Towards Intelligent Decision Support Systems in Robotics: Investigating the Role of Self-Confidence Calibration in Joint Decision-Making

Raunak Bhattacharyya*[1], Duc An Nguyen*[1], Clara Colombatto[2], Stephen Fleming[2], Ingmar Posner[1], Nick Hawes[1]

[1]Oxford Robotics Institute, University of Oxford
[2]University College London
{raunakbh, annguyen, ingmar, nickh}@robots.ox.ac.uk, {c.colombatto, stephen.fleming}@ucl.edu

Introduction

Human operators are increasingly working with robots in domains such as nuclear decommissioning (Nagatani et al. 2013; Budd et al. 2020; Chiou et al. 2022), inspection (Hawes et al. 2017; Chiou, Hawes, and Stolkin 2021; Budd et al. 2023), and search and rescue (Casper and Murphy 2003; Dole et al. 2015). Operators in such domains work under time pressure, with incomplete information and communication latencies, such that they often have to make decisions under uncertainty.

An important type of such decision making under uncertainty is choosing between options. For example, a human operator may choose between direct teleoperation or autonomous robot operation (Lee, Mehmood, and Ryu 2016), or may have to select which out of a fleet of robots to assist in the case of robot failure (Ji, Dong, and Driggs-Campbell 2022). Automated agents can also be tasked with making choices. For example, an automated agent may have to decide between operating autonomously or querying the human operator for a demonstration (Rigter, Lacerda, and Hawes 2020), or a decision making agent may select between giving control to the human operator or an automated controller (Costen et al. 2022).

While such implementations often involve decision-making by a single agent (usually the human operator), an open question is whether two decision-makers instead of one may yield better decisions than either individual decision-maker alone. The two decision makers could be both human (Boschetti et al. 2021; Szczurek et al. 2023), or a human with an automated decision-support system (which may be embedded in the robotic platform). Intelligent decision support (IDS) systems are increasingly being used in domains including agriculture (Zhai et al. 2020), maritime transportation systems (Gil et al. 2020), medical diagnosis (Braun et al. 2021), and manufacturing (Turner et al. 2019). The human operator decides whether to accept/reject the IDS input. Given that human operators work under severe cognitive and attentional demands (Norton et al. 2017), this can lead to suboptimal decision making. An open problem therefore is to find a principled method of joint decision making between the human operator and the IDS system.

*These authors contributed equally.

AAAI 2024 Spring Symposium on User-Aligned Assessment of Adaptive AI Systems. Stanford University, Stanford, CA, USA.

We study joint decision making in human dyads with the goal of informing future research on joint decision making in a human-IDS system dyad. Work in human-human dyad joint decision-making in visual-perception tasks has shown that team performance was better than either individual, provided the participants were allowed to exchange confidence estimates in their decisions (Bahrami et al. 2010; Koriat 2012). This is called the Maximum Confidence Slating (MCS) algorithm. This raises the question whether the exchange of confidence estimates between two decision makers can also elevate team performance beyond either individual decision maker alone, in tasks involving human operators controlling robots. Specifically, we focus on the following questions:

1. Is there a benefit of MCS for joint decision making in human-human dyads on a spatial temporal task involving a robot?
2. How well calibrated is the confidence of humans on this task?
3. What is the impact of confidence calibration on the impact of MCS on dyad performance?

We conducted user studies on an online robot navigation experiment with 100 participants. Participants drove two different robots in an online simulation environment. Two different time delays were used to corrupt the control of both robots. The participants had to choose the robot which had the lesser time delay (and therefore was more reliable to be driven) based on their driving experience. They had to also provide their confidence estimate in their choice.

To the best of our knowledge, this is the first experiment where the MCS algorithm for joint decision making has been studied in a dynamic spatio-temporal dynamic task as opposed to a static task. Our results show that:

1. The joint decision is more accurate than either individual's decision when the choice made with higher confidence is selected.
2. The extent of this benefit is regulated by how well calibrated the individuals' confidence is. Pairing poorly calibrated individuals leads to worse performance.

Experiment

We conducted an online study, approved by the University Research Ethics Committee, to investigate joint decision making on a robot controller selection task by a human dyad. Our study, conducted using Prolific, tasked participants to drive two different robots in a given environment, and select the one with lesser delay in controller commands.

A participant was teamed with a pair of simulated robots and tasked with navigating them one by one from a start location to a goal location through a narrow gap. The robot could start at 6 possible initial poses and the doorway could be in 4 different gap configurations, leading to 24 different conditions. The goal was the construction cone.

The participant drove the robot using the W (move forward), A (rotate anticlockwise), S (move back), and D (rotate clockwise) keyboard keys, and was asked to navigate each robot for 6 seconds. The trial number and the elapsed time was displayed to the participant. Each robot had a different amount of delay corrupting the control commands. At the end of the trial, the participant was asked to select which of the two robots they thought had the lower delay. Following this, they were asked to indicate the level of their confidence in the choice by selecting one out of four possible options: lowest, low, high, and highest confidence.

