Trusting Artificial Agents: Communication Trumps Performance

Marin Le Guillou, Laurent Prévot, Bruno Berberian

Trusting Artificial Agents: Communication Trumps Performance

Marin Le Guillou
ONERA, the French Aerospace Lab
TIS Department
Salon-de-Provence, France
marin.le_guillou@onera.fr

Laurent Prévot
Aix Marseille Univ, CNRS, LPL
Aix-en-Provence, France
laurent.prevot@univ-amu.fr

Bruno Berberian
ONERA, the French Aerospace Lab
TIS Department
Salon-de-Provence, France
bruno.berberian@onera.fr

ABSTRACT

Acceptability and trust toward an Artificial Agent (AA) are known to be strongly related to the transparency of its behavior. However, the opacity of the AI techniques implemented to drive the behavior of AAs is growing in parallel with their performance (performance/transparency trade-off). Thus, it is crucial and increasingly required to identify minimal and necessary information for achieving efficient human/AA interaction in order to include them as AAs' requirements at design stage. For this purpose, this paper proposes to bring knowledge and methods from domains accustomed to human behavior studies. Based on ergonomic and cognition literature, this paper tests through a user-study the hypothesis that sharing distal, proximal and motor intentions (what we call intention-based explanations) will improve the acceptability of an AA. The "Overcooked" task from Carroll and colleagues [3] - which requires coordination on goals and motor levels - is used as a test-bed for hypotheses manipulation. Our experimental work consisted in implementing a modified version of the task, analyzing 60 subjects performance, behaviors and feelings in two groups (control and hypothesis-testing) and having them filled an extensive survey. Half of them interact with an agent sharing its intentions while the other half stand in the control group without any information shared by the agent. The results show that intentions sharing leads to a greater acceptability - by means of delegation of control towards the AA - as well as trust. Critically, acceptability and trust seem to be decoupled from team performance.

These results suggest the importance of intention-based explanations as a support for cooperation between the human operator and artificial agents. This work demonstrates the need to take into account human cognition when designing systems requiring acceptable and trustworthy AI techniques.

KEYWORDS

XAI, User-study, Cognition

ACM Reference Format:
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.

1 INTRODUCTION

Academic technical progresses within artificial intelligence (AI) techniques have led to outstanding capabilities demonstrations such as AlphaGo [42]. Consequently industrial companies have started to fund and realize projects integrating these techniques, from personal assistants to autonomous cars. AI engineers focused primarily on pure performance metrics, developing increasingly complex and opaque algorithms. This approach came with a performance/transparency trade-off [10] [6]: the more an AI technique is efficient, the less the ins and outs of its operating are understandable by humans. This problem of transparency is more commonly known as the AI black box problem. The relative importance of the transparency problem is amplified by the use-cases considered for this techniques: they aim at being embedded in artificial agents (AAs) designed in order to act in multi-agent systems including humans. In such scenarios, AAs are assigned to roles from advisors to co-actors. This is the case for autonomous cars which should deal with humans supervisors (the driver) or humans partners cooperating in avoiding accidents (pedestrians, other cars' drivers).

As every technological object, AI faces public and institutional acceptance [12] [11]. The acceptance is strongly linked to the trust AI inspires and consequently to the black-box problem [41]. As a result, growing research towards Explainable AI (XAI) or "how to open the black box" emerged. This way, a consensus emerged regarding the importance of causal explanations for public trust and acceptance [27] [41]

However, public trust and acceptance is not the same as trust and acceptance of the people who directly *work with* the AA. We still know little about what XAI should explain when it comes to make an artificial agent more acceptable and trustworthy from the point of view of the operator who interacts with it. What form should the explanation take when it comes to support human AI cooperation then becomes a critical question that we will call operational XAI.

This paper first presents the first consensus of XAI about causal-based explanations. Then, it is shown how studying the literature of human-human joint-action leads to propose the communication of intention-based explanations from AAs as joint-action enablers. Then, the experimental method of hypothesis testing is used to evaluate the benefits of intention based explanations on joint-action metrics (performance, trust, fluency...). This study shows significant effects of intention based explanations on participant's behavioral patterns as well as on their trust for the AA. Finally, some of the results which are counter-intuitive - especially on performance - plead for a wider use of this method in human-AA joint action research.

2 RELATIVE WORK

The present paper is in line with others calls [27] [10] to benefit of the plentiful literature and experimental methods of human social cognition, human-machine, robot and computer interaction (HMI, HRI, HCI). Tim Miller's contribution [27] [28] [26] about what

should be explained and how is a step forward in the design of recommender system based on the knowledge of social sciences. In these papers, Miller shows why AI researchers developing XAI mainly from their intuition looks like "inmates running the asylum" [28]. Then, papers elicit when people need explanations (mainly when the behavior of the system is different of one's belief of what if would be) and how causal explanations answer to this need. These results about causal-based explanations were later enforced by empirical results [41]. But this work doesn't cover the wider interaction when the AI is used for AAs integrating social work systems. In this regards, efforts are made in the development of "human-aware" AI models but these models often imply to put all the weight of the cooperation at all levels (objectives, methods, action) on the AA's side [3]. However, it is known that coordination is a collective work in such systems [37], and relies on precise information exchanges (either verbal [5] or sensorimotor [16]). Gambling on the availability of the relevant information is risky for systems which are not designed with this interaction requirements. Thus, the next section will dive into ergonomics and cognition theories of human-autonomy teaming and human-human joint-action in order to hypothesize about minimal explanations content that AA designer should think as requirements at design stage.

