IS CONFIDENCE ALL YOU NEED? EXPLORING HUMAN-AI JOINT DECISION-MAKING IN SPATIOTEMPORAL ROBOTIC TASKS

Duc-An Nguyen
Oxford Robotics Institute
University of Oxford
Oxford, United Kingdom
annguyen@robots.ox.ac.uk

Raunak Bhattacharyya
Yardi School of Artificial Intelligence
Indian Institute of Technology Delhi
New Delhi, India
raunakbh@iitd.ac.in

Clara Colombatto
Department of Psychology
University of Waterloo
Ontario, Canada
clara.colombatto@uwaterloo.ca

Steve Fleming
Department of Experimental Psychology
University College London
London, United Kingdom
stephen.fleming@ucl.ac.uk

Ingmar Posner
Oxford Robotics Institute
University of Oxford
Oxford, United Kingdom
ingmar@robots.ox.ac.uk

Nick Hawes
Oxford Robotics Institute
University of Oxford
Oxford, United Kingdom
nickh@robots.ox.ac.uk

ABSTRACT

The growing integration of agentic artificial intelligence technologies into human workflows has introduced a new paradigm of AI-assisted decision-making. While previous research has demonstrated that collaboration between humans and AI can lead to higher accuracy than either working alone, such studies have predominantly focused on static and passive tasks, such as price prediction, recidivism risk assessment, conversation and content moderation. In this study, we explore human-AI joint decision-making in a dynamic spatiotemporal robotic task, where humans tele-operate robots. Using human-subject experiments involving 100 participants, we evaluated a teleoperation task in which participants chose between two mobile robots in a simulation, guided by an AI agent providing its confidence level. Our findings reveal that human meta-decisions - particularly in resolving disagreements between humans and AI - are often suboptimal and confidence-driven frameworks such as Maximum Confidence Slating (MCS) can significantly enhance joint decision-making outcomes ($p < 0.001$). To the best of our knowledge, this is the first application of MCS in a human-AI joint decision-making context. Moreover, we discovered that both well-calibrated and poorly-calibrated AI agents influence human decision accuracy. However, a well-calibrated AI agent that effectively represents its confidence can lead to better decision outcomes, while poorly calibrated AI is more likely to steer users toward negative changes in their decision-making process. These results not only highlight the importance of well-calibrated AI confidence levels beyond performance metrics but also provide important insights into fostering effective collaboration and enhancing human-AI joint decision-making capabilities in complex spatiotemporal tasks.

1 INTRODUCTION

The drive for human-AI agent decision-making arises from the complementary strengths and abilities of humans and computing machines. As we rapidly advance towards an AI-centric future,

these AI agent systems are increasingly being utilized in collaboration with humans across a wide spectrum of decision-making domains, including healthcare, business, and design. Since most of these decision aids are, deployed in static and passive tasks, such as recidivism risk prediction Noti & Chen (2023); Chiang et al. (2023), content moderation Keswani et al. (2021), medical diagnosis Hemmer et al. (2022); Pham et al. (2023), and rental price bidding Wang et al. (2023), little is known about their integration into dynamic, complex domain like robotics. These operational settings often involve high-stakes situations, particularly in teleoperated robotics used for search and rescue operations, medical surgery, military and defense tasks, nuclear and hazardous material handling, and space exploration and maintenance. In this scenarios, decision-making processes are pivotal for achieving optimal control, navigation, and task execution and can have profound impacts on human lives.

As semi- or fully autonomous teleoperated robots become increasingly integrated into various industries - such as manufacturing (e.g., humanoid robots), healthcare (e.g., surgical robots), transportation (e.g., delivery robots), space exploration, and military applications - the need for precise and reliable real-time decision-making in selecting capable robots for critical missions is crucial. Poor decisions can lead to inefficiencies, safety risks, or even failures in overall robotic operations. Given these high stakes, understanding how humans and AI can best collaborate in selecting capable mission-critical robots is essential and remain challenging, particularly in ensuring that these joint decision-making systems work effectively to improve outcomes in real-time robotic operations. In addition to that, optimizing this partnership to improve overall decision accuracy and efficiency in scenarios involving robot control with different competence levels remains an open problem.

Early robotic systems involving teleoperation relied solely on human operators as the decision-makers. Human decision-making, however, often relies on intuition, which can be susceptible to biases, emotions, and incomplete information Casper & Murphy (2003); Norton et al. (2017); Nguyen et al. (2020b). This can result in the gradual degradation of decision-making performance, particularly in complex, fast-paced, and long-term teleoperation scenarios. AI, while powerful, often lacks the contextual understanding and intuition that humans have. Recent advancements have sought to address these limitations by integrating AI agents with the aim of enhancing decision accuracy through human-AI collaborative decision-making processes, particularly in scenarios involving high cognitive demands Hawes et al. (2017); Kunze et al. (2018); Rigter et al. (2020); Tankelevitch et al. (2024). At the outset, human-AI joint decision-making has been managed by a setup with human-initiative control Hawes et al. (2007); Chiou et al. (2021); Nguyen et al. (2025a), where the final decision to accept or reject the AI recommendation lies with the human operator. Few have delved into the domain of joint decision-making, where both human and AI have an active role, especially in the context of real-time robotics control.

Our research takes this human - AI joint decision making studies a step further by integrating confidence into the decision-making process of the dynamic robotic task. We created a simulation environment where humans interact with robots with varying delay factor in completing spatiotemporal rescue tasks, and confidence-calibrated AI agents were there to assist human in robot decision-making. The task in focus involves selecting the optimal rescue robot for specific control scenarios where the robots have different delay characteristics. This delay injection concept reflects the reality that control signal transmission is aleatorically affected by the environment in which the robot is operating. In real-time robotic operations, decision-making process is critical, as delays can significantly affect performance, safety, and task completion. Even small delays in control can lead to substantial deviations from desired outcomes, such as imprecise movements, reduced efficiency, or even accidents. Ensuring that the right robot is chosen is vital for maintaining the reliability and effectiveness of robotic systems. Our setup enables an in-depth exploration of human-AI decision-making dynamics in spatiotemporal robotic tasks and examines how shared confidence levels influence human-AI team decisions. Through this experiment, we evaluated human collaboration with AI agents in making critical decisions and assessed meta-decision-making methods, such as Maximum Confidence Slating (MCS), to improve team output.

