Metacognition for Joint Inference about Robot Control



Nguyen Duc An

Linacre College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2025

Acknowledgements

I'm deeply grateful to have had such a wonderful group of people supporting me throughout this DPhil journey.

I owe my sincerest thanks to Professor Nick Hawes and Professor Raunak Bhattacharya for their guidance and support during my research at Oxford. Thank you, Nick and Raunak, for every opportunity, for believing in me, and for teaching me not just how to begin my research, but how to gradually take the lead and grow into an independent researcher. I feel incredibly lucky to have been part of such a dream team.

Special thanks to my collaborators - Professor Ingmar Posner, Professor Steve Fleming, and Professor Clara Colombatto - for sharing your expertise in cognitive science and for your valuable feedback on my papers and ideas. I am also grateful to the EPSRC Programme Grant From Sensing to Collaboration (EP/V000748/1) and Amazon Web Services (AWS) for their generous support, which made much of my research (and conference travels) possible. A big thank you to Matt Hansen and the AWS team for working with us to enable the software platform - something we could never have completed so smoothly without your support.

My years in Oxford have been filled with wonderful friends and colleagues. I want to thank Bruno, Mihaela, Karan, Clarissa, Alex, Carl, Cephas, Grid, and everyone in the GOALS Research Group and the wider ORI for all the chats, sparring, and laughter. Our conversations were always thought-provoking and often left me with renewed motivation and excitement to do great research. Talking with you pulled me back to the surface whenever I slipped into "robot mode", lost in code, experiments, or writing for hours on end.

My sincerest gratitude goes to my fiancée, Minh Anh, who has stood by me through every stage of this journey. Her kindness, understanding, and support have been a constant source of strength. Finally, I am endlessly grateful to my family for their love and support in every way. I could not have made it through this DPhil, or anything ever in fact, without them. Thank you to Mum, Dad, and Chi! Dad, thank you for sharing your experiences from your own PhD years; your insight often lit the way whenever I needed to be understood. Mum, thank you for always discussing math with me in the most cheerful way possible. I cannot express how much I appreciate everything you've done quietly to support me throughout this journey.

Abstract

With the increasing prevalence of semi and fully autonomous robots across various fields, including safety-critical domains like nuclear surveillance and decommissioning, healthcare (e.g., surgical robots), space exploration, and military applications, as well as daily life settings such as manufacturing (e.g., humanoid robots) and transportation (e.g., delivery robots), human operators are repeatedly faced with making choices between different options of robot control, such as selecting a robot for completing the task or deciding whether to operate a robot manually or autonomously. The continuous cognitively demanding process of making these choices over time, under various sources of environmental uncertainty, can lead to cognitive fatigue and errors, with poor decisions resulting in inefficiencies, safety risks, or even failures in robotic operations. Furthermore, robotic systems operate in physical environments with temporal constraints and safety implications that amplify the consequences of poor decisions. The spatial and temporal dynamics inherent to robotic tasks create compound uncertainty that traditional collaboration methods struggle to address effectively. Addressing these challenges requires methods that adaptively combine human and AI/robot decision-making. This hinges on the ability of each partner to not only perform the task but also to accurately assess and communicate their own competence and uncertainty, a process known as metacognition. This metacognitive awareness becomes the critical input for joint inference, a decision-making process where judgments from both partners are combined to yield a single, more robust choice.

In this research, we explore how sharing metacognitive information can enhance joint inference in human-robot collaboration during teleoperation missions. We systematically investigate this framework through a progressive, three-stage study grounded in a dynamic robotics task that requires operators to perform active, real-time inference about a robot's latent state. First, we establish a foundational understanding of dyadic collaboration by examining human-human teams sharing metacognitive signals: confidence. Through a large user study (N=100), we demonstrate that joint inference using the heuristic Maximum Confidence Slating (MCS), where the choice of the more confident partner is chosen, improves team accuracy beyond that of the best-performing individual. The study also identifies key moderators of this success, revealing that performance gains are regulated by the partners' skill similarity and their collective metacognitive sensitivity.

Second, this metacognitive signal-sharing framework is extended to human–AI dyads. A subsequent user study (N=100) confirms that the MCS heuristic remains effective for joint inference between humans and their AI partners when confidence is shared, outperforming both individual human performance and the algorithmic joint inference baseline. This stage also reveals the critical role of the AI's own metacognitive sensitivity profile, showing that its sensitivity not only impacts team accuracy but also shapes the human partner's decision-making dynamics.

Finally, to overcome the limitations of static rules, we introduce an adaptive joint inference framework based on contextual bandits. We demonstrate that an adaptive policy, which learns to arbitrate between the human and AI by considering contextual factors such as task difficulty and agent-specific metacognitive traits, extending beyond confidence alone, consistently outperforms static heuristics. A key finding is that this adaptive approach is most effective in high-difficulty scenarios, enabling human–AI complementarity precisely when it is most needed. The primary contribution of this thesis is a validated, context-aware framework for adaptive decision arbitration that moves beyond simple assistance to effectively operationalize human-robot complementarity.

Contents

List of Figures **xi**

List of Tables **xv**

List of Abbreviations **xvii**

1 Introduction **1**

1.1 Overview 1
1.2 Contribution 6
1.3 Structure of Thesis 9
1.4 Publications 11

2 Background **13**

2.1 Metacognition 14
2.1.1 Definition and Foundational Frameworks 14
2.1.2 Quantifying Metacognition: Sensitivity & Calibration 18
2.2 Maximum Confidence Slating 23
2.3 Multi-Armed Bandit 24
2.4 Discussion 28

3 Literature Review **29**

3.1 Confidence Sharing in Human Collaboration 30
3.1.1 Confidence Matching 30
3.1.2 Weighted Confidence Sharing 31
3.1.3 Maximum Confidence Slating 33
3.2 Human-Agent Collaboration 34
3.2.1 Human-AI Joint Inference 34
3.2.2 Human-Robot Joint Inference 37
3.3 Engineering Metacognition in AI Systems 42
3.4 Metacognition Experimental Paradigms 44
3.5 Discussion 46

4 Confidence-Based Heuristics for Human-Human Joint Inference 49
4.1 Introduction . . . 49
4.2 Methodology . . . 53
4.2.1 Experiment Setup . . . 53
4.2.2 Procedure . . . 54
4.2.3 Data Collection . . . 55
4.2.4 Metrics . . . 55
4.3 Results . . . 56
4.3.1 Impact of Confidence on Accuracy . . . 57
4.3.2 Effect of Performance Discrepancy . . . 58
4.3.3 Influence of Metacognitive Sensitivity . . . 61
4.3.4 Dyadic Metacognitive Sensitivity Correlation . . . 65
4.4 Discussions . . . 67

5 Confidence-Based Heuristics for Human-AI Joint Inference 71
5.1 Introduction . . . 71
5.2 Methodology . . . 75
5.2.1 Problem Statement . . . 75
5.2.2 Environment and Task . . . 75
5.2.3 AI Decision Support System (AI-DSS) . . . 77
5.2.4 Experimental Procedure . . . 79
5.2.5 Human-AI Joint Inference . . . 80
5.2.6 Performance Metrics . . . 82
5.3 Results . . . 82
5.3.1 Human Inference Dynamics . . . 82
5.3.2 Joint Inference Performance . . . 87
5.4 Discussions . . . 92

6 A Contextual Bandit Framework for Adaptive Joint Inference 95
6.1 Introduction . . . 95
6.2 Methodology . . . 98
6.2.1 The Human &AI-DSS Dyad . . . 99
6.2.2 Human-Subject Experiment . . . 100
6.2.3 Virtual Dyad Pairings . . . 101
6.2.4 Performance Metrics . . . 102
6.2.5 Methods for Human-AI Joint Inference . . . 102
6.3 Results . . . 109
6.3.1 Adaptive Joint Inference Outperforms Heuristics . . . 109
6.3.2 Performance Across Various Difficulty Conditions . . . 110
6.3.3 Latent Variable Modeling of Unobservable Context . . . 112
6.4 Discussion . . . 112

7 Conclusion **115**
7.1 Summary of Contributions 115
7.1.1 Generalizability to Real-World Robotic Tasks 118
7.1.2 Impact on HRI Dynamics 118
7.2 Limitations 119
7.3 Future Work 120