3 THEORETICAL BACKGROUND AND HYPOTHESIS

The introduction of any kind of new agent in a work system is going to move its equilibrium through a transition phase. In particular, the coordination framework is renewed and new classes of problem are created by the failures in the interactions between agents. The magnitude of these problems is dependent of the relative importance of the new agent in the work system. But these problems have been identified and handled since automated systems have been given responsibilities such as plane autopilots. The human factor literature is rich of models allowing to handle them, though there is a need for stepping-up as AAs provided with AI are expected to be handled with more and more critical - and interaction demanding - tasks.

3.1 Shared Mental Models and Team Situation Awareness

One of the first occurrence of the term "human-machine cooperation" is Hoc [13] in 2000, while the first theories of shared control on action between humans and automation can be found in [40]. The necessity for human and artificial partners to develop a "team situation awareness, (TSA)" [9] and shared mental models (SMM) has emerged as a consensus, with respective definitions from the American Psychological Association (APA) being:

- Team Situation Awareness: "For every member of the team, the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future, required for his/her job"
- Shared Mental Model: "in ergonomics, a mental model of a work system that is held in common by the members of a team. Ideally, team members should have a shared mental picture of the system and its attributes, a shared knowledge of all relevant tasks, and a shared understanding of the team's progress toward its goal. Coordination, efficiency, and accuracy will increase as team members converge on a common mental model that is accurate and complete yet flexible. Also called team mental model."

Then, requirements for cooperating artificial agents have been proposed: the Situation Awareness-based Agent Transparency model by [4] or the transparency on intentional, task, analytical, environment and teamwork agent's models by [20]. Agents developed according to these requirements were tested in user studies [24] [21]. Although, these models are very demanding both in terms of raw-material (information) and in mental resources to value them, so at least two reasons encourage to look for finer cognitive mechanisms with minimal information requirements:

- the performance/transparency trade-off of AI techniques presented in [10]
- in time-constrained situations faced to cognitively demanding tasks - e.g. the cooperation between humans and robots rescuers after an earthquake - people need *heuristics* [8] about their partner behavior. *Heuristics* are "rules of thumb" used to simplify complex cognitive tasks, and create action reflexes.

Therefore identifying the minimal information that would trigger these heuristics is particularly relevant. Interestingly, problems found in human/AA interaction are very similar to the ones which have led to a large literature in cognition regarding joint-action:

> "any form of social interaction whereby two or more individuals coordinate their actions in space and time to bring about a change in the environment" [37]

The next subsection sums-up the main learnings from joint-action literature regarding human/AA interaction.

3.2 Intentions sharing in joint-action

Many researchers in the fields of psychology and cognitive science are exploring the underlying mechanisms of a joint action task. An extensive review of ergonomic concepts cognitive mechanisms of cooperation can be found in [18]. Two key mechanisms for joint-action have been identified:

- shared-task representation (STA) which is for each cooperator the mental representation of the task of every teammember in a plan accomplishment [16]
- mentalization, which is "the ability to understand one's own and others' mental states, thereby comprehending one's own and other' intentions and affects" (APA). This is a process by which useful heuristics can be built.

Both mechanisms are supported by behavioral and electrophysiological studies [39][33][47]. Interestingly, several authors support the role of communication of intentions in the production of joint action [36][43][2] [25]. For example, Sebanz et al. [38] described the ability to make prediction about the intention of other interacting partners as a critical part in a successful performance of a joint action task. However, if the human operator has access to a multitude of cues when it comes to understanding the intentions of a human partner, access to the intentions of artificial agents seems

quite difficult, especially considering the opacity of the systems discussed earlier.

3.3 The role of trust in joint-action

Trust, defined as "an individual's willingness to make himself or herself vulnerable to the actions of another person with the expectation of positive outcomes" [23] is known to be essential in the use of automation under one's control: Muir [29] [30] demonstrated that the more an automation is trusted, the less the operator will retake manual control on it. On the other hand, the study of human-human interactions shows evidences for the role of interpersonal trust in cooperative activities. For example, trust is known to be associated with joint-efforts (as opposed to individual efforts) in the achievement of a common-goal [7]. More generally, Lyu et al. [22] review studies measuring the effects of trust, showing positive effects on creativity, intention to quit, knowledge exchange and creation and others in work context. This literature taken together justifies the use of trust as a key dependent variable when assessing the quality of human-AA joint-action. Although - to the best of our knowledge - lab studies of the impact of trust on cognitive joint-action mechanisms are lacking even if first studies [35] shows an impact of self-confidence on joint-action.

3.4 General hypothesis

In this paper, we propose that intention-based explanations of the behavior of an artificial agent is a critical element when it comes to create more cooperative systems. In our view, intention-based explanations could suit to the problem of human-AA joint-action as causal-based explanations do towards public and institutional acceptance. Here, intentions are considered as "an initial representation of a goal or state to be achieved, which precedes the initiation of the behavior itself" [31]. Particularly, we take as a starting point the hierarchical model of intentions that made a distinction between Distal intentions, Proximal intentions, and Motor intentions - DPM model [32]. Using a joint action task, the following study explore how communicating system's intention improve acceptability and trust in artificial partner.