Our study yields the following significant findings:

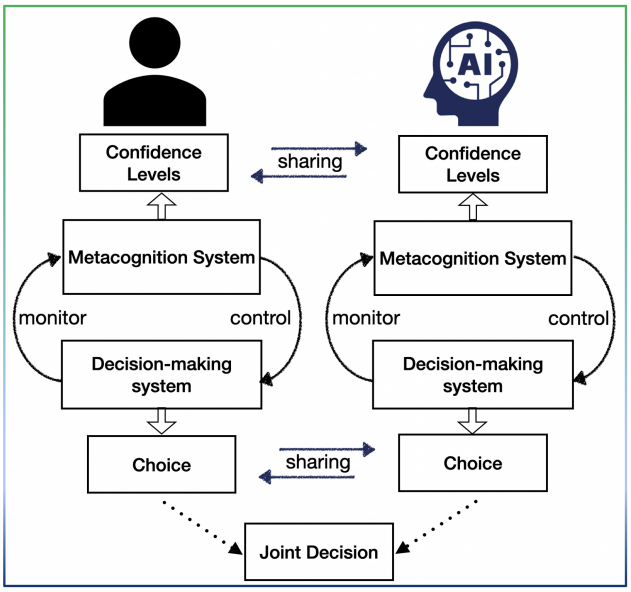

Figure 1: **Joint Decision-Making Frameworks:** The diagram shows how confidence levels relate to metacognition and decision-making, with humans and AI exchanging confidence level to make joint decisions.

1. **Human Suboptimality and Effective of Confidence-Driven Algorithms:** Our findings reveal that human meta-decisions[1], particularly when disagreements arise between humans and AI systems, often underperform compared to algorithmic approaches. Maximum Confidence Slating (MCS), utilizing an accessible metric like confidence ratings, offers an effective alternative for resolving disagreements in joint decision-making. This is significant because it marks the first application of this algorithm in the context of human-AI joint decision-making within robotics.
2. **Impact of AI Confidence-Calibration on Human-AI Collaboration:** Our study also provides insight into the different effects of well-calibrated and poorly calibrated AI agents on human-AI joint decision-making accuracy. Our finding suggests that even suboptimal AI systems can positively impact human decision-making by prompting a self-verification process in humans. However, the importance of high-quality AI confidence-calibration cannot be overstated. Confidence sharing emerges as a critical mechanism, enabling AI to enhance not only the decision outcomes but also the collaborative decision-making process itself. A well-calibrated AI agent that transparently represents its confidence can lead to better decision outcomes, while poorly calibrated AI is more likely to steer users toward negative changes in their decision-making process.

2 Background and Related Work

Research on joint decision-making in human-human dyads began with studies involving two-alternative forced-choice tasks in visual perception. Bahrami et al. demonstrated that confidence plays a crucial role in joint decision-making. More broadly, systems capable of self-assessing their cognitive processes are referred to as metacognitive systems, which manage confidence levels in their own decision-making processes. Bahrami et al. found that human dyads, by sharing metacognitive sensitivity through confidence, achieve significantly better collective output compared to separate individuals. Such a dyadic system is depicted in Fig. 1. In their experiments, participants judged which of two briefly-presented stimuli contained an oddball target. They first made their decision individually, then shared their decisions, and if they disagreed, they discussed the matter until they

[1]A meta-decision refers to a higher-level decision about the decision-making process itself. It's essentially making choices about when to use AI assistance in the overall decision-making process.

reached a joint decision. The results led to the conclusion that "for two observers of nearly equal visual sensitivity, two heads were definitely better than one provided they were given the opportunity to communicate freely." Bang et al. (2017) In discussing the mechanism for the two-heads-better-than-one (2HBT1) effect, the authors assumed that each individual can monitor the accuracy of their performance and can communicate their confidence accurately to the other member.

Further studies by Koriat demonstrated that the benefit of having two heads remains even in the absence of communication between participants in a dyad. Recent works Bang et al. (2014); Massoni & Roux (2017); Bang et al. (2017); Rouault et al. (2018) have concluded that "the decisions advised by the confidence heuristic were just as accurate as those reached through interaction, for individuals of nearly equal reliability." This indicates that the confidence heuristic is a very important and promising tool for enhancing joint decision-making ability in dyadic systems. With only confidence in decision sharing, the Maximum Confidence Slating (MCS) algorithm was used to select the higher confidence choice made in a dyad and it was shown that the MCS algorithm resulted in higher accuracy than the better performing individual. Studies in human-human joint decision making on varied tasks such as threat detection by observing video feeds Bhattacharyya et al. (2021), detecting fake news Guilbeault et al. (2021), deciding rank ordering between items on a survival situation task Hamada et al. (2020), and breast and skin cancer diagnosis Kurvers et al. (2016), and robotics tele-operation Bhattacharyya et al. (2024); Nguyen et al. (2025b) have also shown that higher confidence decision selection leads to higher accuracy. Additionally, the collective benefit is statistically indistinguishable between conditions with and without feedback in the long-term perceptual task. Given enough practice, the learning process of "without feedback" takes longer to develop but leads to as much collective benefit as "with feedback". Bang et al. (2014)