Appendices

A Pilot Study Experiments **125**

B Software Design and Metadata of Human-Human Experiment **127**
B.1 Experiment Data Labels 128
B.2 Metacognitive Sensitivity Data 129

C Software Design and Metadata of Human-AI Experiment **131**
C.1 Experiment Data Variables 132
C.2 Human Participant Demographics 133

References **135**

List of Figures

1.1 Joint Inference Framework Development 1.1a Human executes inference alone. 1.1b Human making inference jointly with another human partners via MCS heuristic. 1.1c Human making inference jointly with an AI-DSS via MCS heuristic. 1.1d Human making inference jointly with an AI-DSS via Bandit algorithms. 7

2.1 The "Choice" is determined by the decision-making system, and the decision uncertainty via "Confidence Levels" is encoded in the metacognition system. 15

4.1 Software Framework and User Study Scenario. Fig. 4.1a presents the scenario with the red and green curves represent trajectories driven by a participant from our user study who was experiencing low and high delay in control of the mobile robot. Fig. 4.1b presents the software framework for the web-based robot navigation simulator. The simulator is hosted on AWS, and data is collected and stored in a MongoDB database. The participants for the user study were recruited via the Prolific Platform. 53

4.2 Study Interface: After teleoperating the two robots, participants choose which robot they deemed to have lower control delay (Fig. 4.2a), and provide their associated confidence (Fig. 4.2b). 54

4.3 Accuracy in robot selection for high-confidence, higher-performing, low-confidence, lower-performing, and random participants from the paired dyads. 58

4.4 Variation in accuracy gain as participants of different success rates are paired together. Pairing similar performing participants leads to higher benefit. 59

4.5 Variation in accuracy gain as participants of different success rates are paired together. Pairing similar performing participants leads to higher accuracy gain. 60

4.6 Variation in accuracy gain as participants of different performance levels are paired together. Performance level is assessed by the difficulty of the task as established by delay plateau difference. Pairing similar performing participants leads to higher accuracy gain. 61
4.7 Metacognitive sensitivity of two different participants. 4.7a: Plot of trial result vs. confidence rating. The line is the regression fit with 95% confidence interval. 4.7b: ROC curves for both participants. The participant represented by red demonstrates higher metacognitive sensitivity compared to the participant represented in blue. 62
4.8 Choice accuracy (%) of dyads with both participants of good metacognitive sensitivity (288 pairs). 62
4.9 Choice accuracy (%) of dyads with both participants of poor metacognitive sensitivity (242 pairs). 63
4.10 Choice accuracy (%) of dyads with one participant of good metacognitive sensitivity paired with one of poor metacognitive sensitivity (264 pairs). 63
4.11 Variation in accuracy gain as participants of different metacognitive sensitivity are paired together. For good metacognitive sensitivity (red), pairing participants of similar metacognitive sensitivity leads to higher benefit. Pairing participants with similarly poor metacognitive sensitivity (blue) does not yield a benefit. 64
4.12 Trial results versus confidence ratings for an example virtual dyad: Red indicates a dyad formed by pairing participants of good metacognitive sensitivity, while blue represents a dyad formed by pairing participants of poor metacognitive sensitivity. 66
4.13 Variation in dyadic metacognitive sensitivity based on how participants are paired according to their individual sensitivity. 67
4.14 Dyadic metacognitive sensitivity versus accuracy gain. Higher dyadic metacognitive sensitivity leads to an increased accuracy gain. 68

5.1 Software Framework and User Study Scenario. Fig. 5.1a presents the scenario with The red and green curves represent trajectories driven by a participant from our user study who was experiencing low and high delay in control of the mobile robot. Fig. 5.1b presents the software framework for the web-based robot navigation simulator. The simulator is hosted on AWS, and data is collected and stored in a MongoDB database. The participants for the user study were recruited via the Prolific Platform. 76

5.2 Study Interface: After teleoperating the two robots, participants make an initial inference and provide associated confidence (left panel). Then, the AI-DSS provides its inference and confidence (middle panel). If the participant decides to change their initial inference, they provide an updated inference and associated confidence (right panel). The joint human-AI inference is then decided using MCS (top right). 77
5.3 Histograms of human confidence levels associated with correct and incorrect inferences and separated by metacognitive sensitivity of the participants. The delay difference between the two robots was 60 ms in this case. 78
5.4 Higher confidence scores from the AI-DSS result in a higher number of inferences changed by humans. 83
5.5 Dynamics of human inferences: Difference in *confidence levels* between humans and AI-DSS influences whether humans change from incorrect to correct inferences. 84
5.6 Dynamics of human inferences: AI-DSS metacognitive sensitivity influences whether humans change from incorrect to correct inferences. 85
5.7 Dynamics of human confidence: AI *metacognitive sensitivity* influence on human confidence changes and joint inference correctness. 86
5.8 Dynamics of human confidence: AI *metacognitive sensitivity* influence on human confidence changes and joint inference correctness. 87
5.9 Misalignment in human and AI confidence ratings leads to degradation in accuracy of human-initiative joint inference. 88
5.10 Joint human-AI inference accuracy is highest with MCS, outperforming purely human individual inference, joint inference dictated by human, and approaches using Thompson Sampling (TS) bandit. . . 89
5.11 Influence of varying AI type on team inference accuracy. 90
5.12 Inference accuracy obtained by pairing human participants with the poor metacognitive sensitivity AI-DSS. 91

6.1 Software Framework and User Study Scenario. Fig. 6.1a presents the scenario with The red and green curves represent trajectories driven by a participant from our user study who was experiencing low and high delay in control of the mobile robot. Fig. 6.1b presents the software framework for the web-based robot navigation simulator. The simulator is hosted on AWS, and data is collected and stored in a MongoDB database. The participants for the user study were recruited via the Prolific Platform. 98

6.2 Overall Joint Inference Accuracy: The performance of all tested methods across all trials where the human and AI disagreed. The adaptive bandit models incorporating task context (OLRB-Metacog, OLRB-ConfDiff, OLVB) demonstrate higher median accuracies than the static MCS heuristic and the Human baseline. 109
6.3 Joint Inference Accuracy Across Task Difficulty Levels: This figure illustrates the resilience of adaptive models compared to static baselines. While the performance of Human and MCS collapses on High and Extreme difficulty tasks, the top adaptive bandits (OLRB-Metacog, OLRB-ConfDiff, OLVB) maintain consistently high accuracy. 111

B.1 Histogram of confidence ratings over all trials over the entire set of participants. 129
B.2 Histogram of AUROC2 values over the entire set of participants. . . 130
B.3 Corresponding AUROC Curve for an Example Participant. 130

List of Tables

1.1 A description of each chapter's methodological focus and contribution to joint inference research. 9

2.1 Description of background knowledge connection to each chapter's methodological focus. 28

4.1 A description of Chapter 4's contribution to joint inference research. 52

5.1 A description of Chapter 5's contribution to joint inference research. 74

6.1 A description of Chapter 6's contribution to joint inference research. 97

6.2 Summary of Joint Inference Accuracy (Mean ± SD) Across All Conditions. 112

A.1 Success rate for each pilot scenario. 125

B.2 Description of variables used in our human-human joint inference study. 128

C.2 Description of variables used in our human-AI joint inference study. 132

List of Abbreviations

2HBT1 . . . . Two-Heads-Better-Than-One

AI . . . . . Artificial Intelligence

AI-DSS . . . . AI Decision Support System

AUC . . . . . Area Under the Curve

AUROC2 . . . Area Under the Receiver Operating Characteristic Type 2 Curve

BDL . . . . . Bayesian Deep Learning

DLC . . . . . Dummy Low Confidence

DR . . . . . Dummy Random

ECE . . . . . Expected Calibration Error

EEG . . . . . Electroencephalography

ERPs . . . . . Event-Related Potentials

fNIRS . . . . . Functional Near-Infrared Spectroscopy

GSR . . . . . Galvanic Skin Response

HCI . . . . . Human-Computer Interaction

HHI . . . . . Human-Human Interaction

HP . . . . . High-Performing (individual in a dyad)

HRI . . . . . Human-Robot Interaction

i.i.d. . . . . . Independent and Identically Distributed

JS . . . . . Jensen-Shannon Divergence

LLMs . . . . . Large Language Models

LP . . . . . Low-Performing (individual in a dyad)