4 USER-STUDY PRESENTATION

4.1 Overcooked task

To meet our research goals, the task for the user-study had to met two main requirements:

- Allow coordination between the human participant and the AA at the task level and at the spatial level
- Being challenging enough to be considered by the (X)AI community and simple enough to meet the cognition user-study methods

In that respect, the "sequential stag-hunt" from [47] and "overcooked" from [3] were the best candidates, and the recent popularity of the "overcooked" as well as the code availability weighted in favor of the last. The adaptation of the available code to scalable user-studies with different missions, bloc and experimental conditions combinations is a significant contribution of this research work.

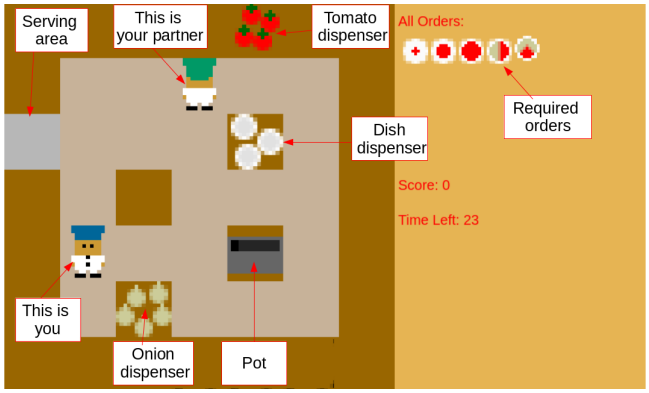

Figure 1: Our overcooked environment

Figure 1 represents the task: two cooks share the same kitchen provided with assets (onion, tomato and dish dispensers, pot and serving area). User-study participants play the blue hat and are asked to score the maximum recipes among the ones presented in the all-orders panel within 50 seconds. A typical game sequence, if we consider a solo player working on the 2 onions + 1 tomato recipe would be : onion dispenser → pot → onion dispenser → pot → tomato dispenser → pot + start cooking → dish dispenser → pot → serving area → start with another recipe...

In this first study, participants played with an artificial agent which behavior is directed by a basic algorithm adapted from Carroll et al.'s [3] planning agent. The game is played at 10 fps (10 game tick per seconds, this has a direct influence on the agent's playing speed).

4.2 Participants and group design

32 women and 28 men participants were recruited through Prolific recruitment platform, aged minimum 20 and maximum 35. They were randomly splitted with gender equality respect between the control group U(nxplained) and the test group E(xplained). Group U was only provided with the information available on figure 1, whereas the participants from group E were provided with additional information about their partners' intentions as shown in Figure 2.

4.3 Procedures

After giving their consent for the use of their data, participants were introduced (see supplementary material for detailed instructions [19]) to the task and had to play 3 tutorial games. Then, both group played 5 blocs of 10 missions lasting 50 seconds each. The 10 missions layouts can be seen in Figure 3. In order to moderate the effect of training on particular layouts, successive 90° rotations were applied at bloc 2,3,4 and a transposition at bloc 5 on every layout. Participants could take a break at the end of each bloc. The median time spent by every participant in the study was 1h02mn and they were rewarded by Prolific with £7.50.

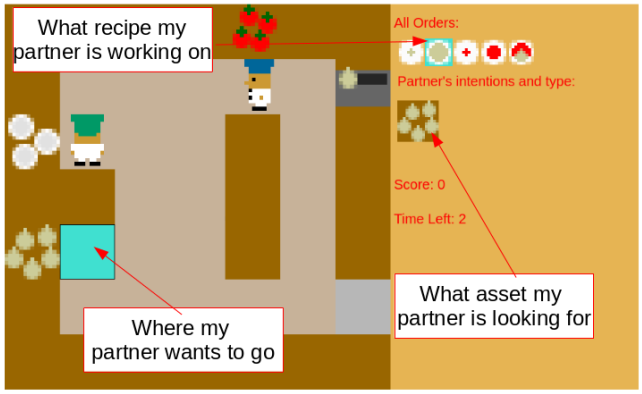

Figure 2: The additional information provided to participants in group E

4.4 Measures and hypotheses on dependent variables

The main goal of the study is to evaluate to what extent the information about intentions provided to the group E brings an advantage compared to group U on perception (trust, acceptability) and performance (score) levels. Regarding perception, subjective measures (questionnaire) have been taken from Hoffman's work [14] for purposes of coherence with the existing knowledge on human/AA fluency. The questionnaire is presented at the end of the last bloc (bloc 5). Participants answer "Please rate your agreement with the following affirmations" on a 5 points Likert's scale (strongly disagree - disagree - neutral - agree - strongly agree) for the 23 affirmations, all available in the supplementary material. This questionnaire makes it possible to get scores on higher level axes:

- Trust: evaluates the trust the robot evokes
- AA's relative contribution: evaluates AA's relative contribution to the team performance. According to Vantrepotte [44, experiment 4, 134-161], the choice to delegate control to a system may be considered as a proxy for AA's acceptability.
- Working Alliance for H-R Teams: adapted from Horvath et al.'s Working Alliance Inventory [15], evaluates the quality of the working alliance between the human and the artificial partner.
- Human/AA fluency: evaluates the overall fluency between the human and the robot

On the other hand, the average score on each bloc will be used to measure each group performance regardless of missions. It is presumed that explanations will bring a performance improvement without additional cost [24]. Additionally, improvement over blocs is expected.

