Avis de Soutenance

Monsieur Yazan MUALLA

Présentera ses travaux en soutenance

Soutenance prévue le ***lundi 30 novembre 2020*** à 14h00
Lieu : Université de Technologie de Belfort-Montbéliard 13 rue Ernest Thierry-Mieg 90010 Belfort, FRANCE
Salle : D-207

Titre des travaux : Expliquer le comportement de robots distants à des utilisateurs humains : une approche orientée-agent

Ecole doctorale : SPIM - Sciences Physiques pour l'Ingénieur et Microtechniques

Section CNU : 27

Unité de recherche : CIAD - Connaissance et Intelligence Artificielle Distribuées

Directeur de thèse : Stephane GALLAND

Codirecteur de thèse : Christophe NICOLLE ☑ HDR ☐ NON HDR

Soutenance : ☑ Publique ☐ A huis clos

Membres du jury :

Nom	Qualité	Etablissement	Rôle
M. Stephane GALLAND	Professeur des Universités	Université de Technologie de Belfort Montbéliard	Directeur de thèse
M. Eric T. MATSON	Professeur	Purdue University	Rapporteur
M. Flavien BALBO	Professeur des Universités	École Nationale Supérieure des Mines de Saint-Étienne	Rapporteur
M. Christophe NICOLLE	Professeur des Universités	Université de Bourgogne	Co-directeur de thèse
Mme Marie-Pierre GLEIZES	Professeur des Universités	Université Paul Sabatier de Toulouse	Examinatrice
M. Laurent VERCOUTER	Professeur des Universités	Normandie Université - INSA Rouen Normandie	Examinateur
M. Abderrafiaa KOUKAM	Professeur des Universités	Université de Technologie de Belfort Montbéliard	Examinateur

Mots-clés : Intelligence artificielle explicable,Systèmes multi-agents,Interaction homme-machine,,

Résumé de la thèse (en français) :

Avec l'émergence et la généralisation des systèmes d'intelligence artificielle, comprendre le comportement des agents artificiels, ou robots intelligents, devient essentiel pour garantir une collaboration fluide entre l'homme et ces agents. En effet, il n'est pas simple pour les humains de comprendre les processus qui ont amenés aux décisions des agents. De récentes études dans le domaine l'intelligence artificielle explicable, particulièrement sur les modèles utilisant des objectifs, ont confirmé qu'expliquer le comportement d'un agent à un humain favorise la compréhensibilité de l'agent par ce dernier et augmente son acceptabilité. Cependant, fournir des informations trop nombreuses ou inutiles peut également semer la confusion chez les utilisateurs humains et provoquer des malentendus. Pour ces raisons, la parcimonie des explications a été présentée comme l'une des principales caractéristiques facilitant une interaction réussie entre l'homme et l'agent. Une explication parcimonieuse est définie comme l'explication la plus simple et décrivant la situation de manière adéquate. Si la parcimonie des explications fait l'objet d'une attention croissante dans la littérature, la plupart des travaux ne sont réalisés que de manière conceptuelle. Dans le cadre d'une méthodologie de recherche rigoureuse, cette thèse propose un mécanisme permettant d'expliquer le comportement d'une intelligence artificielle de manière parcimonieuse afin de trouver un équilibre entre simplicité et adéquation. En particulier, il introduit un processus de formulation des explications, sensible au contexte et adaptatif, et propose une architecture permettant d'expliquer les comportements des agents à des humains (HAExA). Cette architecture permet de rendre ce processus opérationnel pour des robots distants représentés comme des agents utilisant une architecture de type Croyance-Désir-Intention. Pour fournir des explications parcimonieuses, HAExA s'appuie d'abord sur la génération d'explications normales et contrastées, et ensuite sur leur mise à jour et leur filtrage avant de les communiquer à l'humain. Nous validons nos propositions en concevant et menant des études empiriques d'interaction homme-machine utilisant la simulation orientée-agent. Nos études reposent sur des mesures bien établies pour estimer la compréhension et la satisfaction des explications fournies par HAExA. Les résultats sont analysés et validés à l'aide de tests statistiques paramétriques et non paramétriques.

Abstract (in English):

With the widespread use of Artificial Intelligence (AI) systems, understanding the behavior of intelligent agents and robots is crucial to guarantee smooth human-agent collaboration since it is not straightforward for humans to understand the agent's state of mind. Recent studies in the goal-driven Explainable AI (XAI) domain have confirmed that explaining the agent's behavior to humans fosters the latter's understandability of the agent and increases its acceptability. However, providing overwhelming or unnecessary information may also confuse human users and cause misunderstandings. For these reasons, the parsimony of explanations has been outlined as one of the key features facilitating successful human-agent interaction with a parsimonious explanation defined as the simplest explanation that describes the situation adequately. While the parsimony of explanations is receiving growing attention in the literature, most of the works are carried out only conceptually. This thesis proposes, using a rigorous research methodology, a mechanism for parsimonious XAI that strikes a balance between simplicity and adequacy. In particular, it introduces a context-aware and adaptive process of explanation formulation and proposes a Human-Agent Explainability Architecture (HAExA) allowing to make this process operational for remote robots represented as Belief-Desire-Intention agents. To provide parsimonious explanations, HAExA relies first on generating normal and contrastive explanations and second on updating and filtering them before communicating them to the human. To evaluate the proposed architecture, we design and conduct empirical human-computer interaction studies employing agent-based simulation. The studies rely on well-established XAI metrics to estimate how understood and satisfactory the explanations provided by HAExA are. The results are properly analyzed and validated using parametric and non-parametric statistical testing.