MAB . . . . . Multi-Armed Bandit

MC . . . . . Monte Carlo

MCS . . . . . Maximum Confidence Slating

MOMDPs . . Mixed Observability Markov Decision Processes

NASA-TLX . NASA-Task Load Index

OLRB . . . . . Online Linear Regression Bandits

OLVB . . . . . Online Latent Variable Bandit

POMDP . . . Partially Observable Markov Decision Process

RL . . . . . Reinforcement Learning

ROC2 . . . . . Type 2 Receiver Operating Characteristic

SA . . . . . Shared Autonomy

SC . . . . . Skin Conductance

SDT . . . . . Signal Detection Theory

SMM . . . . . Shared Mental Model

SVI . . . . . Stochastic Variational Inference

TS . . . . . Thompson Sampling

UAV . . . . . Unmanned Aerial Vehicle

UQ . . . . . Uncertainty Quantification

WCS . . . . . Weighted Confidence Sharing

1 Introduction

Contents

1.1 Overview 1
1.2 Contribution 6
1.3 Structure of Thesis 9
1.4 Publications 11

1.1 Overview

The integration of advanced robots - embodied AI agents with perception, reasoning, and actuation - into human-centric domains is creating a new frontier in engineering and design [1]. These systems are increasingly deployed in dynamic, high-stakes environments alongside humans (e.g, search and rescue, nuclear surveillance, space exploration) where failures may not always be immediately apparent [2–9]. In such settings, delays, errors, and other latent faults can disrupt coordination, making effective human-embodied AI collaboration difficult. In search and rescue operations or nuclear surveillance, environmental changes can degrade sensor reliability without obvious indicators [10, 11]. Medical robotic systems face patient-specific variations that create diagnostic uncertainty [12–14]. These scenarios share a common thread: the need for real-time decision-making under uncertainty, where neither human nor

AI agent has perfect information. Unlike static AI applications, robotic systems operate in physical environments with temporal constraints and safety implications that amplify the consequences of poor decisions. The spatial and temporal dynamics inherent to robotic tasks create compound uncertainty that traditional human-AI collaboration methods struggle to address effectively [15]. Addressing these challenges requires methods that adaptively combine human and AI decision-making. A fundamental component of this is joint inference, a single-step decision-making process where judgments from both partners are combined.

To make this collaboration operational, the system needs a way to weigh the human's intuitive assessment against its own analysis. The key is **confidence**. Confidence is a relatively intuitive and communicable metric for both humans and AI systems to express the perceived reliability of their information or assessments. In the human neuroscience and psychology field, confidence ratings are commonly used as a proxy for *metacognitive monitoring*, the process by which individuals (or agents) evaluate the quality or correctness of their decisions or cognitive states. For example, in decision-making tasks, a high confidence rating typically indicates that the individual believes their choice is accurate. Humans also use confidence judgments to adapt their behavior in *metacognitive control* processes. For instance, when confidence is low, a person might seek additional information, revise their decision, or defer to a partner or AI system. Similarly, AI systems can adjust their decision thresholds or solicit external input when their internal confidence estimates are low. It acts as a proxy for the quality of the underlying data or inference process.

These monitoring and control metacognitive processes, or the ability of "thinking about thinking" of any agent, whether human or AI, to evaluate the certainty of its own conclusions is an aspect of what's known as **metacognition**. Insights from fundamental studies of metacognition, particularly in fields like neuroscience and psychology, offer a solid foundation for understanding how to build teams - including human-human teams - that can make better decisions together. These metacognitive ideas, especially the concept of sharing self-assessments like confidence, motivate

and guide the approach in our research. Specifically, the quality of each agent's self-assessment becomes critically important. This involves aspects like "metacognitive sensitivity" (how well an agent can distinguish between its correct and incorrect judgments) and "metacognitive calibration" (how well an agent's stated confidence matches its actual accuracy). Despite the recognized importance of these nuanced metacognitive factors in human decision-making, there has been a noticeable lack of research applying them systematically to improve joint inference within mixed-initiative human-AI systems. Therefore, this thesis aims to address this gap by focusing on how joint inference capabilities in human-robot collaboration can be improved by explicitly considering these two axes of metacognition: metacognitive sensitivity and metacognitive calibration. The core objective is to incorporate metacognition in developing and evaluating heuristics and algorithms that enhance mixed-initiative inference between humans and AI Decision Support Systems (AI-DSS) models, particularly for dynamic spatiotemporal robotics missions.

To ground our investigation, we focus on a representative latent fault: operational delay in robot teleoperation. This delay manifests as a subtle, time-varying difference in control latency between two robots and serves as an example of hidden system degradation that emerges as a continuous deviation from expected performance, rather than as a discrete failure event such as a total sensor blackout. Such degradations are often subtle, continuous, and time-varying, arising from factors such as software performance variability, network fluctuations, or gradual component degradation. They cannot be detected by examining sensor values at a single moment; instead, differential delay must be inferred through temporal interaction, pattern recognition, and active inference over time, making it a demanding task that strains static decision-making strategies [16, 17]. This scenario connects directly to real operational challenges in distributed robotic systems, where network conditions, processing loads, or hardware variations can create performance asymmetries that are not immediately apparent but may significantly affect mission effectiveness. The choice of control delay as the primary latent variable also offers advantages over other diagnostic tasks, such as discrete sensor failures or object classification

errors. Unlike binary hardware faults, which are often detectable through system status flags, control delay is closely tied to the interaction loop between the human and the machine. Whereas classification tasks primarily test static perceptual accuracy, delay requires the operator to monitor the fluency of the control-to-feedback loop over time. While specifically a temporal phenomenon, this delay serves as a high-fidelity proxy for broader categories of hidden system degradation. The challenge generalizes to any robotic failure where the fault is not a discrete, binary event, such as a total sensor blackout, but rather a subtle, continuous deviation from expected performance.

We frame the task of identifying which robot exhibits such lower latent faults (control delay) as a joint inference problem: a collaborative process in which human operators and their partners, whether other humans or AI systems, form independent inferences that are then combined into a final team decision. Our investigation leverages metacognition principles to structure and support this process. To this end, we employ active robot control in a simulated operating environment as a testbed for developing and evaluating models of human–robot collaboration. The novelty of our approach can be distinguished in three key ways.

Firstly, by drawing insights from the well-established domains of neuroscience and psychology, our work seeks to provide more than just a task-specific solution. We aim to establish a foundational, principled framework for understanding and improving joint inference in human-robot systems. This approach allows our findings to be generalizable beyond the specific task of detecting lower operational delays, offering a new lens through which to analyse a wide array of human-AI collaboration challenges. Consequently, we anticipate that our results will not only validate our proposed algorithms but also provide concrete evidence for a metacognition-based approach, paving the way for future research into adaptive human-AI teams.

Secondly, our task design extends beyond traditional Signal Detection Theory (SDT) paradigms, which examines static discrimination between fixed stimuli in controlled settings [18]. In contrast, the “signal” in our context is a latent, time-varying operational delay that must be inferred through active control and continuous

observation of robot behavior. This shifts the challenge from passive classification to active inference, where human decisions are shaped not only by perceptual input but also by their own control actions and the feedback those actions generate.

Lastly, while human–AI collaboration has been studied in contexts involving static prediction and recommendation real-world robotic missions, however, demand that human supervisors infer unobserved variables like control delay within unfolding temporal processes, all while actively interacting with the system through control actions. This convergence of challenges-latent state, temporal dynamics, and action-based inference-has not been systematically tackled together in the existing literature. Our work directly addresses this gap, positioning teleoperation delay as a representative case through which broader insights into human–AI teaming can be developed.

Our findings reveal a progression from simple heuristics to adaptive frameworks. We first show that a confidence-based rule - Maximum Confidence Slating (MCS) - is effective for and significantly improving joint inference accuracy when both applied to human-human and human-AI dyads. Building on this foundation, we introduce adaptive contextual bandit algorithms that arbitrate decisions sequentially, outperforming the static MCS baseline, particularly under high-difficulty task conditions. These adaptive models demonstrate that incorporating metacognitive signals enables robust performance without relying on external indicators of task difficulty, offering a principled pathway for real-world deployment.

In summary, our research demonstrates a trajectory from validating a foundational metacognition-based heuristic to establishing an adaptive, resilient framework for human–AI collaboration in robotic missions. By addressing latent variables, temporal processes, and action-based inference together, we provide both a concrete solution to the challenge of teleoperation delay and a generalizable framework for understanding human–AI teamwork. These results highlight the value of metacognitive signals as a foundation for adaptive decision-making, paving the way for more effective, trustworthy, and resilient human–AI teams in dynamic robotic environments.