When studying human behavior, it is essential to account for the potential impact of different sources of variability on the measures collected. Classically, statistical methods are used to reduce the effect of these sources of variability on the answers to the questions of interest and help draw more definitive scientific conclusions. The type of statistical method that will be used to analyze the data will depend on the experimental design, the specific questions that the experiment is intended to answer, and the nature of the metrics being collected.

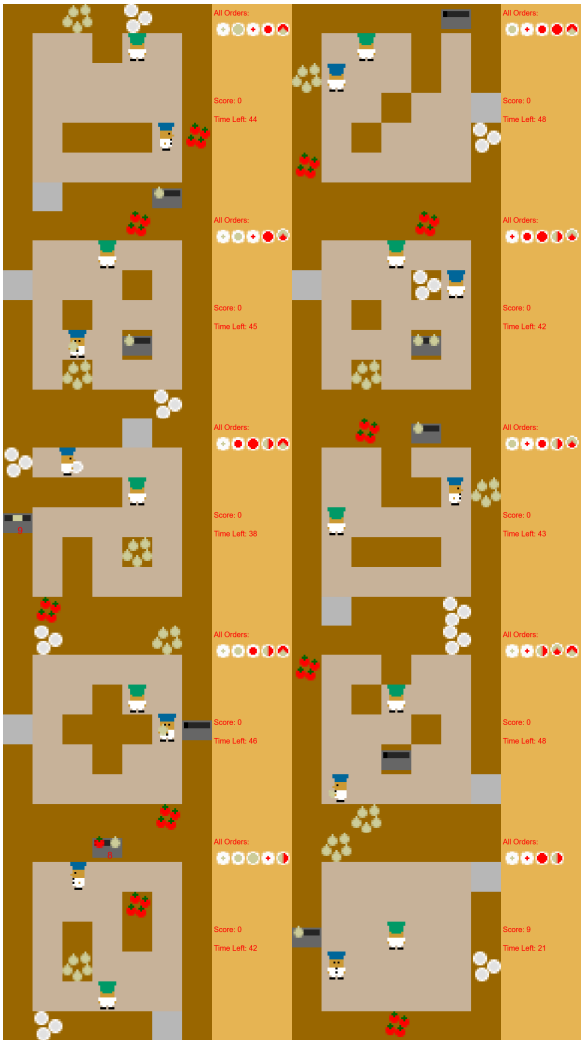

Figure 3: Layouts and recipes of the 10 missions

In this work, we used various tests account for the robustness of our results. The significance level was set at 0.05, meaning that we considered every test result/statistic with a p-value $< .05$ to be "significant." In other words, we were willing to take a 5 percentage risk in concluding that there was a difference when there was not.

Behavioral data were analysed using a 2×5 repeated-measures ANOVA with group (Unexplained vs. Explained) as between subject factors and bloc (from bloc 1 to bloc 5) as within-participants factor. Subjective data were analysed using a Mann-Whitney T-test.

5 RESULTS

5.1 Subjective measures

A confirmatory factor analysis performed on the data confirms the coherence of Hoffman's categories [14] on the collected data for HR fluency, Robot's Relative contribution and Working Alliance

Index (see supplementary material for details). Results of the T-Tests on Hoffman's categories are reported in figure 4, and detailed per-group descriptive statistics as a well as individual questions results can be found in supplementary material.

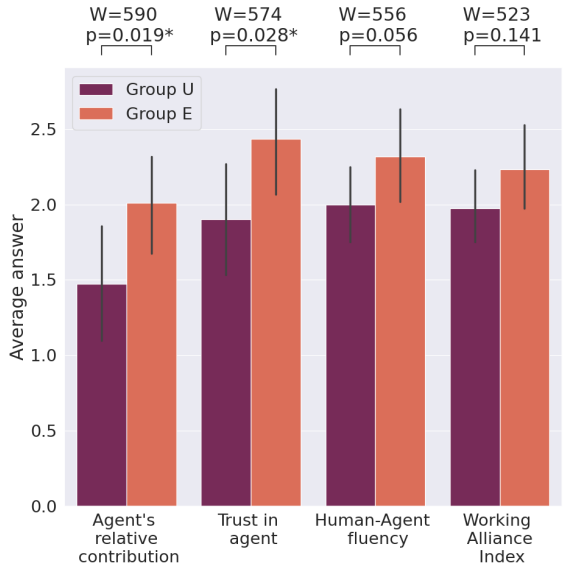

Figure 4: Average scores to Hoffman's subjective categories [14] depending of group. Mann-Whitney test-results are reported. * stands for statistical significance (p<0.05). Error bars represent 95% error interval

Interesting results can be noted:

- Participants from E(xplained) group perceive significantly better AA's relative contribution to the team work(W=590.500, p=0.019).
- Participants from E(xplained) group trust significantly more AA's for efficient action (W=574.500, p=0.028).
- There is no statistical effect of explanations on Working Alliance Index
- There is a weak evidence (W=556.000, p=0.056) for a better perceived human/AA fluency in E(xplained) group.