The initial pose of the robot as well as the door configuration was sampled randomly at the beginning of every trial. One of the robots was randomly assigned a delay of 50 ms whereas the delay assigned to the other robot varied based on the result of the previous trials. We imposed a staircase procedure where the task was made harder after 2 consecutive successful choices by decreasing the delay on the robot by 20 ms (thus making the difference in behavior more difficult to distinguish), and the task was made easier after 1 failure by increasing the delay by 20 ms.

Procedure

Participants were prefiltered on Prolific such that they were not allowed to participate if they were under 18 years of age. Each Prolific participant was given a link which took them to our web page. The web page contained an overview of the study, ethics approval information, and the option to participate. If they accepted to participate, they were taken to a more detailed overview of the study. Here, they were shown a video which contained descriptions of various components on the interface, and an example runthrough of two trials. Following the video, they were given 5 practice trials to familiarize themselves with driving the robot.

After the practice round, each participant went through 100 trials. Each trial began with a prompt box. On clicking OK, the timer began and the participant was able to control the robot using the keyboard keys. After 6 seconds had elapsed, the robot was reset to the same initial pose and another prompt box appeared asking whether the participant was ready to drive the other robot. On clicking the button, the participant was able to drive. At the end of the trial, a box appeared which asked the participant to choose the robot with lesser delay. Two radio buttons were provided with robot 1 and robot 2. Following this, another box appeared which asked the participant to select their confidence level in the task. 4 radio buttons appeared with the options as: Lowest confidence, Low confidence, High confidence and Highest confidence. After this submission, a new initial robot pose and door configuration was sampled and the next trial began.

Results

The maximum confidence slating algorithm for joint decision making shows significantly higher accuracy than the higher performing individual. To the best of our knowledge, this is the first study to have shown this on a spatio-temporal dynamic task as opposed to static tasks. This provides a principled way to select between choices made by a dyad: select the choice made with higher confidence.

How well the two individuals' confidence is calibrated has a significant impact on the benefit of maximum confidence selection. We found that the effect does not hold when both participants had poor confidence calibration. These results provide an addition to the existing literature on the benefit of MCS on joint decision making that found that participants need to be of similar skill level (Bahrami et al. 2010; Koriat 2012). We found that pairing dyads according to how well their confidence is calibrated also impacts the benefit obtained by doing MCS on the dyad.

These results supplement the existing research on human-IDS joint decision making. One desideratum for human-IDS pairs is that humans are well calibrated to both their own as well as the IDS decisions (Green and Chen 2019). On a collaborative truss design task, it was shown that good human decision makers were those who varied their probability of accepting the IDS recommendation according to their self-confidence and their confidence in the IDS system (Chong et al. 2023). On a recidivism risk prediction task, it has been shown that humans show well calibrated confidence in their decision making (Alufaisan et al. 2021). The role of confidence estimates provided by IDS systems has also been studied in the explanations and trust space. It was found that confidence estimates provided by IDS systems improved people's trust calibration (Zhang, Liao, and Bellamy 2020). However, they found that there was no significant benefit to joint decision making when the human operators were provided with IDS system confidence estimates. We conjecture that this may be because the human operator was still in charge of the final decision and the higher-confidence decision was not taken.

Conclusion

We studied human-human joint decision making in dyads on a robotics task. We investigated the impact of maximum confidence choice selection on the accuracy of a two alternative forced choice task where the task was to select between two robots, which had different control delays. We found that the accuracy of the dyad joint decision was significantly higher than the higher performing individual. However, when both dyad members had poorly calibrated confidence in their decisions, this benefit was no longer observed. To the best of our knowledge, this is the first study on joint decision making for a spatio-temporal dynamic task involving actively controlling robots. The results suggest that IDS systems should provide confidence estimates along with their decisions, as well as adaptively calibrate confidence according to the human partner. These considerations are important as human-IDS systems become increasingly prevalent to enable seamless autonomy.

References

Alufaisan, Y.; Marusich, L. R.; Bakdash, J. Z.; Zhou, Y.; and Kantarcioglu, M. 2021. Does explainable artificial intelligence improve human decision-making? In *AAAI Conference on Artificial Intelligence (AAAI)*.

Bahrami, B.; Olsen, K.; Latham, P. E.; Roepstorff, A.; Rees, G.; and Frith, C. D. 2010. Optimally interacting minds. *Science*, 329(5995): 1081–1085.

Boschetti, G.; Bottin, M.; Faccio, M.; and Minto, R. 2021. Multi-robot multi-operator collaborative assembly systems: a performance evaluation model. *Journal of Intelligent Manufacturing*, 32: 1455–1470.