In the domain of human-AI interaction, extensive research explores how humans are influenced by AI through various factors, including the accuracy of AI models: Yin et al. (2019); Lai & Tan (2019); Hoel et al. (2020); Asmar & Kochenderfer (2022); Srivastava et al. (2022), confidence: Zhang et al. (2020); Rechkemmer & Yin (2022), the type of AI explanations and the ways that they are presented: Yang et al. (2020); Bansal et al. (2021); Corso et al. (2019), humans' mental models about AI: Bansal et al. (2019a;b); Glaese et al. (2022); See et al. (2019b;a), and the level of agreement between humans and AI models: Lu & Yin (2021); Sanneman & Shah (2020); El-Sayed et al. (2024), metacognition in generative AI and human interaction: Tankelevitch et al. (2024); Meimandi et al. (2024); Zhang et al. (2024); Nguyen et al. (2020a); Kim et al. (2024); Lee et al. (2024). Nonetheless, most studies have concentrated on static and passive tasks. In contrast, dynamic tasks involve continuously changing variables and real-time interactions, which introduce a higher level of complexity. Unlike static tasks, where decisions rely on fixed data points, dynamic tasks require instantaneous decision-making and adjustments based on fluctuating and spatio conditions. Consequently, the domain of human-AI interaction in dynamic, spatiotemporal tasks, ubiquitous in human-robot interaction, has yet to be thoroughly explored.

To address this gap in knowledge, our work investigates the impact of AI suggestions in a spatiotemporal environment. Motivated by the use of confidence heuristics from human psychology, this study is, to the best of our knowledge, the first to examine the Maximum Confidence Slating (MCS) algorithm within a human-AI joint decision-making context in the robotics application domain.

3 METHODOLOGY

In this section, we first provide a formal definition of the AI-assisted decision-making framework, specifically focusing on the confidence-sharing setting studied in this paper. Next, we describe the design of the AI system with confidence-sharing capabilities. Building up on this, we introduce the meta-decision approach for resolving human-AI disagreements using the Maximum Confidence Slating (MCS) algorithm. Finally, we briefly present the framework that supports the entire experiment and the metrics used to evaluate our implementation.

3.1 PROBLEM FORMULATION

We consider the following AI-assisted decision-making setting in this paper: A human decision maker must complete a sequence of T binary robot selection tasks in a varying environmental context with the aid of an AI model. in each trial t ($1 \leq t \leq$ T). The initial pose of the robot and the door

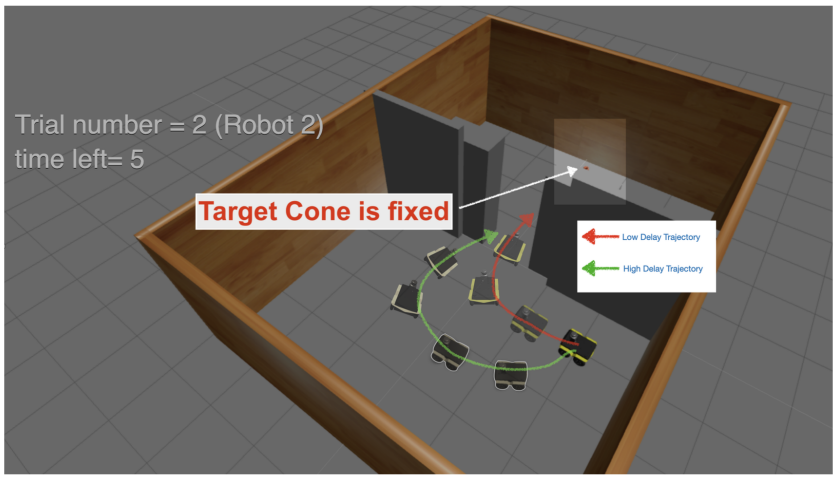

Figure 2: **Experiment Setup**: Online robot navigation simulator using ROS and web interface. Red and green curves show trajectories driven by a study participant controlling the Jackal robot under low and high delay conditions, respectively.

configuration are sampled randomly at the beginning of each trial, while the goal position, to which the human must teleoperate the robot within a time limit, remains fixed. An example of a rescue robotics teleoperation task in a ROS environment is shown in Fig. 2. The delay characteristics of each robot are randomly preset using a staircase procedure. In this procedure, the task difficulty is adjusted adaptively: after two consecutive successful choices, the delay difference between the two robots is decrease by 20 ms, making the behaviour difference harder to distinguish. Conversely, if a failure occurs, the delay is increase by 20 ms, making the task easier. This adaptive adjustment ensures that the accuracy achieved by any participant is around 70%, preventing the task from being too easy (resulting in ceiling performance) or too difficult (leading to performance close to random). This approach helps avoid participants being either overly confident or underconfident about their performance Levitt (1971).

The participant teleoperates each robot and provides their choice of robot that they deem to have a greater advantage in reaching the preset rescue point more promptly (D_h), along with their confidence (C_h). Additionally, the decision-maker receives an AI decision recommendation (D_{AI}), which may or may not be the same as ($D_{correct}$), the correct decision of this trial, along with their confidence (C_{AI}). With all this information, the human decision maker needs to make a final decision along with their confidence ($D_{h-final}$, $C_{h-final}$) which allows human to accept or reject the AI agent recommendation, as well as adjust their confidence. The set of robot decisions is $\mathcal{D} \in \{\texttt{Robot 1}, \texttt{Robot 2}\}$, and confidence is $\mathcal{C} \in \{\texttt{Highest Confidence}, \texttt{High Confidence}, \texttt{Low Confidence}, \texttt{Lowest Confidence}\}$. When the final decision is made, the decision maker *will not* be informed of its correctness. Through this setup, we aim to study the meta-decision-making ability of humans in a human-AI agent dyad and the benefits of confidence sharing between humans and AI agents in making joint meta-decisions. We employ the Maximum Confidence Slating algorithm to make meta-decisions when the human and AI agent choices conflict. This occurs in the absence of feedback, so the human may not have complete information about the AI's performance. The Maximum-Confidence Slating (MCS) algorithm then facilitates joint decision-making by selecting the decision with the highest confidence from two sources: a human decision maker (D_h) and an AI agent (D_{AI}). The MCS algorithm is datailed in Alg. 2.