5.2 Behavioral measures

Score means and standard-deviation regarding blocs and group are presented in figure 5. The hypothesis of a direct performance improvement can be visually discarded, and the repeated-measures ANOVA test excludes (F=0.815, p=0.370) any stastistically significant effect of group on score, while confirming a training effect (F=24.485, p<0.001) as blocs go along.

5.3 Exploratory analysis

The analysis of the correlations between the variables (figure 7) shows that the score correlates negatively with AA's contribution and positively with actions. On the other side, HR-fluency, Trust and AA's contribution correlate positively. This suggests that human's individual action is more efficient than trying to coordinate

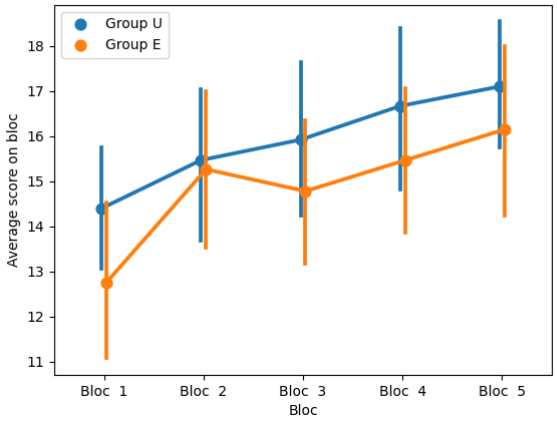

Figure 5: Scores by bloc and group. Error bars represent 95% error interval

with the AA in the configuration of the study. Beside the dependent variables linked to hypotheses, the data collected allows an exploratory analysis. First of all, the number of actions (press on keyboard) performed by every participants have been recorded. A repeated-measure ANOVA performed on this data shows that group E is less active than group U (F=4.151, p=0.046). Then, given the importance of AA's relative contribution for AA's acceptability (see section 4.4), it is interesting to look for an objective measure. For this purpose, a correlation test is performed between the sum of scores and the sum of "interact" actions performed by the participant and the agent. "Interact" action gather the actions of pulling/putting an onion/tomato/dish, start-cooking and deliver which can be considered as actions with high added-value. Pearson's r=0.900 (p<0.001) (versus r = 0.652 for the total number of actions performed) (see figure 7). This comparison presented in figure 8 shows a stronger effect of explanations on subjective than on objective AA's relative contribution.

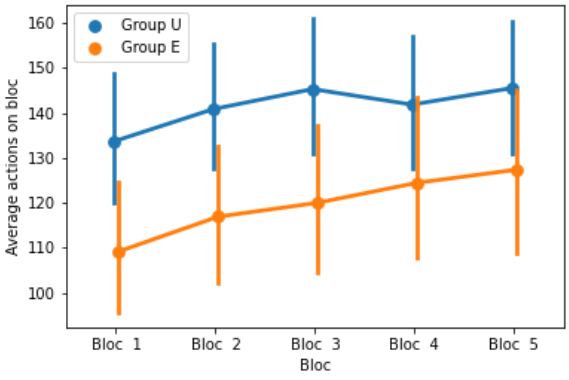

Figure 6: Participant's actions (keyboard press) by group and bloc. Error bars represent 95% error interval

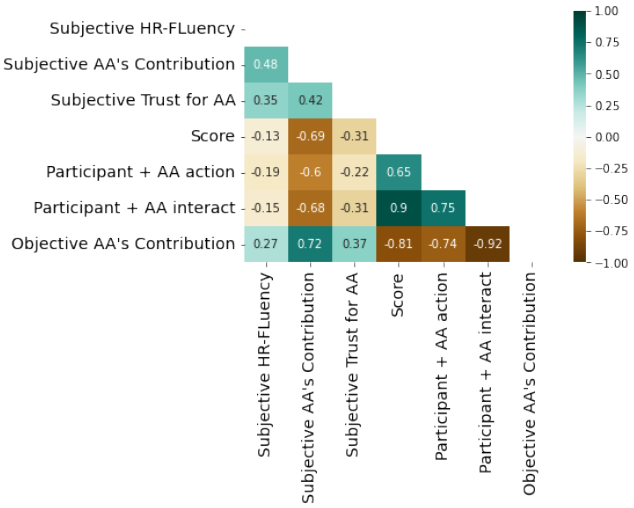

Figure 7: Between variables correlations (Pearson's r)

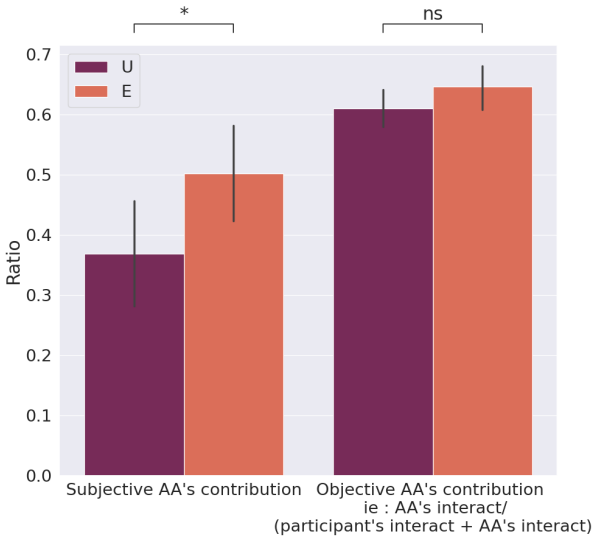

Figure 8: Comparison of perceived and objective AA's contribution across groups.Error bars represent 95% error interval

6 DISCUSSION

The goal of the present study was to explore how communicating the intentions of an artificial agent improves acceptability and trust towards it when engaged in cooperative activities. Three main results were found. As expected, communicating system's intentions is associated with increased trust in the artificial partner. Second, communicating system intention is also associated with an increase, albeit marginal, in the perceived fluency of the human-agent interaction. Finally, while the number of total actions performed by the participants decreases when communicating the system's intention, the number of high value-added actions ("interact", see section 5.3) remains constant.