Braun, M.; Hummel, P.; Beck, S.; and Dabrock, P. 2021. Primer on an ethics of AI-based decision support systems in the clinic. *Journal of Medical Ethics*, 47(12): e3–e3.

Budd, M.; Duckworth, P.; Hawes, N.; and Lacerda, B. 2023. Bayesian reinforcement learning for single-episode missions in partially unknown environments. In *Conference on Robot Learning (CoRL)*.

Budd, M.; Lacerda, B.; Duckworth, P.; West, A.; Lennox, B.; and Hawes, N. 2020. Markov decision processes with unknown state feature values for safe exploration using gaussian processes. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Casper, J.; and Murphy, R. R. 2003. Human-robot interactions during the robot-assisted urban search and rescue response at the world trade center. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 33(3): 367–385.

Chiou, M.; Epsimos, G.-T.; Nikolaou, G.; Pappas, P.; Petousakis, G.; Mühl, S.; and Stolkin, R. 2022. Robot-assisted nuclear disaster response: Report and insights from a field exercise. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Chiou, M.; Hawes, N.; and Stolkin, R. 2021. Mixed-Initiative variable autonomy for remotely operated mobile robots. *ACM Transactions on Human-Robot Interaction*, 10(4): 1–34.

Chong, L.; Raina, A.; Goucher-Lambert, K.; Kotovsky, K.; and Cagan, J. 2023. The evolution and impact of human confidence in artificial intelligence and in themselves on ai-assisted decision-making in design. *Journal of Mechanical Design*, 145(3): 031401.

Costen, C.; Rigter, M.; Lacerda, B.; and Hawes, N. 2022. Shared Autonomy Systems with Stochastic Operator Models. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Dole, L. D.; Sirkin, D. M.; Murphy, R. R.; and Nass, C. I. 2015. Robots need humans in the loop to improve the hopefulness of disaster survivors. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*.

Gil, M.; Wróbel, K.; Montewka, J.; and Goerlandt, F. 2020. A bibliometric analysis and systematic review of shipboard Decision Support Systems for accident prevention. *Safety Science*, 128: 104717.

Green, B.; and Chen, Y. 2019. The principles and limits of algorithm-in-the-loop decision making. *ACM Transactions on Human-Computer Interaction*, 3(CSCW): 1–24.

Hawes, N.; Burbridge, C.; Jovan, F.; Kunze, L.; Lacerda, B.; Mudrova, L.; Young, J.; Wyatt, J.; Hebesberger, D.; Kortner, T.; et al. 2017. The strands project: Long-term autonomy in everyday environments. *IEEE Robotics & Automation Magazine*, 24(3): 146–156.

Ji, T.; Dong, R.; and Driggs-Campbell, K. 2022. Traversing supervisor problem: An approximately optimal approach to multi-robot assistance. In *Robotics: Science and Systems (RSS)*.

Koriat, A. 2012. When are two heads better than one and why? *Science*, 336(6079): 360–362.

Lee, K.-H.; Mehmood, U.; and Ryu, J.-H. 2016. Development of the human interactive autonomy for the shared teleoperation of mobile robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Nagatani, K.; Kiribayashi, S.; Okada, Y.; Otake, K.; Yoshida, K.; Tadokoro, S.; Nishimura, T.; Yoshida, T.; Koyanagi, E.; Fukushima, M.; et al. 2013. Emergency response to the nuclear accident at the Fukushima Daiichi Nuclear Power Plants using mobile rescue robots. *Journal of Field Robotics*, 30(1): 44–63.

Norton, A.; Ober, W.; Baraniecki, L.; McCann, E.; Scholtz, J.; Shane, D.; Skinner, A.; Watson, R.; and Yanco, H. 2017. Analysis of human–robot interaction at the DARPA Robotics Challenge Finals. *The International Journal of Robotics Research (IJRR)*, 36(5-7): 483–513.

Rigter, M.; Lacerda, B.; and Hawes, N. 2020. A framework for learning from demonstration with minimal human effort. *IEEE Robotics and Automation Letters (RA-L)*, 5(2): 2023–2030.

Szczurek, K. A.; Prades, R. M.; Matheson, E.; Rodriguez-Nogueira, J.; and Di Castro, M. 2023. Multimodal multi-user mixed reality human–robot interface for remote operations in hazardous environments. *IEEE Access*, 11: 17305–17333.

Turner, C. J.; Emmanouilidis, C.; Tomiyama, T.; Tiwari, A.; and Roy, R. 2019. Intelligent decision support for maintenance: an overview and future trends. *International Journal of Computer Integrated Manufacturing*, 32(10): 936–959.

Zhai, Z.; Martínez, J. F.; Beltran, V.; and Martínez, N. L. 2020. Decision support systems for agriculture 4.0: Survey and challenges. *Computers and Electronics in Agriculture*, 170: 105256.

Zhang, Y.; Liao, Q. V.; and Bellamy, R. K. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Conference on Fairness, Accountability, and Transparency*.