An example of the choice-making interface design is shown in Fig. 3.

3.2 AI Recommendation Model

The focus of our study is to explore how humans benefit from AI agent suggestions in this telerobotics setting through confidence sharing. To equip the AI agent with confidence-sharing ability,

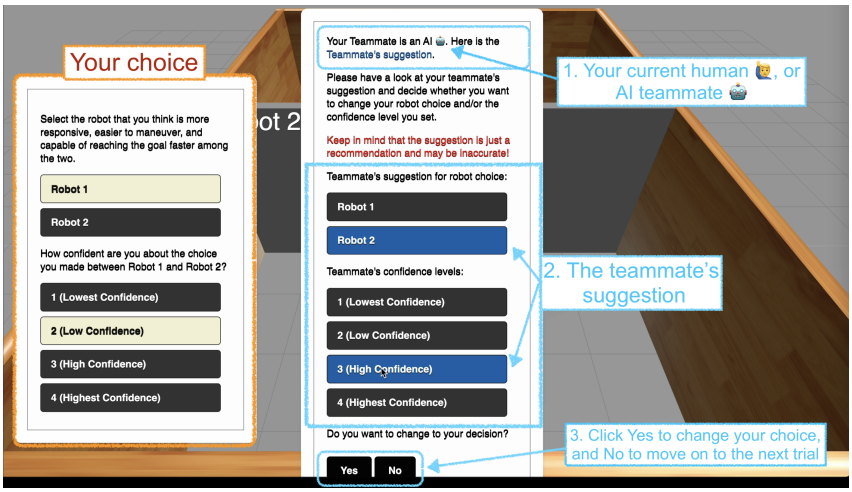

Figure 3: **Study Interface**After completing the robots teleoperation, participants saw three panels sequentially: the left panel for their initial robot choice and confidence, the middle panel for the AI's suggestions and confidence. If the participant decides to change their initial choice, a third panel, similar to the left, will be shown to collect their final choice and confidence.

we collected one round of human data in a separate experiment where humans performed the same tele-robotic task without AI assistance. We used that data to train the AI agent for this study.

For each set of robot delay characteristics, which determine the difficulty of the trials, we built discrete distributions of human confidence levels (`Highest Confidence`, `High Confidence`, `Low Confidence`, `Lowest Confidence`) separately for both correct and incorrect decision scenarios. The algorithm to calculate the confidence calibration of humans Nguyen et al. (2025c) is presented in Alg. 1. Based on this, we designed two types of AI agents: "Good" (well-calibrated) and "Bad" (poorly-calibrated). The Good AI agent learns from the discrete confidence distribution of humans with well-calibrated confidence, while the Bad AI agent learns from the confidence distribution of participants with poorly calibrated confidence, based on data from our previous experiments. The AI agents then sample their confidence from these discrete distributions in specific task setups.

In this stage of our AI agent development, the AI is designed to know the correct robot choice 70% of the time across all trials where it assists humans. Finally, given its current knowledge of human confidence and decision accuracy, the AI agent recommends choices to human users, along with its confidence level for each separate tele-operation trial.

3.3 Algorithm for Resolving Disagreements

Our studies specifically focused on decision-making scenarios where the human participant and AI agent held opposing choice of robot. To maximize the best outcome in these cases, we employed the Maximum Confidence Slating (MCS) Alg. 2, which is extensively studied by prior research in the human-human psychology domain as the meta-decision maker Massoni & Roux (2017); Bang et al. (2017); Rouault et al. (2018).

3.4 Evaluation Metrics

The performance metrics used in our research were accuracy, and the rates of positive and negative changes in human decision-making after observing suggestions from both good and poor confidence-calibtated AI agents. We also assessed the accuracy of human decision-making both before and after receiving AI agents suggestions. Additionally, the performance of the dyad's meta-decision using the MCS algorithm was evaluated based on the accuracy metric.

4 HUMAN-SUBJECT EXPERIMENT

We conducted an online study, approved by the University Research Ethics Committee, to collect human data on AI-assisted decision-making in the robot controller selection task.

4.1 PARTICIPANTS

We collected responses from 100 participants in total via Prolific Platform. We excluded participants who were either unable to perform the task or were inattentive, based on the following criteria: 1) those with an accuracy of less than 65%, and 2) those who gave the same confidence rating response for over 95 of 100 trials. This threshold was chosen to ensure that all participants included in the analyses were performing the task above chance, as is standard procedure in the literature studying human self-confidence Rouault et al. (2018).

Out of the 100 participants who participated in our experiment, 80 passed the above criteria and their data was used for the subsequent analysis. Demographics included 47 males and 33 females. The mean age of participants was 37, with a standard deviation of 11. The average time taken for the experiment was 1.0 hours, with a standard deviation of 0.2 hours, paying an average of 7.6 USD.

4.2 EXPERIMENTAL SETUP

The study was designed to allow participants to take part online from any location. Each participant was paired with two identical simulated rescue robots, each programmed with different delay characteristics in the control. They were tasked with navigating the robots, one at a time, from a start location to a rescue goal through a narrow gap with a closing door as an obstacle. The interface is shown in Fig. 2. The rescue robots randomly start in 1 of 6 possible initial positions, and the doorway has 4 different gap configurations, resulting in 24 unique environment setting. The fixed rescue goal location is marked by a traffic cone.

At the start of each trial, the robot's initial position and door configuration were randomly selected. Participants controlled the robots using the keyboard. They drove each robot in random order within 6 seconds limit, the minimum driving time needed to make a decision between the two robots.