The first key result concerns the relationship between communication of system intentions and trust in the partner. We hypothesized in section 3.2 that the communication of intentions by an artificial partner will increase trust toward it. The results obtained here confirm this hypothesis by showing the role of the readability of partners' intentions in the trust that we will develop for this partner. This result is all the more important as this communicating system's intentions is not associated with an increase in the team's performance. In other words, the participants seem to be more confident in the system when the intentions of the system are presented even though the performance obtained tends to decrease. While most research focuses solely on the performance of the system, this result indicates that the information returned by the system about its own behavior may be just as important in building trust. However, this is not to deny the role played by the system's performance on the operators' trust.

Today, many studies emphasize the role of predictive mechanisms when controlling our own actions [34], but also when coordinating with others[17][45][34]. Furthermore, the ability to infer the intentions of others seems to contribute to this predictability [38]. Therefore, it is likely that the communication of artificial partners' intentions has an impact on the ability to cooperate and thus on trust in the partner. Behavioral results and subjective feedbacks point in this direction. Subjectively, the participants report a greater fluidity of interaction when they have access to the intentions of the artificial partner. Regarding objective measurements, we observed that participants behaved more efficiently in the presence of the partner's intentions since our results show a decrease in the number of total actions without a decrease in the number of relevant actions (interact). At last, the poor correlation between scores and the perceived HR-fluency show that fluency is not enough for performance and agent's ability to cooperate in its behavior is also essential, confirming [3] results.

The last relevant point concerns the acceptability of the system and the role of intention communication on this acceptability. In this study, we did not have a direct measure of the acceptability of the system. We decided to use the delegation of authority as a proxy for acceptability, according [44, experiment 4, 134-161]. Subjective results on robot's contribution (see figure 8) show that participants provided with explanations have the feeling - without any operational benefit (see results on scores 5) - to give more control to their artificial partner. This result is consistent with [1] which shows that AI's explanations influence humans to accepts suggestions and plans from the AA. This inclination to attribute more control to their artificial partners is likely to indicate a better acceptability of the system. Moreover, results on participant's activity (see figure 6) may indicate an increase of willingness to coordinate (ie: adapting to AA's declared intentions) more than a wait-and-see attitude which would be translated into objective release of control in figure 8.

Further work will need to show how these variables evolves in situations where coordination is either forced or the artificial agent more efficient. Such tests can be done by proposing missions more demanding in term of coordination and/or by changing AA's behavior or execution speed. Moreover, easier comparisons between

objectives and subjective dependent variables should be allowed at design stage (ie: rating AA's relative contribution in percentage, so it can be easily compared to its objective contribution). It would be of particular interest to exchange with researchers engaged in artificial agent design in order to ally the best of both expertises. Indeed, it is maybe easier to design models allowing the extraction of such intention-based explanations than eliciting causal explanations such as "the agent moved right a timestep t because the environment showed *feature A* in combination with *feature B*... etc". For example, considering deep reinforcement learning techniques - very challenging for *why* explanations - design practices such as hierarchical reinforcement learning [46] look at first sight compatible with intention sharing.

Finally, the scope of this paper is limited by the choice of the task, which remain relatively simple compared to potential real-life situation such as search and rescue AAs. However, showing that intention-based explanations has such an influence on behavior and trust in this simple environment justifies to scale-up regarding the tasks. It would be very interesting to see how the DPM model and the relative importance of each of the distal, proximal and motor intentions evolves with complexity? Do intention-based explanations remain realistic regarding the cognitive load of human partners? Our first thought is that the relative importance of distal intentions will grow with complexity. Of course, this need to be tested in further studies, along with the influence of team experience on intention-based explanations (and every level of intentions) relevance.

7 CONCLUSION

The whole paper shows that there are both theoretical (section 3) and empirical reasons (section 5) showing that being able to provide intention-based explanation from the artificial agent to its human partners is essential for trust and acceptability of the system. We do not pretend that explaining the *why* of AA's actions (regarding the environment state) is useless. Only that allowing its partners to build and exploit a mental model of its behavior by reading its intentions will result in more trust and acceptability toward it.

The second main contribution of the paper is methodological: as Miller [27] advocated, AI research and engineering would be mad ignoring decades of behavioral studies on human-human and human-automation interactions when working on AAs destined to interact with humans (and they all are). The results presented in this paper also shows that even when interactions hypotheses are formulated from strong cognitive models, user-studies must be conducted rigorously for evaluation: the results on performance presented here are a good reminder, because they go against the initial hypothesis (there is no performance improvement) and also against the visual analysis (which looks like there is a deprecation whereas statistically it is impossible to reject the null hypothesis).

