre de ce sujet de thèse, on distinguera 2 groupes pouvant avoir des objectifs différents : les *explainers* (ceux qui expliquent, qui doivent rendre des comptes) et les *explainees* (à qui l'on doit expliquer, à qui l'on doit rendre des comptes). Ces 2 groupes peuvent avoir des objectifs différents et être collaboratifs, *adversarials* (contradictoire, antagoniste) voire intermédiaires. L'exemple le plus évident d'antagonisme est le cas du *fair washing*, c'est-à-dire lorsque le groupe des *explainers* cherche à cacher les raisons inavouables de leur décision aux *explainees* en leur fournissant des explications fallacieuses. Mais l'inverse peut également se produire : les *explainees* peuvent chercher à obtenir un maximum d'explications et ainsi obtenir des informations confidentielles auxquelles ils ne devraient pas avoir accès ou encore récolter suffisamment d'informations pour leur permettre de détourner le système de décision à leurs fins.

Le cadre collaboratif peut se trouver lorsque les *explainers* veulent s'assurer que les *explainees* ont bien compris les explications données. C'est un cas typique en médecine où le médecin doit s'assurer que son diagnostic et les conséquences de celui-ci sont accessibles au patient.

Objectif scientifique

L'objectif scientifique de ce projet est de s'appuyer sur et d'étendre différents formalismes permettant de créer un système permettant :

1. De tester et de valider des moteurs d'explications au moyen de moteurs de jeu afin de tenter de les mettre en défaut et ainsi de mettre en exergue le caractère potentiellement injuste du

système d'explication. Par exemple, si le système présente les caractéristiques d'un candidat lui aussi permettant de justifier le refus d'un autre candidat par un jury (comme par exemple dans [**BCL18**]), certaines formes de transitivités sur les caractéristiques des candidats doivent être préservées.

2. De tester à quel point les requêtes d'un *explainee* sont fondées et n'essayent pas de mettre en défaut le système, de remettre systématiquement en cause la décision prise pour le système d'IA, ou encore de l'attaquer afin d'avoir des informations qu'il n'a pas le droit d'avoir.
3. À un niveau *meta*, le système doit permettre de tester des protocoles de validations des API d'explication et éventuellement soit les remettre en cause, soit permettre l'aboutissement d'un consensus.

Cadre formel et justification de l'approche scientifique

L'idée principale de ce sujet de thèse est de représenter les mécanismes d'explication et de persuasion via une représentation logique des jeux et l'argumentation. La logique peut être utilisée pour déterminer ce qui accepté par chaque agent participant à la discussion sur la base des justifications présentées pendant l'argumentation. Le protocole d'argumentation doit lui-même pouvoir être décrit et remis en cause lors de l'argumentation. Certains systèmes tels que ceux proposés dans [**Bre01**] ont de telles propriétés. Des analogies peuvent également être trouvées avec de récents travaux traitant des règles de votes non décidées.

Afin de capturer l'aspect stratégique du jeu d'explication, il est nécessaire de le modéliser à l'aide de langages déclaratifs utilisés pour le jeu, tels que ceux de la famille de GDL (Game Description Language) [**LGH06,Thie17**]. Les jeux modélisés par ces langages, Turing complets, peuvent être joués par des moteurs de jeu génériques (par exemple [**KLP17**]).

Validation

Il semble tout à fait naturel de faire intervenir des laboratoires de sciences humaines afin de valider le système développé. Parmi eux, il est naturel de penser au le laboratoire Costech de l'UTC qui s'intéresse plus particulièrement aux relations "Homme - Technique – Société".

References

[**BDKM21**] E. Bonzon, J. Delobelle, S. Konieczny, **N. Maudet**, "A parametrized ranking-based semantics compatible with persuasion principles", Argument Comput. 12(1): 49-85 (2021).

[**Bre01**] G. Brewka, "Dynamic Argument Systems: A Formal Model of Argumentation Processes Based on Situation Calculus," in *Journal of Logic and Computation*, vol. 11, no. 2, pp. 257-282, 2001.

[**BCL18**] **K. Belahcène**, Y. Chevaleyre, C. Labreuche, **N. Maudet**, V. Mousseau, W. Ouerdane, "Accountable Approval Sorting", in Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI 2018): 70-76.

[**KLP17**] F. Koriche, **S. Lagrue**, É. Piette, S. Tabary, "Constraint-Based Symmetry Detection in General Game Playing". In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017): 280-287.

[**LGH06**] N. Love; M. Genesereth; T. Hinrichs, "General game playing: game description language specification. Tech. Rep. LG-2006-01". Stanford University. 2006.

Stephan Schiffel, Michael Thielscher: "Reasoning About General Games Described in GDL-II" In Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence. AAAI 2011.

[Mill17] Tim Miller. Explanation in Artificial Intelligence: Insights from the Social Sciences. Art. Int.

Journal. 2017
[**RE21**] S. Rey, U. Endriss, Ronald de Haan, "Shortlisting Rules and Incentives in an End-to-End Model for Participatory Budgeting". In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI 2021): 370-376
[**Thie17**] M. Thielscher, "GDL-III: A Description Language for Epistemic General Game Playing" in Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. IJCAI 2017.

Autour du sujet

Adéquation à l'initiative SCAI

Le projet porte clairement sur le domaine d'excellence *Mathematics*, *Computer Science & Robotics* de SCAI, en particulier sur les thématiques de systèmes multi-agents, de prise de décision et d'intelligence artificielle explicable. Elle fait intervenir deux laboratoires de l'alliance Sorbonne université (le LIP6 et Heudiasyc) dont la recherche est reconnue dans le domaine de l'IA.

Complémentarité de l'équipe encadrante

Khaled Belahcène est un maître de conférence à l'UTC, membre du laboratoire Heudiasyc, et spécialiste des problèmes d'explication et de recevabilité des algorithmes de décision.

Sylvain Lagrue est professeur des universités à l'UTC, membre du laboratoire Heudiasyc, et spécialiste de représentation logique des connaissances et des raisonnements, en gestion de l'incertain en IA, et en prise de décision dans les jeux.

Nicolas Maudet est professeur des universités à Sorbonne Université, membre du laboratoire LIP6, et spécialiste en systèmes multi-agents, en choix social computationnel, et en argumentation.

Profil recherché

Le/la candidat·e idéal·e est en possession d'un master 2 ou diplôme d'ingénieur, en informatique ou en mathématiques appliquées. Il/elle possède une appétence pour les problèmes théoriques et pratiques, mais aussi sociétaux. Il/elle disposera d'un bon niveau de programmation et de bonnes capacités de synthèse et de formalisation.

On attendra de la personne recrutée une autonomie certaine et qu'elle soit capable de prises d'initiatives et de fortes aptitudes au travail en équipe. Un bon niveau d'anglais et de français est requis (au moins B2 dans chacune des langues).