1.2 Contribution

To advance human-robot teaming beyond simple assistance, this thesis addresses the problem of joint inference about a robot's control delays by focusing on two key axes of metacognition: metacognitive sensitivity and metacognitive calibration. Here, simple assistance refers to the traditional AI-assisted decision-making paradigm in which the AI functions as a static advisor, offering recommendations that a human operator may accept or reject while remaining the default final arbiter regardless of current competence. This research moves beyond that model by proposing a peer-to-peer framework in which joint inference becomes an adaptive, bidirectional process: the initiative, the authority to make the final choice, can shift between agents. The goal is to develop a joint inference framework that enhances human-AI mixed-initiative collaboration in dynamic robot missions. We investigate this within the specific task of inferring the latent delay difference between two actively controlled robots in simulation.

Our research is guided by the following key questions:

RQ1: Human & Human Complementarity. To what extent can human dyads achieve joint inferencing complementarity by sharing confidence about robot control delays? How do individual differences in metacognitive sensitivity affect the performance of this collaboration?

RQ2: Human & AI-DSS Complementarity. Can the principles of confidence-sharing be extended to human & AI-DSS dyads to achieve similar complementarity in the same robotics task? How does the metacognitive profile of the AI and its alignment with the human partner affect team performance?

RQ3: Adaptive Joint Inference. Can adaptive, sequential decision-making algorithms enhance joint inference performance beyond static heuristics? What constitutes the most effective contextual information for these algorithms to optimize team performance, particularly in challenging scenarios?

The chronological progression of the research in developing the joint inferencing framework is illustrated in Figure 1.1.

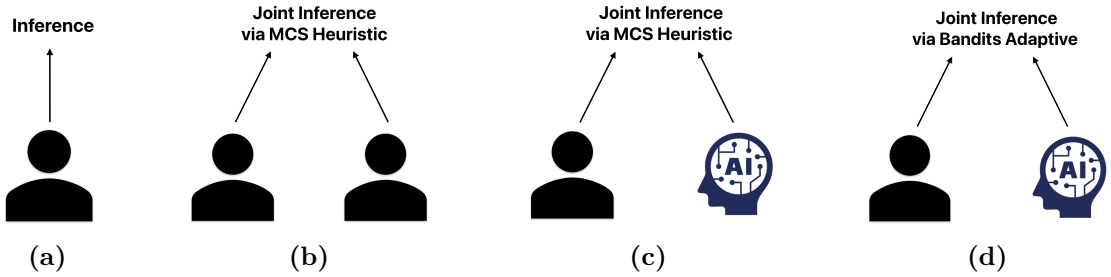


Figure 1.1: Joint Inference Framework Development
1.1a Human executes inference alone.
1.1b Human making inference jointly with another human partners via MCS heuristic.
1.1c Human making inference jointly with an AI-DSS via MCS heuristic.
1.1d Human making inference jointly with an AI-DSS via Bandit algorithms.

In response to these questions, within our research, we employ a simple yet controllable AI Decision Support System (AI-DSS) where confidence outputs are sampled from human data distributions. This methodology allows for precise experimental control over the AI's metacognitive characteristics, and isolate its effects, while serving as a robust placeholder for future integration with more advanced diagnostic models. The main contributions of this thesis are as follows:

Complementarity in the Joint Inferencing of Human & Human dyads:

We establish that confidence-based heuristics can achieve human-human complementarity in a latent variable inference task and identify the metacognitive factors that govern its success. We first demonstrate that enabling joint inference via the Maximum Confidence Slating (MCS) heuristic improves a dyad's performance beyond that of its better-performing member. This finding is substantiated by a large user study ($N = 100$). We further show that this benefit is regulated by specific factors: smaller gains are seen in dyads with large skill discrepancies, while higher collective metacognitive sensitivity and calibration are correlated with greater joint inferencing accuracy.

Complementarity in the Joint Inferencing of Human & AI-DSS dyads:

We provide the first demonstration of extending this metacognitive framework to human-AI dyads in this task, revealing the critical role of AI-DSS metacognition. We show that the MCS heuristic again leads to higher joint inference accuracy compared to both individual human performance and traditional AI-assisted paradigms. This finding is substantiated by another user study ($N = 100$). Crucially, the AI-DSS's own metacognitive sensitivity not only regulates team accuracy but also influences the human partner's own inference and confidence dynamics. This work also highlights the robustness of the MCS heuristic, as it remains effective even in non-ideal pairings where both the human and the AI exhibit poor metacognitive sensitivity.

Advancing Adaptive Joint Inference with Contextual Bandits:

We develop an adaptive joint inference framework using contextual bandits that significantly outperforms static heuristics. Moving beyond MCS, we demonstrate that sequential joint inference guided by contextual bandit algorithms consistently surpasses the performance of the MCS heuristic.

We also identify a subset of specific contextual information required to unlock the potential of adaptive algorithms. While confidence is an important contextual input, the research found that relying solely on confidence as the contextual information for the bandit algorithms did not yield a significant advantage over simpler heuristic methods like MCS. This suggests that while confidence is valuable, its isolated use may not fully capture the complexities needed for optimal joint inference in a sequential setting. However, a substantial improvement in the performance of the contextual bandit algorithms was observed when confidence was combined with task difficulty levels as components of the contextual vectors. A key contribution here is the significant accuracy gain achieved by the contextual bandit method in tasks of high difficulty, an advantage not consistently demonstrated by the MCS heuristic alone. This highlights the potential of contextual bandits to facilitate human &

AI-DSS complementarity precisely where it is most needed - in challenging scenarios where individual performance or simpler heuristics might otherwise falter.

The description of each chapter's methodological focus and contribution to joint inference research is summarised in Table. 1.1.

Table 1.1: A description of each chapter's methodological focus and contribution to joint inference research.

Dimension	**Chapter 4**	**Chapter 5**	**Chapter 6**
Collaboration Type	Human-Human dyads	Human-AI dyads	Human-AI dyads
Heuristic?	Yes	Yes	No
Adaptive?	No	Partially	Yes
Bandit Approach?	None	Thompson Sampling (MAB)	Contextual & Latent Bandit
Focus of Study	Metacognitive sensitivity & heuristics in joint inference	AI-assisted decision making under uncertainty	Unified framework with context awareness

1.3 Structure of Thesis

In Chapter 2.3 we introduce the core concepts of metacognition, detailing how it can be quantified through measures of *sensitivity* and *calibration*. We then formally define the Maximum Confidence Slating (MCS) heuristic and provide a technical overview of the multi-armed and contextual bandit algorithms that are central to our later contributions.

Building on this foundation, in Chapter 3.5, we discuss the relevant literature, including prior research in human-human interaction, mixed-initiative human-robot collaboration, and algorithmic joint inference. This review serves to identify the research gaps that our work aims to address.

The empirical work begins in Chapter 4.4, which presents our first investigation into confidence-based heuristics for joint inference in human-human dyads. This chapter directly addresses our first research question (RQ1) by testing the efficacy of the MCS heuristic in our experimental task and identifying the metacognitive factors that regulate its success.

In Chapter 5.4, we extend these findings to the more complex domain of human-AI collaboration. This chapter tackles our second research question (RQ2), applying

the same confidence-based principles to human & AI-DSS dyads and analysing how the AI's metacognitive profile influences team performance.

To overcome the limitations of static heuristics, Chapter 6.4 introduces an adaptive framework for joint inference. This chapter addresses our final research question (RQ3) by developing and evaluating a novel approach using contextual bandit algorithms. We demonstrate that this sequential method significantly outperforms the MCS heuristic and identify the specific contextual information required for its success.

Finally, Chapter 7 concludes the thesis by summarizing the key contributions, discussing their broader implications for the field of human-robot collaboration, and outlining promising directions for future research.