Finally, this work calls for more integration between research fields. In the particular case studied of overcooked, it would be particularly interesting to ally innovation in AI techniques from [3] with the design guidelines suggested here.

ACKNOWLEDGMENTS

This work is funded under PhD grants by ONERA, the French Aerospace Lab and Région SUD.

REFERENCES

[1] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–16. https://doi.org/10.1145/3411764.3445717

[2] Stephen Butterfill. 2016. Joint action: a minimalist approach. In *The Routledge Handbook of Philosophy of the Social Mind*. Routledge, 373–385.

[3] Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. 2019. On the Utility of Learning about Humans for Human-AI Coordination. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc. https://papers.nips.cc/paper/2019/hash/f5b1b89d98b7286673128a5fb112cb9a-Abstract.html

[4] Jessie Y C Chen. 2018. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science* (2018), 25.

[5] Herbert H. Clark. 1996. Joint actions. In *Using Language*. Cambridge University Press, Cambridge, 59–91. https://doi.org/10.1017/CBO9780511620539.004

[6] Arun Das and Paul Rad. 2020. *Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey*. Technical Report arXiv:2006.11371. arXiv. http://arxiv.org/abs/2006.11371 arXiv:2006.11371 [cs] type: article.

[7] Kurt T. Dirks. 1999. The effects of interpersonal trust on work group performance. *Journal of Applied Psychology* 84, 3 (1999), 445–455. https://doi.org/10.1037/0021-9010.84.3.445

[8] Gerd Gigerenzer and Wolfgang Gaissmaier. 2011. Heuristic decision making. *Annual review of psychology* 62, 1 (2011), 451–482.

[9] Jamie C. Gorman, Nancy J. Cooke, and Jennifer L. Winner. 2006. Measuring team situation awareness in decentralized command and control environments. *Ergonomics* 49, 12-13 (Oct. 2006), 1312–1325. https://doi.org/10.1080/00140130600612788

[10] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. *Science Robotics* 4, 37 (Dec. 2019), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

[11] Abhinav Hasija and Terry L. Esper. 2022. In artificial intelligence (AI) we trust: A qualitative investigation of AI technology acceptance. *Journal of Business Logistics* 43, 3 (2022), 388–412. https://doi.org/10.1111/jbl.12301 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jbl.12301.

[12] High-Level Expert Group on AI. 2019. *Ethics guidelines for trustworthy AI*. Report. European Commission, Brussels. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[13] Jean-Michel Hoc. 2000. From human – machine interaction to human – machine cooperation. *Ergonomics* 43, 7 (July 2000), 833–843. https://doi.org/10.1080/001401300409044

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

[15] Adam O. Horvath and Leslie S. Greenberg. 1989. Development and validation of the Working Alliance Inventory. *Journal of Counseling Psychology* 36, 2 (April 1989), 223–233. https://doi.org/10.1037/0022-0167.36.2.223

[16] Günther Knoblich, Stephen Butterfill, and Natalie Sebanz. 2011. Psychological Research on Joint Action. In *Psychology of Learning and Motivation*. Vol. 54. Elsevier, 59–101. https://doi.org/10.1016/B978-0-12-385527-5.00003-6

[17] D. Kourtis, N. Sebanz, and G. Knoblich. 2013. Predictive representation of other people's actions in joint action planning: An EEG study. *Social Neuroscience* 8, 1 (Jan. 2013), 31–42. https://doi.org/10.1080/17470919.2012.694823

[18] Marin Le Guillou, Laurent Prévot, and Bruno Berberian. 2022. Bringing together ergonomic concepts and cognitive mechanisms for human - AI agents cooperation. *International Journal of Human–Computer Interaction* (2022). https://doi.org/10.1080/10447318.2022.2129741

[19] Marin Le Guillou, Laurent Prévot, and Bruno Berberian. 2023. Supplementary material for AAMAS 2023 paper : Trusting Artificial Agents: Communication Trumps Performance. https://doi.org/10.5281/ZENODO.7670961 Type: dataset.

[20] Joseph B. Lyons. 2013. Being transparent about transparency: A model for human-robot interaction. In *2013 AAAI Spring Symposium Series*.

[21] Joseph B. Lyons, Garrett G. Sadler, Kolina Koltai, Henri Battiste, Nhut T. Ho, Lauren C. Hoffmann, David Smith, Walter Johnson, and Robert Shively. 2017. Shaping Trust Through Transparent Design: Theoretical and Experimental Guidelines. In *Advances in Human Factors in Robots and Unmanned Systems*, Pamela Savage-Knepshield and Jessie Chen (Eds.). Vol. 499. Springer International Publishing, Cham, 127–136. https://doi.org/10.1007/978-3-319-41959-6_11 Series Title: Advances in Intelligent Systems and Computing.

[22] Serena C. Lyu and Donald L. Ferrin. 2018. Determinants, Consequences and Functions of Interpersonal Trust within Organizations. In *The Routledge Companion to Trust* (1 ed.), Rosalind H. Searle, Ann-Marie I. Nienaber, and Sim B. Sitkin (Eds.). Routledge, New York : Routledge, 2017. |, 65–104. https://doi.org/10.4324/9781315745572-7

[23] Roger C Mayer and James H Davis. 1995. An Integrative Model of Organizational Trust. *The Academy of Management Review* , 20, 3 (July 1995), 709–734.