We introduced a time delay in the control of each robot. One robot had a fixed delay of 50 ms between key press and action, while the other robot's delay started at 70 ms and varied based on previous trials. These delays set the task difficulty level. The order of robot appearance was shuffled randomly across 100 trials, and a staircase procedure adjusted this difficulty. This experiment setup was initialized and optimized through multiple iterations of separate pilot studies with $N = 10$ samples each.

4.3 EXPERIMENTAL PROCEDURE

Participants were prefiltered on Prolific to ensure they were at least 18 years old. Selected participants received a link to a study overview page with ethics information and participation details. After agreeing to participate, they viewed a detailed study overview, including a video demonstration and example trial runs. They then completed 5 practice trials to familiarize themselves with controlling the robot using a keyboard controller. In the main experiment, participants conducted 100 driving trials. At the conclusion of each trial, they were asked to select the robot they believed had the lower control delay. The AI agent, which had prior knowledge of the task difficulty as well as a learned distribution of real human calibration confidence levels at that specific task difficulty, would then provide the human participant with a recommendation on the correct choice, along with its confidence. After receiving the AI agent's advice, participants were given the opportunity to either follow the recommendation or re-evaluate their original selection.

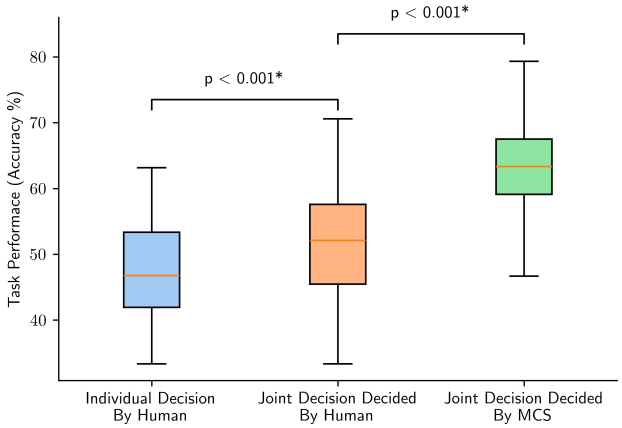

Figure 4: **Joint Decision-Making Performance**: Joint decision-making human-AI team achieved higher accuracy than human decision-making alone ($p < 0.001$). Moreover, the MCS combining human and AI decisions outperformed joint decisions decided human ($p < 0.001$).

5 RESULTS

5.1 RQ1: HOW CAN WE ACHIEVE THE BEST OUTCOME WHEN HUMANS AND AI DISAGREE ON A CHOICE?

One of our main focuses is on trials where there is a disagreement between the human participant and the AI agent regarding the choice of robot.Maximum Confidence Slating (MCS) is a more effective approach to joint decision-making between humans and AI agents, compared to relying solely on human final decisions.

In Fig. 4, "Joint Decision Decided By Human" represents cases where, in the event of a disagreement, the final decision is made by the human participant, while "Joint Decision Decided By MCS" refers to cases where the final decision is determined by the Maximum Confidence Slating (MCS) algorithm. Although the final decision made by humans after receiving AI recommendations shows well improvement compared to their original decision without AI support (52.5% vs. 46.7%), the accuracy of the MCS algorithm is significantly higher than when the human makes the final decision ($p < 0.001$). The MCS algorithm achieved an average accuracy of 63.3%, which is substantially higher than the 52.5% accuracy achieved by humans.

This finding indicates that human-based decision-making is inefficient, and that there is a more optimal method exists for resolving disagreements between humans and AI agents. By implementing Maximum Confidence Slating, based on a simple and easily accessible metric confidence we have found an effective method for resolving decision-making disagreements between humans and AI agents.

5.2 RQ2: HOW DO HUMANS BENEFIT FROM GOOD AND BAD AI IN SPATIOTEMPORAL TASK?

As mentioned in Section. 3.2, we specifically designed two types of AI agents for this study: Good and Bad, both acting as teammates advising humans. Out of 100 participants, we randomly assigned the Good AI agent to 50 participants and the Bad AI agent to the remaining 50. Fig. 5 illustrates the accuracy of human decision-making when paired with Good versus Bad AI agents, both before and after receiving AI recommendations.

We found that pairing humans with both Bad and Good AI agents increased their final decision accuracy, with $p < 0.05$ and $p < 0.005$, respectively. This counterintuitive result suggests that even suboptimal AI agents can positively impact human decisions, possibly by encouraging a self-verification process. Introducing 'confidence' sharing between the human and AI agents, the collective benefit was enhanced. These results align with findings from many studies on static and passive tasks, but this is the first time such an effect has been confirmed in a spatiotemporal robotic task.

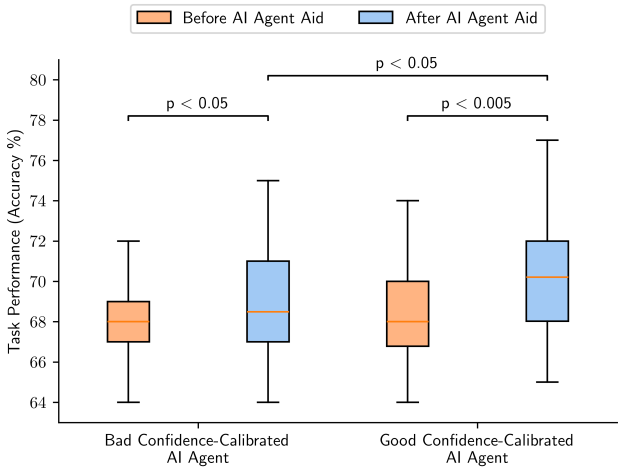

Figure 5: **Task Performance Analysis**: Pairing humans with both Bad and Good Confidence-Calibration AI agents significantly improved their final decision accuracy. Additionally, humans paired with the Good AI agent achieved higher final decision accuracy ($p < 0.05$).