1.4 Publications

The work in this thesis has been published as the following:

- Duc-An Nguyen, Raunak Bhattacharyya, Nick Hawes. *Beyond Confidence: A Bandit-Based Framework for Joint Human-AI Inference in Robot Missions.* In: IEEE International Conference on Systems, Man, and Cybernetics. 2026.
- Duc-An Nguyen, Clara Colombatto, Steve Fleming, Ingmar Posner, Nick Hawes, Raunak Bhattacharyya. *Towards Joint Human-AI Inference for Robot Missions: A Confidence-Based Approach.* In arXiv preprint arXiv:2508.03293 (2025).
- Duc-An Nguyen, Raunak Bhattacharyya, Clara Colombatto, Steve Fleming, Ingmar Posner, Nick Hawes. *Is Confidence All You Need? Exploring Human-AI Joint Decision-Making in Spatiotemporal Tasks.* In: The Thirteenth International Conference on Learning Representations (ICLR) Workshop on Human-AI Coevolution. 2025.
- Duc-An Nguyen, Raunak Bhattacharyya, Clara Colombatto, Steve Fleming, Ingmar Posner, Nick Hawes. *Group Decision-Making in Robot Teleoperation: Two Heads are Better Than One.* In: The Proceedings of the Twentieth ACM/IEEE International Conference on Human-Robot Interaction (HRI). 2025.
- Raunak Bhattacharyya*, Duc-An Nguyen*, Clara Colombatto, Steve Fleming, Ingmar Posner, Nick Hawes. *Towards Intelligent Decision Support Systems in Robotics: Investigating the Role of Self-Confidence Calibration in Joint Decision-Making.* In: The Proceedings of the AAAI Spring Symposium. 2024.

*These authors contributed equally to the work.

Publications extending this thesis:

- Le Tuan Minh Trinh, Le Minh Vu Pham, Thi Minh Anh Pham, Duc-An Nguyen. *Metacognitive Sensitivity for Test-Time Dynamic Model Selection*. In: The Thirty-Ninth Neural Information Processing Systems Conference (NeurIPS). 2026.

- Thi Minh Anh Pham, Duc-An Nguyen, Cephas Svosve, Vasileios Argyriou, Georgios Tzimiropoulos. *PRE: Vision-Language Prompt Learning with Reparameterization Encoder*. In: The Twelfth International Conference on Learning Representations (ICLR). 2024.

2
Background

In this chapter, we present the theoretical background that underpins our analysis of joint inference. We begin with an overview of metacognition, focusing on two measurable constructs: metacognitive sensitivity (the ability to distinguish correct from incorrect decisions) and metacognitive calibration (the alignment between confidence and accuracy).

We then outline two frameworks for the joint inference: Maximum Confidence Slating (MCS), a heuristic for aggregating group decisions, and the Multi-Armed Bandit (MAB) framework, which models sequential decision-making under uncertainty. These concepts provide the foundation for the following chapters.

- **Chapter 4** applies metacognition and MCS to study dyads joint inference in robot teleoperation, as described in [19] (Nguyen et al.).

- **Chapter 5** extends this foundation by incorporating MAB to investigate joint human–AI inference, as presented in [20] (Nguyen et al.).

- **Chapter 6** explores more advanced Bandit algorithms, including contextual and latent variable bandits, to examine adaptive strategies for joint inference, as presented in [21] (Nguyen et al.).

Contents

2.1 Metacognition 14
2.1.1 Definition and Foundational Frameworks 14
2.1.2 Quantifying Metacognition: Sensitivity & Calibration . 18
2.2 Maximum Confidence Slating 23
2.3 Multi-Armed Bandit 24
2.4 Discussion 28

2.1 Metacognition

2.1.1 Definition and Foundational Frameworks

Metacognition is often defined as “thinking about thinking” or “cognition about cognition” [22]. This encapsulates a vast and complex field of study dedicated to understanding how sentient beings monitor, evaluate, and regulate their own mental processes. It is the capacity for introspection that allows an individual to move beyond merely executing a cognitive task to actively managing their own mental performance, shaping how effectively they learn, solve problems, and make decisions. It was not until the 1970s, however, that the concept was formalized within the domain of psychology. Developmental psychologist John Flavell is widely credited with coining the term “metacognition” and providing its first systematic definition. Moving beyond philosophical inquiry, Flavell framed metacognition as a scientifically tractable phenomenon, defining it as "the active monitoring and consequent regulation and orchestration of these cognitive processes in relation to the cognitive objects or data on which they bear, usually in service of some concrete goal or objective" [23]. Later, in 1990, the “Two-Level Architecture of Mind” model proposed by Nelson-Narens provides a powerful and testable framework by structuring the mind into two distinct but intimately connected hierarchical levels [24]. This structure moved the field beyond purely descriptive accounts by specifying the functional relationships and information flow that enable self-regulation.

The two levels of the model are:

- **The Object-Level:** This is the "ground floor" of cognition, where primary cognitive processes operate directly on the world or on internal representations. Examples of object-level processes include perceiving a visual stimulus, decoding text while reading, retrieving a memory, or performing a mathematical calculation. This is where "thinking" happens.

- **The Meta-Level:** This is the higher-order level where "thinking about thinking" takes place. The meta-level contains a dynamic, albeit imperfect, model of the object-level. Its function is to observe and exert executive control over the object-level processes to ensure a cognitive goal is achieved. [25]

The explanatory power of the Nelson-Narens model lies in its explicit characterization of the two-way flow of information between these levels, which it terms **Monitoring** and **Control**. Figure 2.1 represents the connection of Object-Level and Meta-Level and the two-way flow of information between these levels.

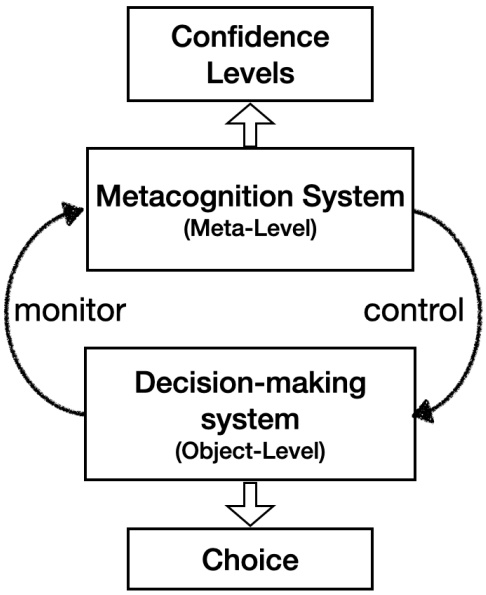


Figure 2.1: The "Choice" is determined by the decision-making system, and the decision uncertainty via "Confidence Levels" is encoded in the metacognition system.

- **Monitoring:** This describes the flow of information *from* the object-level *to* the meta-level. The meta-level actively monitors the operations occurring at the object-level, giving rise to subjective metacognitive experiences and judgments. For example, as a student reads a textbook (object-level process), their meta-level monitors their comprehension. This monitoring process generates a judgment, such as a feeling of understanding or a feeling of confusion. These judgments form the basis of metacognitive awareness.

- **Control:** This describes the flow of information *from* the meta-level *to* the object-level. Based on the information it has gathered via monitoring, the meta-level can initiate, continue, modify, or terminate object-level processes. If the meta-level's monitoring indicates that comprehension is poor, it can exert control by directing the object-level to take corrective action, such as re-reading the difficult paragraph, slowing down the reading pace, or looking up an unfamiliar term in a dictionary. These actions are metacognitive control.

In the context of the robot control delay joint inference mission we are considering, the two levels of the Nelson-Narens model (Object and Meta) and its two information flows (Monitoring and Control) can be defined as follows:

1. **Object-Level:** The human operator's decision process (e.g., the cognitive act of choosing "Robot A is faster").

2. **Meta-Level:** The confidence or uncertainty about that decision.

3. **Monitoring:** The meta-level monitors the fluency of the human operator's internal and active mental process of making a decision based on available evidence. If the evidence is overwhelming and points clearly in one direction (e.g., Robot A is significantly more sluggish than Robot B), the monitoring process returns a 'high fluency' signal, which the meta-level interprets as a high feeling of confidence. If the evidence is contradictory or sparse (e.g., the delay difference between the two robots is small and difficult to perceive, or results from different trials are inconsistent), the monitoring returns a 'low

fluency' or 'high conflict' signal, interpreted as uncertainty or a low feeling of confidence.

4. **Control:** Based on the information from monitoring (e.g., "I am very uncertain"), the meta-level must initiate, modify, or terminate the object-level process. What does it mean to "control" the decision-making process in this case? If confidence is high, the meta-level's control action is to terminate the deliberation and output the final choice: "Final decision: Robot A is faster." If confidence is low, the meta-level's control action is to not make a final decision. It identifies that the current cognitive process is insufficient. The control command would be something like: "Abort decision. Seek more information or suggestion from AI-DSS."