[24] Joseph E. Mercado, Michael A. Rupp, Jessie Y. C. Chen, Michael J. Barnes, Daniel Barber, and Katelyn Procci. 2016. Intelligent Agent Transparency in Human–Agent Teaming for Multi-UxV Management. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 58, 3 (May 2016), 401–415. https://doi.org/10.1177/0018720815621206

[25] John Michael and Elisabeth Pacherie. 2015. On Commitments and Other Uncertainty Reduction Tools in Joint Action. *Journal of Social Ontology* 1, 1 (Jan. 2015), 89–120. https://doi.org/10.1515/jso-2014-0021 Publisher: De Gruyter.

[26] Tim Miller. 2018. Contrastive Explanation: A Structural-Model Approach. *arXiv:1811.03163 [cs]* (Nov. 2018). http://arxiv.org/abs/1811.03163 arXiv: 1811.03163.

[27] Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* (2019). http://arxiv.org/abs/1706.07269 arXiv: 1706.07269.

[28] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. *arXiv:1712.00547 [cs]* (Dec. 2017). http://arxiv.org/abs/1712.00547 arXiv: 1712.00547.

[29] Bonnie M. Muir. 1994. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics* 37, 11 (Nov. 1994), 1905–1922. https://doi.org/10.1080/00140139408964957

[30] Bonnie M. Muir and Neville Moray. 1996. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics* 39, 3 (March 1996), 429–460. https://doi.org/10.1080/00140139608964474

[31] Elisabeth Pacherie. 2000. The Content of Intentions. *Mind and Language* 15, 4 (Sept. 2000), 400–432. https://doi.org/10.1111/1468-0017.00142

[32] Elisabeth Pacherie. 2008. The phenomenology of action: A conceptual framework. *Cognition* 107 (2008), 179–217.

[33] Narender Ramnani and R. Christopher Miall. 2004. A system in the human brain for predicting the actions of others. *Nature Neuroscience* 7, 1 (Jan. 2004), 85–90. https://doi.org/10.1038/nn1168

[34] Aïsha Sahaï, Elisabeth Pacherie, Ouriel Grynszpan, and Bruno Berberian. 2017. Predictive Mechanisms Are Not Involved the Same Way during Human-Human vs. Human-Machine Interactions: A Review. *Frontiers in Neurorobotics* 11 (Oct. 2017), 52. https://doi.org/10.3389/fnbot.2017.00052

[35] Remi Sanchez, Anne-Catherine Tomei, Pascal Mamassian, Manuel Vidal, and Andrea Desantis. 2023. *The role of confidence in decision-making and its pupillary correlate during the interaction with a partner.* preprint. Neuroscience. https://doi.org/10.1101/2023.02.24.529874

[36] John R. Searle. 1990. Consciousness, explanatory inversion, and cognitive science. *Behavioral and Brain Sciences* 13, 4 (Dec. 1990), 585–596. https://doi.org/10.1017/S0140525X00080304 Publisher: Cambridge University Press.

[37] Natalie Sebanz, Harold Bekkering, and Günther Knoblich. 2006. Joint action: bodies and minds moving together. *Trends in Cognitive Sciences* 10, 2 (2006), 70–76.

[38] Natalie Sebanz and Guenther Knoblich. 2009. Prediction in joint action: what, when, and where. *Topics in Cognitive Science* 1, 2 (April 2009), 353–367. https://doi.org/10.1111/j.1756-8765.2009.01024.x

[39] Natalie Sebanz, Günther Knoblich, and Wolfgang Prinz. 2005. How two share a task: corepresenting stimulus-response mappings. *Journal of Experimental Psychology: Human Perception and Performance* 31, 6 (2005), 1234. ISBN: 1939-1277 Publisher: American Psychological Association.

[40] T.B. Sheridan and W.L. Verplank. 1978. Human and computer control of undersea teleoperators. *Human and Computer Control of Undersea Teleoperators* (1978).

[41] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies* 146 (Feb. 2021), 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

[42] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 7587 (Jan. 2016), 484–489. https://doi.org/10.1038/nature16961 Citation Key Alias: silverMasteringGameGo2016.

[43] Raimo Tuomela. 2006. Joint Intention, We-Mode and I-Mode. *Midwest Studies In Philosophy* 30, 1 (2006), 35–58. https://doi.org/10.1111/j.1475-4975.2006.00127.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1475-4975.2006.00127.x.

[44] Quentin Vantrepotte. 2022. *L'incertitude et la machine : la métacognition comme nouveau format d'explication des interactions humain-système.* Ph.D. Dissertation. Université Paris Sciences et Lettres.

[45] Cordula Vesper, Robrecht P. R. D. van der Wel, Günther Knoblich, and Natalie Sebanz. 2013. Are you ready to jump? Predictive mechanisms in interpersonal coordination. *Journal of Experimental Psychology: Human Perception and Performance* 39, 1 (2013), 48–61. https://doi.org/10.1037/a0028066

[46] Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. 2017. Feudal networks for hierarchical reinforcement learning. In *International Conference on Machine Learning*. PMLR, 3540–3549.

[47] Wako Yoshida, Ben Seymour, Karl J Friston, and Raymond J Dolan. 2010. Neural Mechanisms of Belief Inference during Cooperative Games. *The Journal of Neuroscience* (2010), 8.