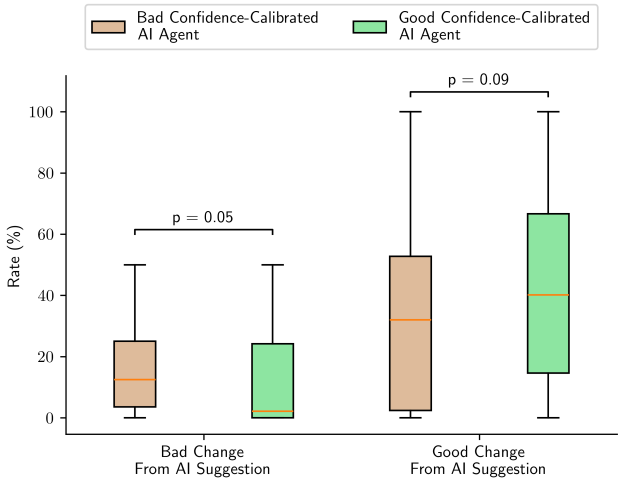

Figure 6: **Rates of Positive and Negative Decision-Making Changes by AI Agent Type**: The Bad AI agent results in more bad/negative changes, while the Good AI agent leads to more good/positive changes.

Furthermore, we recorded the rates of positive and negative changes in decision-making. We defined a negative (or bad) change as when a human changes from a correct answer to an incorrect one based on AI suggestions, and a positive (or good) change as when a human changes from an incorrect answer to a correct one after AI agent suggestions. Based on the data, we analyzed how suggestions from Bad and Good AI agent influence human final decisions. The statistical results in Fig. 6 shows that Bad AI agent leads to more negative changes ($t(61) = 1.61, p = .05$), while Good AI agent leads to more positive changes ($t(61) = 1.34, p = .09$). This indicates the importance of ensuring that AI agents are of high quality (i.e., well-calibrated) to maximize positive outcomes. In this context, a high-quality AI agent is one with well-calibrated confidence levels.

6 Discussion & Conclusion

We studied human-AI dyad joint decision-making on a robot teleoperation task. Our research reveals the suboptimality of human meta-decision making, particularly when resolving disagreements

between humans and AI. This discovery highlights the potential for more effective approaches, such as Maximum Confidence Slating (MCS), which leverages confidence as an easily accessible metric. The application of MCS in this context represents a novel contribution to the field of human-AI joint decision-making in robotics. Our study demonstrates that pairing humans with AI, even when the AI is not perfectly confidence-calibrated, can enhance overall decision accuracy. This finding suggests that AI's role extends beyond providing correct answers; it also enhances the decision-making process through confidence sharing. However, we emphasize the critical importance of well-calibrated AI confidence levels. Our findings indicate that a high-quality confidence-calibration AI agent can lead to improved decision outcomes. This highlights the necessity for AI systems to prioritize not only precision but also confidence-aware frameworks to maximize their collaborative potential and influence in complex collaboration environments.

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. Our study lays groundwork for further research on other approaches to improve decision-making in human-AI collaborative decision-making in a more complex collaboration environments like robotics domain, advancing to safer, more efficient, and more reliable robotics operations.

7 Ethical Statement

This research was conducted with strict adherence to ethical guidelines governing human-AI interaction studies. All human participants involved in the experiment provided informed consent, with a full understanding of the study's purpose, procedures, and any potential risks or benefits. Data collected from the experiment were anonymized and used solely for the purpose of this research. The study was reviewed and approved by the University Research Ethics Committee, confirming that all ethical standards were met.

References

Dylan M. Asmar and Mykel J. Kochenderfer. Collaborative decision making using action suggestions, 2022. URL https://arxiv.org/abs/2209.13160.

Bahador Bahrami, Karsten Olsen, Peter E Latham, Andreas Roepstorff, Geraint Rees, and Chris D Frith. Optimally interacting minds. *Science*, 329(5995):1081–1085, 2010.

Bahador Bahrami, Karsten Olsen, Dan Bang, Andreas Roepstorff, Geraint Rees, and Chris Frith. Together, slowly but surely: the role of social interaction and feedback on the build-up of benefit in collective decision-making. *Journal of Experimental Psychology: Human Perception and Performance*, 38(1):3, 2012.

Dan Bang, Riccardo Fusaroli, Kristian Tylén, Karsten Olsen, Peter E Latham, Jennifer YF Lau, Andreas Roepstorff, Geraint Rees, Chris D Frith, and Bahador Bahrami. Does interaction matter? testing whether a confidence heuristic can replace interaction in collective decision-making. *Consciousness and cognition*, 26:13–23, 2014.

Dan Bang, Laurence Aitchison, Rani Moran, Santiago Herce Castanon, Banafsheh Rafiee, Ali Mahmoodi, Jennifer YF Lau, Peter E Latham, Bahador Bahrami, and Christopher Summerfield. Confidence matching in group decision-making. *Nature Human Behaviour*, 1(6):0117, 2017.

Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-ai team performance. In *AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pp. 2–11, 2019a.

Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *AAAI Conference on Artificial Intelligence*, volume 33, pp. 2429–2437, 2019b.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *2021 Conference on Human Factors in Computing Systems (CHI)*, pp. 1–16, 2021.

Raunak Bhattacharyya, Duc An Nguyen, Clara Colombatto, Stephen Fleming, Ingmar Posner, and Nick Hawes. Towards intelligent decision support systems in robotics: Investigating the role of self-confidence calibration in joint decision-making. In *AAAI Spring Symposium Series (AAAI 2024)*, 2024.

Saugat Bhattacharyya, Davide Valeriani, Caterina Cinel, Luca Citi, and Riccardo Poli. Anytime collaborative brain–computer interfaces for enhancing perceptual group decision-making. *Scientific reports*, 11(1):17008, 2021.

Jennifer Casper and Robin R. Murphy. 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, 2003.

Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. Are two heads better than one in ai-assisted decision making? comparing the behavior and performance of groups and individuals in human-ai collaborative recidivism risk assessment. In *Conference on Human Factors in Computing Systems (CHI)*, 2023.