The progression of metacognitive theory from descriptive to the Nelson-Narens model's functional architecture represents a critical maturation of the field. This shift was not merely a change in terminology but a conceptual leap that made metacognition amenable to rigorous, quantitative, and neuroscientific investigation. The model's clear distinction between object-level performance and meta-level evaluation provided a theoretical blueprint for empirical work. It allowed neuroscientists to begin searching for the distinct neural hardware that implements the meta level and it provided the basis for quantitative psychologists to develop precise metrics to measure the fidelity of the "monitoring" signal. In this way, the Nelson-Narens model serves as the essential theoretical bridge connecting the psychology of self-reflection to the modern cognitive neuroscience of metacognition.

To bridge these cognitive models with robotic collaboration, this thesis adopts the concept of *mixed-initiative*. In this context, it is defined as a collaborative paradigm where the distribution of control and decision-making authority is not fixed but is negotiated dynamically. The "initiative" is shared between the human and AI based on their respective metacognitive signals; rather than the AI simply helping the human, the system arbitrates which partner should take the lead for a specific inference based on their observed or inferred competence.

2.1.2 Quantifying Metacognition: Sensitivity & Calibration

Building upon its foundational frameworks, metacognition research has expanded into a diverse and multifaceted field. This expansion has proceeded along two primary axes: a horizontal exploration of metacognition's role across a wide range of cognitive domains, and a vertical investigation into its biological implementation in the brain with the neuroscientific evidence that has begun to uncover the neural architecture of self-monitoring. We however, focus on quantifying metacognition in order to interpret and conduct empirical research in human-AI joint inference via sharing metacognition information.

To move metacognition from a qualitative concept to a quantitative science, researchers have developed sophisticated methods for measuring the accuracy and properties of self-monitoring. This endeavor has largely focused on two primary, complementary measurement axes: **metacognitive sensitivity**, which captures the fidelity of self-insight, and **metacognitive calibration**, which captures the alignment of subjective confidence with objective reality.

Axis I: Metacognitive Sensitivity

Metacognitive sensitivity is formally defined as an individuals ability to distinguish between their own correct and incorrect decisions [26]. It measures the degree of *correspondence* or association between a person's subjective confidence and their objective accuracy [27]. An individual with high metacognitive sensitivity will reliably assign higher confidence to their correct judgments and lower confidence to their incorrect ones.

The framework for measuring metacognitive sensitivity is derived from Signal Detection Theory (SDT). The application of SDT to metacognition involves conceptualizing the cognitive task and the metacognitive evaluation as two separate, nested decision problems [28].

- **Type 1 and Type 2 Tasks:** The primary task, such as the robot performing object identification (e.g., classifying an object as a 'cube' or a 'sphere'), is the

Type 1 task. The robot's performance on this task is quantified by the Type 1 sensitivity measure, d'. The subsequent metacognitive judgment, where the robot reports its confidence in that identification (e.g., outputting a confidence score from 0 to 1), is the Type 2 task. The goal of the Type 2 task is for the robot's internal model to discriminate between the states of having made a correct versus an incorrect Type 1 identification.

- **Type 2 ROC Curves:** To measure Type 2 sensitivity in a bias-free way, one can construct a Type 2 Receiver Operating Characteristic (ROC2) curve. This is analogous to a standard Type 1 ROC curve. For each possible confidence level (e.g., on a 1-4 Likert scale), one calculates the **Type 2 Hit Rate**, defined as the probability of giving a confidence rating ($\mathtt{k}$) at or above that level given a correct Type 1 response ($p(\mathtt{confidence} = \mathtt{k}|\mathtt{correct})$), and the **Type 2 False Alarm Rate**, defined as the probability of giving a confidence rating at or above that level given an incorrect Type 1 response ($p(\mathtt{confidence} = \mathtt{k}|\mathtt{incorrect})$). Plotting the Type 2 Hit Rate against the Type 2 False Alarm Rate for all confidence criteria yields the Type 2 ROC curve [29]. The calculation of `AUROC2` value is simplfied by the following equation:

$$\mathtt{AUROC2} = \sum_{k=1}^{n-1} (\mathtt{FAR}_{\mathtt{k+1}} - \mathtt{FAR}_{\mathtt{k}}) \cdot \frac{\mathtt{HR}_{\mathtt{k+1}} + \mathtt{HR}_{\mathtt{k}}}{2}, \tag{2.1}$$

where n is the number of confidence levels on the Likert scale. `HR` and `FAR` are the Type 2 Hit Rate and False Alarm Rate of the inference at each confidence level ($\mathtt{k}$). The Area Under this Curve (AUC) provides a non-parametric, bias-free measure of metacognitive sensitivity, which we called **AUROC2**.

The AUROC2 implementation algorithm is shown in Algorithm 1. An AUROC2 of 0.5 indicates chance-level sensitivity, while an AUC of 1.0 indicates perfect sensitivity.

Algorithm 1 Metacognitive Sensitivity (AUROC2)

Input:

`correct`: vector of size $1 \times n_{\text{trials}}$, (0: error and 1: correct)
`conf`: vector of size $1 \times n_{\text{trials}}$, confidence ratings $(1, \ldots, N_{\text{ratings}})$
N_{ratings}: number of available confidence levels

Initialize arrays `HR` and `FAR` of size N_{ratings}
Initialize $i \leftarrow N_{\text{ratings}} + 1$
for $k = 1$ **to** N_{ratings} **do**
    $\texttt{HR}[i-1] \leftarrow \text{count}(\texttt{conf} = k \wedge \texttt{correct}) + 0.5$
    $\texttt{FAR}[i-1] \leftarrow \text{count}(\texttt{conf} = k \wedge \neg\texttt{correct}) + 0.5$
    $i \leftarrow i - 1$
▷ *Normalize to create probability distributions*
$\texttt{HR} \leftarrow \texttt{HR} / \sum(\texttt{HR})$
$\texttt{FAR} \leftarrow \texttt{FAR} / \sum(\texttt{FAR})$
▷ *Compute cumulative sums*
$\texttt{csum_HR} \leftarrow [0, \text{cumsum}(\texttt{HR})]$
$\texttt{csum_FAR} \leftarrow [0, \text{cumsum}(\texttt{FAR})]$
Initialize $j \leftarrow 1$
for $k = 1$ **to** N_{ratings} **do**
    $\texttt{S}[j] \leftarrow (\texttt{csum_HR}[k+1] - \texttt{csum_FAR}[k])^2 - (\texttt{csum_HR}[k] - \texttt{csum_FAR}[k+1])^2$
    $j \leftarrow j + 1$
$\texttt{auroc2} \leftarrow 0.5 + 0.25 \times \sum(\texttt{S})$
return `auroc2`

Axis II: Metacognitive Calibration

While sensitivity measures how well confidence *discriminates* between correct and incorrect responses, calibration measures how well the *absolute level* of confidence maps onto the absolute probability of being correct. Metacognitive calibration is formally defined as the degree of fit between a person's subjective judgment of performance and their objective, actual performance [30, 31].

In the domain of experimental psychology, early and simple approaches to calibration often involve calculating a discrepancy score, which is the difference between a person's mean confidence rating and their mean accuracy across a set of tasks. A positive score indicates overall overconfidence, a negative score indicates underconfidence, and a score near zero suggests good calibration at the aggregate level.

In the domain of artificial intelligence, particularly in the context of probabilistic

classification models, the concept of calibration is of paramount importance. A model is considered well-calibrated if its predicted confidence scores can be directly interpreted as true probabilities of correctness. For instance, if a model assigns a confidence of 80% to a set of predictions, it should be accurate on approximately 80% of those predictions. This property is not merely a desirable academic feature; it is a critical requirement for the trustworthy and reliable deployment of AI systems in real-world, high-stakes environments.

Among the various metrics proposed to evaluate model calibration, the Expected Calibration Error (ECE) has become one of the most widely used due to its conceptual simplicity and intuitive interpretation. ECE directly measures the discrepancy between a model's average confidence and its average accuracy, providing a single scalar value that quantifies the overall miscalibration.

