Explanation of the Artificial Agent: Exploring the content and format of human interaction with AI

Soutenance de thèse – Le GUILLOU Marin

22 Septembre 2023 à 14h00

Salle Blériot, École de l'air et de l'espace, BA 701, Salon de Provence

Devant le jury composé de :

Pr. Franck MARS, Laboratoire des Sciences du Numérique, Nantes, Rapporteur
Dr. Aurélie CLODIC, Laboratoire d'analyse et d'architecture des systèmes, Toulouse, Rapporteuse
Pr. Elisabeth PACHERIE, Institut Jean Nicod, Paris, Examinatrice
Dr. Florence DE GRANCEY, THALES AVS, Toulouse, Examinatrice
Pr. Noel NGUYEN-TRONG, Laboratoire Parole et Langage, Aix-en-Provence, Examinateur
Pr. Laurent PREVOT, Laboratoire Parole et Langage, Aix-en-Provence, Directeur de thèse
Dr. Bruno BERBERIAN, ICNA/ONERA, Salon de Provence, Co-encadrant de thèse

Résumé

La révolution technologique promise par les récents progrès de l'intelligence artificielle permet d'imaginer une ère dans laquelle les humains seront assistés par des agents artificiels (AA) dans de nombreuses tâches. Or l'opacité de l'IA embarquée dans les AAs freine aujourd'hui les possibilités de coopération. Il apparait dès lors critique de comprendre comment soutenir la coopération entre les humains et ces agents artificiels. Cette thèse contribue à cet effort en proposant une approche originale du problème. Se basant sur les connaissances inhérentes au contrôle de l'action, notamment de l'action conjointe, la thèse explore la nature de l'information qu'un agent artificiel doit transmettre pour soutenir la coopération avec son partenaire humain. L'hypothèse est faite que la communication des précurseurs des intentions de l'agent artificiel (Intention Based Explanations ou IBEs) est un élément clé de la coopération entre l'homme et l'agent artificiel.

Utilisant une tâche coopérative (Overcooked), nous avons mené une série d'expériences dans lesquelles nous avons étudié l'impact de la communication des intentions (IBEs) de l'agent artificiel sur la performance de l'équipe humain-agent et l'expérience subjective des participants. Nos résultats suggèrent un impact positif des IBEs sur la confiance des participants envers les actions de l'agent artificiel. Les résultats d'une seconde expérience montrent que les IBEs influencent le comportement des participants vers plus de coopération - même s'ils ne sont pas efficaces en termes de performance, tout en répliquant l'impact des IBEs sur la confiance des participants. Dans une troisième expérience, nous démontrons que l'amélioration de la performance individuelle de l'agent réduit l'effet des IBE à la fois sur le plan comportemental et subjectif. Enfin, une quatrième étude indique que les bénéfices en termes subjectifs (notamment la confiance dans l'AA) liés à l'utilisation des IBEs disparaissent si les intentions proximales ou motrices sont présentées de manière séparée.

La thèse apporte des éléments permettant de considérer les IBEs comme des activateurs de l'action conjointe humain-agent artificiel. Les IBEs ont l'avantage d'être des activateurs d'action conjointe "en ligne", alors que les approches plus traditionnelles des explications causales sont des processus

hors ligne. En outre, les IBE ont l'avantage potentiel d'être moins coûteux que les explications causales, compte tenu du compromis performance/transparence des techniques d'IA. Les solutions opérationnelles proviendront probablement d'une combinaison équilibrée des deux approches, associée à des capacités de coopération efficaces du côté de l'agent – éventuellement par le biais d'une lecture adéquate des intentions de l'humain.

Mots clés

Intelligence Artificielle Explicable (XIA), Coopération Humain - Agent Artificiel, Equipe Humain-Autonomie, Modèle Hiérarchique des Intentions, Mécanismes Prédictifs

Vous êtes invité à rejoindre la web-conférence JITSI via le lien ci-dessous :

https://rdv.onera.fr/Soutenancedeth%C3%A8seMarinLEGUILLOU