Manolis Chiou, Nick Hawes, and Rustam Stolkin. Mixed-initiative variable autonomy for remotely operated mobile robots. *ACM Transactions on Human-Robot Interaction*, 10(4):1–34, 2021.

Anthony Corso, Peter Du, Katherine Driggs-Campbell, and Mykel J. Kochenderfer. Adaptive stress testing with reward augmentation for autonomous vehicle validatio. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pp. 163–168, 2019. doi: 10.1109/ITSC.2019.8917242.

Seliem El-Sayed, Canfer Akbulut, Amanda McCroskery, Geoff Keeling, Zachary Kenton, Zaria Jalan, Nahema Marchal, Arianna Manzini, Toby Shevlane, Shannon Vallor, Daniel Susser, Matija Franklin, Sophie Bridgers, Harry Law, Matthew Rahtz, Murray Shanahan, Michael Henry Tessler, Arthur Douillard, Tom Everitt, and Sasha Brown. A mechanism-based approach to mitigating harms from persuasive generative ai, 2024. URL https://arxiv.org/abs/2404.15058.

Amelia Glaese, Nat McAleese, Maja Trebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

Douglas Guilbeault, Samuel Woolley, and Joshua Becker. Probabilistic social learning improves the public's judgments of news veracity. *Plos One*, 16(3), 2021.

Daisuke Hamada, Masataka Nakayama, and Jun Saiki. Wisdom of crowds and collective decision-making in a survival situation with complex information integration. *Cognitive Research: Principles and Implications*, 5(1):1–15, 2020.

Nick Hawes, Aaron Sloman, Jeremy Wyatt, Michael Zillich, Henrik Jacobsson, Geert-Jan M Kruijff, Michael Brenner, Gregor Berginc, and Danijel Skocaj. Towards an integrated robot with multiple cognitive functions. In *AAAI*, volume 7, pp. 1548–1553, 2007.

Nick Hawes, Christopher Burbridge, Ferdian Jovan, Lars Kunze, Bruno Lacerda, Lenka Mudrova, Jay Young, Jeremy Wyatt, Denise Hebesberger, Tobias Kortner, et al. The strands project: Long-term autonomy in everyday environments. *IEEE Robotics & Automation Magazine*, 24(3):146–156, 2017.

Patrick Hemmer, Sebastian Schellhammer, Michael Vössing, Johannes Jakubik, and Gerhard Satzger. Forming effective human-ai teams: building machine learning models that complement the capabilities of multiple experts. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2022.

Carl-Johan Hoel, Katherine Driggs-Campbell, Krister Wolff, Leo Laine, and Mykel J. Kochenderfer. Combining planning and deep reinforcement learning in tactical decision making for autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 5(2):294–305, 2020. doi: 10.1109/TIV.2019.2955905.

Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. Towards unbiased and accurate deferral to multiple experts. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2021.

Stephanie Kim, Jacy Reese Anthis, and Sarah Sebo. A taxonomy of robot autonomy for human-robot interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 381–393, 2024.

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

Lars Kunze, Nick Hawes, Tom Duckett, Marc Hanheide, and Tomáš Krajník. Artificial intelligence for long-term robot autonomy: A survey. *IEEE Robotics and Automation Letters*, 3(4):4023–4030, 2018.

Ralf HJM Kurvers, Stefan M Herzog, Ralph Hertwig, Jens Krause, Patricia A Carney, Andy Bogart, Giuseppe Argenziano, Iris Zalaudek, and Max Wolf. Boosting medical diagnostics by pooling independent judgments. *National Academy of Sciences*, 113(31):8777–8782, 2016.

Vivian Lai and Chenhao Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Conference on Fairness, Accountability, and Transparency*, pp. 29–38, 2019.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, Rose E. Wang, Minae Kwon, Joon Sung Park, Hancheng Cao, Tony Lee, Rishi Bommasani, Michael Bernstein, and Percy Liang. Evaluating human-language model interaction, 2024. URL https://arxiv.org/abs/2212.09746.

HCCH Levitt. Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical society of America*, 49(2B):467–477, 1971.

Zhuoran Lu and Ming Yin. Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In *2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2021.

Sébastien Massoni and Nicolas Roux. Optimal group decision: A matter of confidence calibration. *Journal of Mathematical Psychology*, 79:121–130, 2017.

Kiana Jafari Meimandi, Matthew L Bolton, and Peter A Beling. Human-agent teaming: A system-theoretic overview. *Authorea Preprints*, 2024.

A Nguyen, Raunak Bhattacharyya, Clara Colombatto, Steve Fleming, Ingmar Posner, and Nick Hawes. Group decision-making in robot teleoperation: Two heads are better than one. ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2025a.

Duc-An Nguyen, Jude Nwadiuto, Hiroyuki Okuda, and Tatsuya Suzuki. Model structure identification of hybrid dynamical systems based on unsupervised clustering and variable selection. *IFAC-PapersOnLine*, 53(2):1090–1095, 2020a.

Duc-An Nguyen, Jude Nwadiuto, Hiroyuki Okuda, and Tatsuya Suzuki. Modeling car-following behavior in downtown area based on unsupervised clustering and variable selection method. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 3714–3720. IEEE, 2020b.

Duc-An Nguyen, Raunak Bhattacharyya, Clara Colombatto, Steve Fleming, Ingmar Posner, and Nick Hawes. Group decision-making in robot teleoperation: Two heads are better than one. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 489–500, 2025b.

Duc-An Nguyen, Raunak Bhattacharyya, Clara Colombatto, Steve Fleming, Ingmar Posner, and Nick Hawes. Group decision-making in robot teleoperation: Two heads are better than one. In *IEEE/ACM International Conference on Human-Robot Interaction (HRI'25)*, March 2025c.

