Communcating AI intentions to boost Human AI cooperation

Bruno BERBERIAN [a,1], Marin LE GUILLOU [a,b] , and Marine PAGLIARI [a,c]
[a] *Information Processing and Systems Department, ONERA, France*
[b] *Laboratoire Parole Language, Aix-Marseille Universitée, CNRS, France*
[c] *Institut Jean Nicod, Département d'Études Cognitives, École Normale Supérieure, CNRS, PSL University, France.*
ORCiD ID: Berberian https://orcid.org/0000-0002-3908-4358

Abstract. Interacting with Artificial Intelligence (AI) profoundly changes the nature of human activity as well as the subjective experience that agents have of their own actions and their consequences. We propose to mitigate this effect by making AI systems more intelligible to human operators. We hypothesized that the readability of system intentions is a key element of their predictability and, by extension, of the human operator's abilities to interact effectively with highly automated systems. We conducted experiments to explore the impact of the communication of AI intentions during joint human-AI interaction (intention-based explanations). Trust human operators' have towards such algorithms as well as their sense of control across different dimensions (performance, action fluency, contribution) was measured. Overall, our results suggest that adding intention-based explanations during human-AI interaction indeed support cooperation between the human operator and artificial agents.

Keywords. Human-AI cooperation, eXplainable Artificial Intelligence (XAI), Joint actions, AI intentions, intention-based explanations

1. Introduction

Recent technological evolutions have introduced a rupture in our interactions with technology. From simple tools, artificial agents have become full-fledged teammates characterized by a more or less high level of autonomy in terms of decision making, adaptation, and communication [1]. Expanding the current role of the machine transforms the cooperative architecture, introducing new coordination requirements for operators to ensure that their own actions and those of the automated agent are synchronized and consistent. Several researchers have studied to what extent and under what conditions autonomous agents and humans can work together in a team. Notably, many studies assert that human-system coordination requires the development of an adequate mental representation of the operation of the system with which the human interacts [2]. This refers to the concept of a mental model, and corresponds to a mental description of a system's purpose and structure, explanations of how the system works and its observed states, and predictions of its future states [3]. However, the emergence of such a representation is strongly compromised by the introduction of AI (i.e., systems based on machine Learning techniques or deep learning algorithms). If communication is necessary to create a shared representation [4], most AI systems are

1 Corresponding Author: Berberian, Bruno.berberian@onera.fr

silent about how or why they have produced a given output. Collaborative AI design will require designing AI systems capable of communicating about their own operations. Yet little is known about how humans perceive and evaluate algorithms and their results, why a human might trust or distrust an algorithm, and how we can empower humans to cooperate with such systems [5]. A prerequisite for the design of collaborative tools is the identification of the information that must be provided to enable the human operator to work cooperatively with the automation, a question related to the field of investigation called eXplainable Artificial Intelligence (XAI).
The goal of explainable AI is to provide the user with an explanation of why a machine learning system produced a particular result. A lot of work is focused on this issue, both through the design of more transparent algorithms, but also through explainability tools [6]. Yet, XAI remains distant from scientific models of human cognition and is primarily driven by the underlying structure of AI algorithms [7] without considering the potential benefit of replicating the essential parameters of successful human-human interactions. We propose to draw on theories of motor control and in particular the work done in the area of joint action to better understand how to support cooperation between humans and AI. A joint action is an activity involving two or more agents who coordinate their action plans to achieve an external result together [8]. It relies on the synchronization of each partner's actions throughout the execution, and particularly on a set of cognitive processes that support "joint" action planning. Recently, Pesquita and collaborators [9] proposed that joint action planning is closely related to the ability to predict my partner's actions. This planning could be based on a motor plan incorporating predictions about the actions of the individual and their partner [10]. If prediction seems to be central to the coordination between two agents, then the question arises of the information that drives this predictive ability. Notably, coordination relies primarily on the ability to infer the intentionality of others. Sharing agents' intentions before and during the action is a critical element in achieving joint action [11]. Thus, we hypothesize that the readability of system intentions is a key element of their predictability and, by extension, of the human operator's abilities to interact effectively with highly automated systems.

2. Experimental contributions: Sharing AI intention to improve human AI cooperation

To explore the role of AI intention communication on the quality of human-AI cooperation, we conducted three experiments in which we implemented AI intention communication or not, and evaluated the impact of this communication on different dimensions, both at the behavioural and subjective levels. Intentions are considered as "an initial representation of a goal or state to be achieved, which precedes the initiation of the behavior itself" [12].

