Explicabilité des protocoles d'argumentation

L'étude des systèmes multi-agents (SMA) est un domaine de recherche qui s'est largement développé en IA ces dernières années. Les SMA permettent d'étudier formellement des systèmes dans lesquels plusieurs agents (rationnels) interagissent afin d'atteindre un ou plusieurs objectifs (éventuellement antagonistes). Ces interactions peuvent prendre différentes formes. Dans cette thèse, nous nous intéressons au cadre de la théorie de l'argumentation formalisant les échanges d'informations entre agents par des arguments. Ces arguments peuvent être en conflit les uns avec les autres lorsque les agents ont des croyances différentes. La théorie de l'argumentation étudie comment établir des conclusions à partir des arguments avancés par les agents. L'argumentation est ainsi au cœur des processus de délibération.

Nous considérons un cadre dans lequel chaque agent est muni d'un système d'argumentation décrivant ses arguments et leurs relations. Nous nous intéressons en particulier aux protocoles d'argumentation permettant de réguler, dans des débats, les échanges d'arguments entre des agents ayant chacun leur propre opinion (arguments, croyances, objectifs, etc.). Les agents contribuent au débat étape par étape, guidés par leur évaluation individuelle de l'état actuel de la discussion, et sans coordination avec d'autres agents. Plusieurs protocoles ont été définis dans la littérature afin d'organiser les échanges [1,2].
L'un des objectifs de tels débats argumentatifs est de permettre une prise de décision consensuelle entre des agents ayant des points de vue différents. Ces échanges sont également susceptibles de permettre l'apprentissage et éventuellement l'adoption de nouvelles croyances (arguments) de la part des autres agents.
Une application directe de ces travaux est liée à la prise de décision collective dans des sociétés d'agents (aménagements dans une ville, financements de projets au sein d'une université ou d'un laboratoire...). Ces dernières années, plusieurs systèmes de débat se sont développés sur Internet afin de permettre à des utilisateurs d'échanger des opinions (par exemple Debatepedia et Debategraph). Le succès de ces plates-formes, dans leur forme actuelle, semble suggérer qu'elles peuvent devenir une source d'échange et d'information importante, tout comme Wikipedia l'est actuellement.

Récemment, la notion d'intelligence artificielle explicable (XAI) a connu un regain d'attention de la part des chercheurs. Cette résurgence est motivée par le fait que beaucoup d'applications d'IA ont une utilisation limitée ou ne sont pas du tout appropriées, en raison de préoccupations éthiques et d'un manque de confiance de la part de leurs utilisateurs. L'explicabilité en argumentation reste un sujet très peu étudié. Pourtant pour permettre l'adoption par les utilisateurs de systèmes à base d'argumentation, il est indispensable de pouvoir expliquer les conclusions des débats. Il est en effet nécessaire d'être capable de justifier la décision proposée ou le résultat retourné dans la plupart des applications faisant appel à des techniques utilisant l'argumentation ; notamment lorsque ces applications sont destinées à un public non-spécialisé.
Cependant, à notre connaissance, aucun travail n'a été fait pour permettre d'expliquer, aussi bien à un public expert ou non-expert en argumentation, l'acceptabilité de l'issue d'un débat. Un des objectifs de cette thèse est aussi de définir des méthodes permettant d'expliquer les différents résultats retournés par les protocoles d'argumentation qui seront étudiés.

Les problématiques qui seront abordées dans cette thèse peuvent être décomposées en trois parties :

1. Le premier volet de la thèse consistera à enrichir les protocoles argumentatifs existants en permettant notamment aux agents de voter pour ou contre des arguments. L'ajout d'un mécanisme de vote permettrait de se rapprocher des protocoles existants dans les plateformes développées en ligne et qui autorisent pour la plupart aux utilisateurs de voter pour ou contre un argument donné. De tels mécanismes de vote posent de nombreuses questions en particulier au niveau de la sémantique utilisée : la signification des votes est souvent peu claire, et peut varier fortement en fonction de la plateforme considérée. Un vote exprime-t-il qu'un argument est valide, qu'il est pertinent, ou qu'il doit être accepté ? Toutes ces interprétations ont un sens et peuvent donner lieu à des sémantiques différentes. Des questions sont également soulevées au sujet du protocole en lui-même : par exemple, le vote est-il autorisé juste après la présentation d'un argument, ou seulement à la toute fin du débat, une fois que tous les arguments ont été donnés ? Les utilisateurs peuvent-ils supprimer des votes ?
   Les questions liées à la prise de décision stratégique dans les arguments avancés par les agents [3] ou dans les votes pourront également être étudiées.
2. Dans le but d'expliquer les résultats d'un protocole, il sera nécessaire de répertorier et de quantifier l'impact des éléments impliqués dans l'acceptabilité des arguments et l'issue du débat. Ces éléments concernent les arguments via les relations existantes (attaque ou soutien), les votes (positifs ou négatifs) attribués par les utilisateurs sur les arguments et/ou les attaques, mais aussi, si l'information est disponible, les données sur les agents participant au débat. Étendre ou définir des méthodes basées sur les indices de pouvoir (par exemple la valeur de Shapley) pourrait être utile pour définir plusieurs familles de méthodes d'évaluation de l'influence de ces différents éléments.
3. Le dernier volet de la thèse aura pour but de tester et d'évaluer aussi bien les protocoles que les méthodes d'explication sur des données réelles (débats en ligne dont les données sont disponibles). Cela nécessitera de fournir, de manière automatique (ou semi-automatique), de "bonnes" explications concernant les résultats issus des protocoles d'argumentation. Une piste à suivre proviendrait du travail de Tim Miller [4] qui regroupe et analyse de nombreux travaux existants issus des sciences sociales portant sur le thème de l'explicabilité entre humains.

Encadrement

Le doctorant sera accueilli au LIP6 (équipe SMA), Sorbonne Université, sous la direction de Aurélie Beynier (MCF, HDR au LIP6, Sorbonne Université), Elise Bonzon (MCF au LIPADE, Université Paris Cité) et Jérôme Delobelle (MCF au LIPADE, Université Paris Cité).

Références

[1] Elise Bonzon, Nicolas Maudet - On the outcomes of multiparty persuasion - 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2011)

[2] Louise Dupuis de Tarlé, Elise Bonzon, et Nicolas Maudet - Multiagent Dynamics of Gradual Argumentation Semantics - 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022)

[3] Emmanuel Hadoux, Aurélie Beynier, Nicolas Maudet, Paul Weng, Antony Hunter - Optimization of probabilistic argumentation with Markov Decision Models, International Joint Conference on Artificial Intelligence 2015, IJCAI 2015

[4] Tim Miller - Explanation in artificial intelligence: Insights from the social sciences - Artificial Intelligence, 267:1–38, 2019.