The core idea behind ECE is to assess whether a model's confidence aligns with its empirical success rate across different levels of confidence. The calculation is a direct operationalization of this idea. For a set of $\mathtt{n}$ predictions, the ECE is formally defined as the weighted average of the absolute difference between accuracy and confidence within a series of confidence bins :

$$\mathtt{ECE} = \sum_{\mathtt{m}=1}^{\mathtt{M}} \frac{|\mathtt{B_m}|}{\mathtt{n}} |\mathtt{acc}(\mathtt{B_m}) - \mathtt{conf}(\mathtt{B_m})| \tag{2.2}$$

The computation of ECE involves the following steps:

1. Binning: The confidence scores of all predictions are partitioned into M discrete bins. The confidence score for a given prediction is typically taken as the maximum value of the model's output probability vector (e.g., the max softmax probability). A common approach is "equal-width binning," where the confidence interval is divided into $\mathtt{M}$ equally sized subintervals (e.g., for $\mathtt{M} = 10$, the bins would be $[0, 0.1], (0.1, 0.2], ..., (0.9, 1.0]$). Each prediction is assigned to the bin $\mathtt{B_m}$ that corresponds to its confidence score.

2. Per-Bin Accuracy ($\texttt{acc}(\texttt{B}_\texttt{m})$): For each bin $\texttt{B}_\texttt{m}$, the accuracy is calculated as the fraction of predictions within that bin that are correct. This is formally expressed as:

$$\texttt{acc}(\texttt{B}_\texttt{m}) = \frac{1}{|\texttt{B}_\texttt{m}|} \sum_{\texttt{i} \in \texttt{B}_\texttt{m}} \mathbb{I}(\hat{\texttt{y}}_\texttt{i} = \texttt{y}_\texttt{i})$$

where $|\texttt{B}_\texttt{m}|$ is the number of predictions in bin $\texttt{m}$, $\hat{\texttt{y}}_\texttt{i}$ is the predicted label for sample $\texttt{i}$, $\texttt{y}_\texttt{i}$ is the true label, and $\mathbb{I}(\cdot)$ is the indicator function.

3. Per-Bin Confidence ($\texttt{conf}(\texttt{B}_\texttt{m})$): For each bin $\texttt{B}_\texttt{m}$, the average confidence is calculated by averaging the confidence scores of all predictions within that bin. This is formally expressed as:

$$\texttt{conf}(\texttt{B}_\texttt{m}) = \frac{1}{|\texttt{B}_\texttt{m}|}(\hat{\texttt{p}}_\texttt{i})$$

where $\hat{\texttt{p}}_\texttt{i}$ is the confidence (maximum probability) for sample $\texttt{i}$.

4. Weighted Average of Differences: The absolute difference between the accuracy and confidence for each bin, $\texttt{acc}(\texttt{B}_\texttt{m}) - \texttt{conf}(\texttt{B}_\texttt{m})$ represents the miscalibration within that specific confidence range. The ECE is then the weighted average of these differences, where each bin's contribution is weighted by the proportion of samples that fall into it, $\frac{|\texttt{B}_\texttt{m}|}{\texttt{n}}$.

The ECE implementation algorithm is shown in Algorithm 2.

A perfectly calibrated model would have an ECE of 0, as its accuracy would match its confidence in every bin. A higher ECE value indicates greater miscalibration.

It is crucial to recognize that sensitivity and calibration are complementary and capture fundamentally different aspects of self-evaluation. A person could be perfectly sensitive but poorly calibrated. For example, they might consistently rate their confidence as "4" on a 4-point scale when they are correct and "2" when they are incorrect. This shows perfect discrimination (high sensitivity), but if their overall accuracy is only 50%, they are severely overconfident (poor calibration). Conversely, a person could be perfectly calibrated on average but have zero sensitivity. If their

Algorithm 2 Expected Calibration Error (ECE)

Input:
Model $f : \mathcal{X} \to [0,1]^C$ (predicts probabilities for C classes)
Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ where $x_i \in \mathcal{X}$ is an input and $y_i \in \{1, \ldots, C\}$ is the true label
Number of bins M

Initialize bins $B_m = \emptyset$ for $m = 1, \ldots, M$
Initialize ECE $= 0$
for $i = 1$ **to** N **do**
    Predict probabilities $p_i = f(x_i)$
    Get the predicted class $\hat{y}_i = \arg\max_{c \in \{1,\ldots,C\}} p_{i,c}$
    Determine the confidence score $\text{conf}_i = p_{i,\hat{y}_i}$
    Assign sample i to bin B_m where $\frac{m-1}{M} \leq \text{conf}_i < \frac{m}{M}$
for $m = 1$ **to** M **do**
    if $B_m \neq \emptyset$ **then**
        $|B_m|$ = number of samples in bin B_m
        $\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbb{I}(\hat{y}_i = y_i)$
        $\text{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \text{conf}_i$
        $\text{ECE} += \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)|$
return ECE

overall accuracy is 75%, they could simply state "75% confident" on every single trial. On average, their confidence would match their accuracy perfectly (good calibration), but their ratings would provide no information about their trial-by-trial performance (zero sensitivity). Therefore, a complete assessment of metacognitive monitoring requires measuring both *sensitivity* and *calibration*.

2.2 Maximum Confidence Slating

Among the simplest and most intuitive methods for aggregating judgments is Maximum Confidence Slating (MCS). Born from research into the "two heads are better than one" effect, MCS proposes a simple heuristic rule: in a disagreement, trust the person who is more confident. This heuristic approach bypasses the need for complex interaction or mathematical weighting, offering a computationally robust method for leveraging confidence.

The core principle of Maximum Confidence Slating is that for a given binary decision faced by a group, the final decision is determined solely by the choice of the

single most confident member of that group [32]. It is a "winner-takes-all" heuristic that effectively slates the most confident individual's decision as the group's decision.

We adopt this idea to explore how groups make better decisions together. First, we look at Human–Human settings, and then extend the analysis to Human–AI teams. In both cases, we examine how Maximum Confidence Slating works with each member's ability to judge their own confidence (measured using the metacognitive sensitivity metrics introduced in Section 2.1.2).

2.3 Multi-Armed Bandit

Another method of combining inference and later, decision in a systematic way is via a sequential expected utility maximising approach such as Multi-Armed Bandit (MAB). In this section we introduce the background that includes the overall neccesary introduction for the MAB. This section provides an overview of the multi-armed bandit algorithm and the Thompson Sampling methods implemented in Chapter 5. In Chapter 6, we extend this into more advanced multi-armed bandit algorithms, which have been developed to address metacognitive context and latent variable inference.

A bandit problem is a sequential game between a learner and an environment. The game is played over n rounds, where n is a positive natural number called the horizon. In each round $t \in [n]$, the learner first chooses an action A_t from a given set $\mathcal{A}$, and the environment then reveals a reward $X_t \in \mathbb{R}$. The learner cannot peek into the future when choosing their actions, which means that A_t should only depend on the history $H_{t-1} = (A_1, X_1, \ldots, A_{t-1}, X_{t-1})$. A *policy* is a mapping from histories to actions: a learner adopts a policy to interact with an environment. An *environment* is a mapping from history sequences ending in actions to rewards. The most common objective of the learner is to choose actions that lead to the largest possible cumulative reward over all T rounds, which is $\sum_{t=1}^{T} R_t$.

In the context of the joint Human-AI inference task, the problem of deciding which agent's judgment to trust on any given trial is similarly formalized as a sequential decision-making challenge, specifically, a Multi-Armed Bandit (MAB)

problem. The objective of the meta decision-maker (the arbitration policy) is to select the agent that is most likely to yield a correct team decision.

The meta decision-maker cannot know the true accuracy of the human or the AI beforehand, meaning the action A_t must only depend on the history of observed contexts, actions, and rewards, $H_{t-1} = (X_1, A_1, R_1, \ldots, X_{t-1}, A_{t-1}, R_{t-1})$. A *policy* is a mapping from histories to actions.

The problem is defined by the following specialized components:

- **Arms (A):** A set of $K = 2$ discrete actions, or arms, denoted by $A = \{A_H, A_{AI}\}$ in the case of human-AI joint inference; and similarly $A = \{A_{H_1}, A_{H_2}\}$ in the case of human-human joint inference. These arms correspond to:
  - A_H: Select Human's Inference (I_H)
  - A_{AI}: Select AI's Inference (I_{AI})
- **Time Horizon (T):** A sequence of T discrete time steps, or trials, indexed by $t = 1, 2, \ldots, T$.
- **Action Selection (a_t):** At each time step t, the arbitration policy selects one arm $a_t \in A$ to play. This constitutes the team's final decision, $I_{\text{Team}} = I_{a_t}$.
- **Reward (R_t):** Upon selecting an agent's inference I_{a_t}, the agent receives a binary reward $R_t \in \{0, 1\}$.
  - $R_t = 1$ (Success) if the selected inference I_{a_t} is correct.
  - $R_t = 0$ (Failure) if the selected inference I_{a_t} is incorrect.