Adam Norton, Willard Ober, Lisa Baraniecki, Eric McCann, Jean Scholtz, David Shane, Anna Skinner, Robert Watson, and Holly Yanco. Analysis of human–robot interaction at the darpa robotics challenge finals. *The International Journal of Robotics Research (IJRR)*, 36(5-7):483–513, 2017.

Gali Noti and Yiling Chen. Learning when to advise human decision makers. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2023.

Thi Minh Anh Pham, An Duc Nguyen, Cephas Svosve, Vasileios Argyriou, and Georgios Tzimiropoulos. Pre: Vision-language prompt learning with reparameterization encoder. *arXiv preprint arXiv:2309.07760*, 2023.

Amy Rechkemmer and Ming Yin. When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models. In *2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2022.

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

Marion Rouault, Andrew McWilliams, Micah G Allen, and Stephen M Fleming. Human metacognition across domains: insights from individual differences and neuroimaging. *Personality Neuroscience*, 1:e17, 2018.

Lindsay Sanneman and Julie A Shah. A situation awareness-based framework for design and evaluation of explainable ai. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems: Second International Workshop, EXTRAAMAS 2020, Auckland, New Zealand, May 9–13, 2020, Revised Selected Papers 2*, pp. 94–110. Springer, 2020.

Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D Manning. Do massively pretrained language models make better storytellers? *arXiv preprint arXiv:1909.10705*, 2019a.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. What makes a good conversation? how controllable attributes affect human judgments. *arXiv preprint arXiv:1902.08654*, 2019b.

Sanjana Srivastava, Chengshu Li, Michael Lingelbach, Roberto Martín-Martín, Fei Xia, Kent Elliott Vainio, Zheng Lian, Cem Gokmen, Shyamal Buch, Karen Liu, Silvio Savarese, Hyowon Gweon, Jiajun Wu, and Li Fei-Fei. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In Aleksandra Faust, David Hsu, and Gerhard Neumann (eds.), *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*. PMLR, 2022. URL https://proceedings.mlr.press/v164/srivastava22a.html.

Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. The metacognitive demands and opportunities of generative ai. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–24, 2024.

Xinru Wang, Chen Liang, and Ming Yin. The effects of ai biases and explanations on human decision fairness: a case study of bidding in rental housing markets. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2023.

Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. How do visual explanations foster end users' appropriate trust in machine learning? In *The 25th International Conference on Intelligent User Interfaces*, pp. 189–201, 2020.

Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *2019 Conference on Human Factors in Computing Systems (CHI)*, pp. 1–12, 2019.

Alice Qian Zhang, Ryland Shaw, Jacy Reese Anthis, Ashlee Milton, Emily Tseng, Jina Suh, Lama Ahmad, Ram Shankar Siva Kumar, Julian Posada, Benjamin Shestakofsky, et al. The human factor in ai red teaming: Perspectives from social and collaborative computing. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, pp. 712–715, 2024.

Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Conference on Fairness, Accountability, and Transparency*, 2020.

A APPENDIX

In the appendix, we present detailed descriptions of the two implemented algorithms.

Algorithm 1 Confidence Calibration (AUROC2)

Input:
  $correct$: vector of size $1 \times n_{trials}$, with 0 for error and 1 for correct trials
  $conf$: vector of size $1 \times n_{trials}$, with confidence ratings from 1 to $N_{ratings}$
  $N_{ratings}$: number of available confidence levels
Output: $auroc2$: type-2 area under the ROC curve
Initialize $i \leftarrow N_{ratings} + 1$
for $c \leftarrow 1$ **to** $N_{ratings}$ **do**
  $H2[i-1] \leftarrow \text{count}(conf = c \wedge correct) + 0.5$
  $FA2[i-1] \leftarrow \text{count}(conf = c \wedge \neg correct) + 0.5$
  $i \leftarrow i - 1$
end
Normalize $H2 \leftarrow H2 / \sum H2$
Normalize $FA2 \leftarrow FA2 / \sum FA2$
Compute cumulative sums:
  $csum\_H2 \leftarrow [0, \text{cumsum}(H2)]$
  $csum\_FA2 \leftarrow [0, \text{cumsum}(FA2)]$
Initialize $i \leftarrow 1$
for $c \leftarrow 1$ **to** $N_{ratings}$ **do**
  $A \leftarrow csum\_H2[c+1] - csum\_FA2[c]$
  $B \leftarrow csum\_H2[c] - csum\_FA2[c+1]$
  $k[i] \leftarrow A^2 - B^2$
  $i \leftarrow i + 1$
end
Compute $auroc2 \leftarrow 0.5 + 0.25 \times \sum k$
return $auroc2$

Algorithm 2 Maximum-Confidence Slating (MCS)

Input: Decision set $\mathcal{D} = \{\text{D}_\text{h}, \text{D}_\text{AI}\}$,
  Confidence set $\mathcal{C} = \{\text{C}_\text{h}, \text{C}_\text{AI}\}$, Set of trials T
Output: MCS decisions $\{\text{MCS}(t)\}_{t \in T}$
MCS $\leftarrow \emptyset$
for each trial $t \in T$ **do**
  $C_\text{h} \leftarrow \mathcal{C}[\text{C}_\text{h}, t]$
  $C_\text{AI} \leftarrow \mathcal{C}[\text{C}_\text{AI}, t]$
  $D_\text{h} \leftarrow \mathcal{D}[\text{D}_\text{h}, t]$
  $D_\text{AI} \leftarrow \mathcal{D}[\text{D}_\text{AI}, t]$
  while $D_\text{h} \neq D_\text{AI}$ **then**
    if $C_\text{h} > C_\text{AI}$ **then**
      $\text{MCS}(t) \leftarrow D_\text{h}$
    else if $C_\text{AI} > C_\text{h}$ **then**
      $\text{MCS}(t) \leftarrow D_\text{AI}$
    end if
end for
return $\{\text{MCS}(t)\}_{t \in T}$