2.1. General procedure

To address this issue, we use Overcooked [13], a human-AI joint action testbed popular in the field (Fig. 1). This game asks to coordinate at task level (who does what for the collective purpose) and at motor level (avoiding collisions, etc) to achieve a purely cooperative mission. For each experiment, participants were randomly spitted with gender equality respect in two groups, Explained (group E) and Unexplained (group U). Only participants of group E interacted with an agent sharing its intentions. Throughout

the game, the intentions were presented according to the recipes and actions that the virtual agent would perform (see Figure 1). The goal for the participants is to coordinate with the artificial agent to succeed in making a maximum of recipes in 50s. This general procedure was used for the three different experiments. Change in metrics collected make possible to address several issues from this general procedure.

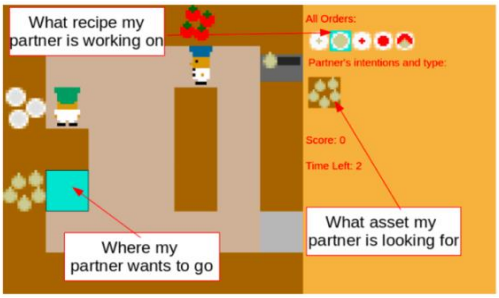

Figure 1 : Our overcooked environment and how intentions are presented in the Explained condition, for more details see [14, 15]

2.2. Experiment 1: impact of communicating AI's intentions on human's trust and performance

Participants - 32 women and 28 men participants participated to the first study. They were randomly spitted with gender equality respect between the two experimental groups (U and E). Both group played 5 blocs of 10 missions lasting 50 seconds each.

Measures - At the end of each block, participants were asked to complete questionnaires (adapted from 16) that included an assessment of their trust in the artificial agent, as well as their perceived contribution of the artificial agent. Average score on each bloc was used to measure each group performance regardless of missions.

Results - Results showed that communication of AI's intentions increased trust and the perception of the AI's contribution to the joint action (Figure 2, left). Interestingly, this communication did not improve the overall performance of the team. However, participants in the group E pressed significantly less ($F=4.151$, $p=0.046$) their keyboard than participants from group U for a similar performance. The results therefore suggest that sharing intentions lead to a different behavioral pattern from participants, more focused on cooperativeness towards the AI. The better assessment of AI's contribution (Figure 2, right) supports this analysis.

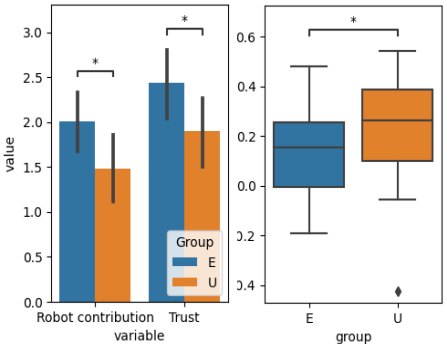

Figure 2 : Left figure represents the perceived relative contributions and trust. Right figure shows the difference between objective AI's contribution fraction and the subjective one. Extensive presentations of these results is availaible at [15]

2.3. Experiment 2: impact of communicating AI's intentions on human's cooperativeness

The hypothesis according to which participants benefitting from AI's intentions will show more cooperativeness is tested in a second experiment. For this purposed, we offered our players the possibility to transmit items to their artificial partner.

Participants - 34 women and 34 men (age 25-30) participated to the second study, again randomly spitted with gender equality respect between the two groups (U and E).

Measure – In addition to the measures used in Experiment 1, we also collected the proportion of action directed towards transmitting an asset to the artificial agent as a measure of the degree of cooperativeness with the artificial agent.

Results - We showed that action proportion directed towards transmission was significantly higher in group E than in group U ($F=9.482$, $p=.003$). As in study 1, participants from group E claimed more trust towards the AI ($F=4.085$, $p=0.047$). Interestingly, this perception did not reflect in team's performance, as performance was significantly poorer in group E.

2.4. Experiment 3: impact of communicating AI's intentions on human's feeling of control and responsibility

Then, we conducted a third study with the same paradigm to explore how communication of AI intentions during joint human-AI interaction affects the sense of control across different dimensions (performance, action fluency, contribution). In this version of the paradigm, participants were faced with two kind of contribution with the artificial agent: in half of the games, they had "symmetric" contribution with the artificial agent, i.e. both the participant and the artificial agent had access to all the ingredients. In the other condition, participants had "asymmetric" contribution, meaning that some ingredients were disposed only in their reachable environment, i.e. they were forced to contribute to the recipes to win the games.

Participants - 100 participants were included in this experiment and were again randomly assigned to one of two groups. Each participant played 4 blocs of 4 missions, with 2 games with symmetric contribution and 2 games with assymetric contribution.

Measure - After each game, participants were asked to complete a questionnaire that included an assessment of the responsibility they felt about the success of the game, as well as their perceived control during the game on a Likert scale in 5 points from not at all to totally.

Results - Our results suggest that, in a situation of joint action with an AI, communication of the AI's intentions increases the level of responsibility of human operators in a situation where they had to take part in the interaction with the agent to achieve the success of the game (Fig. 3). Also, we found that communicating intentions leads to a better judgment of the human operators' performance level by increasing their perceived level of control during a more successful mission, and to a better evaluation of the fluidity of the interaction by decreasing the level of perceived control during non-fluid interactions.

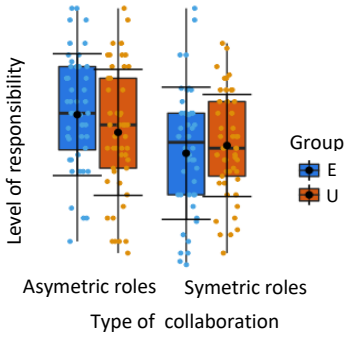

Figure 3: Main result of the third experiment. Communicating AI's intentions to participants lead to higher levels of responsibility towards asymmetric role's

3. Conclusive remarks

These results suggest the importance of intention-based explanations to support human AI cooperation. Most importantly, they show that acceptability and trust seem to be decoupled from team performance and that communication prevails over performance

when it comes to trust in the AI. It is highly likely that the lack of positive impact of the explanations is a result of the poor performance of the proposed AI algorithms. Therefore, delegating control to these agents would lead to a decrease in performance. It is all the more remarkable to observe that despite this poor performance of the agent with which one interacts, one will nevertheless privilege a cooperative behavior as soon as communication allows it. Interestingly, our results also suggest that the addition of intention-based explanations has an effect on the different dimensions of sense of agency by increasing the reliability of this experience of control. Overall, these results suggest interesting avenues of research to improve human-AI interactions and demonstrate the need to take human cognition into account when designing systems that require acceptable and trustworthy AI techniques.

References

[1] Tokadlı G, Dorneich MC, Matessa M. Development Approach of Playbook Interface for Human-Autonomy Teaming in Single Pilot Operations. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2019 Nov 1;63(1):357–61.

[2] O'neill T, McNeese N, Barron A, Schelble B. Human-Autonomy Teaming: A Review and Analysis of the Empirical Literature. Human Factors The Journal of the Human Factors and Ergonomics Society. 2020 Oct 22;1–35.

[3] Rouse, WB, & Morris, NM. On looking into the black box: Prospects and limits in the search for mental models. Psychological bulletin, 100.3 (1986): 349.

[4] McDermott, P, et al. Human-machine teaming systems engineering guide. MITRE CORP BEDFORD MA BEDFORD United States, 2018.

[5] Stoyanovich J, Van Bavel JJ, West TV. The imperative of interpretable machines. Nat Mach Intell. 2020 Apr;2(4):197–9.

[6] Guidotti, R, et al. "A survey of methods for explaining black box models." ACM computing surveys (CSUR). 2018; 51.5: 1-42.

[7] Liao, QV, Gruen, D and Miller, S. Questioning the AI: informing design practices for explainable AI user experiences. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020.

[8] Sebanz N, Bekkering H, Knoblich G. Joint action: bodies and minds moving together. Trends in Cognitive Sciences. 2006 Feb 1;10(2):70–6.

[9] Pesquita A, Whitwell RL, Enns JT. Predictive joint-action model: A hierarchical predictive approach to human cooperation. Psychon Bull Rev. 2018 Oct 1;25(5):1751–69.

[10] Sacheli, LM, Arcangeli, E, & Paulesu, E. Evidence for a dyadic motor plan in joint action. Scientific reports, 2018, vol. 8, no 1, p. 5027.

[11] Michael, J, Pacherie E. On commitments and other uncertainty reduction tools in joint action. Journal of Social Ontology 1.1 (2015): 89-120.

[12] Pacherie E. The Content of Intentions. Mind & Language. 2000;15(4):400–32.

[13] Le Guillou, M., Prévot, L., & Berberian, B. (2023). Supplementary material for AAMAS 2023 paper: Trusting Artificial Agents: Communication Trumps Performance [Data set]. Zenodo. https://doi.org/10.5281/ZENODO.7670961

[14] Marin Le Guillou, Laurent Prévot, and Bruno Berberian. 2023. Trusting Artificial Agents: Communication Trumps Performance. In Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023, IFAAMAS, 8 pages.

[15] Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., & Dragan, A. (2019). On the Utility of Learning about Humans for Human-AI Coordination. Advances in Neural Information Processing Systems, 32.

[16] Hoffman, G. (2019). Evaluating Fluency in Human–Robot Collaboration. IEEE Transactions on Human-Machine Systems, 49(3), 209–218. https://doi.org/10.1109/THMS.2019.2904558.