  For the stochastic bandit, the true underlying mean reward (μ_i) for each arm $i \in \{H, AI\}$ is the agent's true, unknown probability of being correct. The rewards are assumed to be independent and identically distributed (i.i.d.) for each arm across time steps.

- **Objective:** The policy's goal is to maximize the expected cumulative correct decisions over the time horizon T. The objective function remains:

$$\text{Maximize } \mathbb{E}\left[\sum_{t=1}^{T} R_t\right]$$

where the expectation is taken over the randomness in rewards and any randomness in the policy.

The subsequent subsection outlines methods for learning optimal policies, which we will use to identify the best sequential strategy for adaptive joint inference. To implement this, we rely on Thompson Sampling - a Bayesian approach, also known as posterior sampling - to guide the arbitration decision. Its key idea is to select the agent (arm) based on the posterior probability that it is the most accurate, given the data observed so far.

The mechanism provides a natural solution to the exploration-exploitation dilemma in arbitration:

1. **Sample from the Posterior:** For each agent $i \in \{H, AI\}$, draw a random sample of its unknown true accuracy, θ_i, from its current posterior distribution.

2. **Act Greedily on the Sample:** Select the agent (arm) with the highest sampled probability of being accurate.

The posterior distribution for each agent's accuracy reflects the arbitration policy's (meta decision-maker) current belief. An agent that has demonstrated high accuracy consistently will have a posterior concentrated around a high value, leading the policy to exploit that agent frequently. Conversely, an agent with high uncertainty (perhaps due to being played infrequently) will have a wide posterior distribution, giving it a non-trivial probability of producing a high-valued sample, which encourages exploration. As more trials are conducted, the uncertainty decreases, and the policy naturally converges toward the truly optimal agent.

The Bayesian framework requires specifying a prior distribution $p(\theta)$ for the unknown accuracy parameters and a likelihood function $p(\theta|D)$ describing the probability of observing the history of correct/incorrect decisions given the parameters.

Using Bayes' rule, the prior is updated to a posterior distribution $p(\theta|D)$ after observing data: $p(\theta|D) \propto p(D|\theta)p(\theta)$. This posterior then serves as the prior for the next round of observations, allowing for iterative learning.

The reward of making correct, incorrect inference in our setting is binary ($R_t \in \{0, 1\}$). A canonical application of Thompson Sampling for this binary reward setting is the Beta-Bernoulli Model. The Beta distribution is used as the conjugate prior for the Bernoulli likelihood.

The Beta distribution, Beta(α, β), is defined on $(0, 1)$, and its PDF is:

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} x^{\alpha-1}(1 - x)^{\beta-1}$$

In this context, the parameters α and β have a clear physical meaning for each agent:

- α is intuitively interpreted as the total number of times the agent's inference was correct (successes) + 1.
- β is intuitively interpreted as the total number of times the agent's inference was incorrect (failures) + 1.

The complete Thompson Sampling algorithm using this model is summarized below, adapted to use S_i (Successes) and F_i (Failures) for $i \in \{H, AI\}$.

Algorithm 3 Thompson Sampling (Beta-Bernoulli) for Joint Inference

Input: Agent Set $A = \{A_H, A_{AI}\}$, Time horizon T

for $i \in \{H, AI\}$ **do**
    $S_i \leftarrow 0$
    $F_i \leftarrow 0$
for $t = 1$ **to** T **do**
    for $i \in \{H, AI\}$ **do**
        Sample $\theta_i(t) \sim \text{Beta}(S_i + 1, F_i + 1)$
    $a_t \leftarrow \arg\max_{i \in \{H,AI\}} \theta_i(t)$
    $I_{\text{Team}} \leftarrow I_{a_t}$ ▷ *The team decision is the selected agent's inference*
    Observe binary reward $R_t \in \{0, 1\}$ from the chosen agent a_t's inference
    if $R_t = 1$ **then**
        $S_{a_t} \leftarrow S_{a_t} + 1$
    else
        $F_{a_t} \leftarrow F_{a_t} + 1$

2.4 Discussion

This chapter introduced the theoretical constructs and algorithmic frameworks that ground our investigation of joint inference. Metacognitive sensitivity and calibration offer a way to quantify the reliability of confidence as a signal in collaborative settings. MCS provides a principled heuristic for combining such signals, while Bandit algorithms capture the dynamics of sequential decision-making and adaptation.

These concepts establish the basis for the analyses in the following chapters:

- **Chapter 4** applies metacognition (Section. 2.1.1) and Maximum Confidence Slating heuristic (Section. 2.2) to group decision-making in robot teleoperation.
- **Chapter 5** incorporates Maximum Confidence Slating heuristic (Section. 2.2) and multi-arm bandits (Section. 2.3) to study joint human–AI inference in robot teleoperation.
- **Chapter 6** extends to contextual (Section 2.3) and latent variable bandits to explore adaptive joint human-AI inference strategies.

The description of each chapter's methodological focus, contribution to joint inference research, and preliminary background knowledge connection is summarised in Table 2.1.

Table 2.1: Description of background knowledge connection to each chapter's methodological focus.

Dimension	**Chapter 4**	**Chapter 5**	**Chapter 6**
Collaboration Type	Human-Human dyads	Human-AI dyads	Human-AI dyads
Heuristic?	Yes	Yes	No
Adaptive?	No	Partially	Yes
Bandit Approach?	None	Thompson Sampling (MAB)	Contextual & Latent Bandit
Focus of Study	Confidence calibration & heuristics in joint inference	AI-assisted decision making under uncertainty	Unified framework with context awareness
Background	Metacognitive sensitivity (see Section. 2.1.2), Maximum Confidence Slating (see Section. 2.2)	Maximum Confidence Slating (see Section. 2.2, Multi-Armed Bandit (see Section. 2.3)	Metacognition (see Section. 2.1.1), Maximum Confidence Slating (see Section. 2.2), Multi-armed Bandit (see Section. 2.3)

3
Literature Review

In this chapter we review the literature on joint decision-making and inference. We begin with methods in human and human collaboration that rely on confidence sharing. We then explore human and agent collaboration, highlighting a research gap concerning the use of metacognition sharing to achieve complementarity. Finally, we review the literature on engineering metacognition in AI systems, and provide existing experimental frameworks to measure metacognition. This review highlights metacognition as a key signal that supports joint inference in human collaboration and notes that its limited integration in AI and robotic partners represents a promising avenue for advancing human–agent complementarity.

Contents

3.1 Confidence Sharing in Human Collaboration 30
3.1.1 Confidence Matching 30
3.1.2 Weighted Confidence Sharing 31
3.1.3 Maximum Confidence Slating 33
3.2 Human-Agent Collaboration 34
3.2.1 Human-AI Joint Inference 34
3.2.2 Human-Robot Joint Inference 37
3.3 Engineering Metacognition in AI Systems 42
3.4 Metacognition Experimental Paradigms 44
3.5 Discussion 46

3.1 Confidence Sharing in Human Collaboration

A central rationale for studying human–human collaboration is the "two-heads-better-than-one" (2HBT1) effect, in which the joint decision of a dyad exceeds the accuracy of its individual members. [33–37]. This principle is actively leveraged in high-stakes domains where errors have significant consequences. For example, medical tumor boards rely on the joint inference of radiologists, oncologists, and pathologists to form a diagnosis [38, 39], while aircraft cockpit crews must constantly share information to make unified decisions about flight operations, especially in emergencies [40, 41].

Central to this collective intelligence is the communication of metacognitive states, primarily the expression of confidence. Confidence - an individual's subjective estimate of the probability that their decision is correct - acts as the metacognitive signal exchanged to negotiate a shared reality. To understand this exchange, researchers have developed several models and heuristics that describe how partners integrate, weigh, or select between differing opinions based on these signals of certainty.

3.1.1 Confidence Matching

Early methods for combining human-human decisions were introduced by Bang et al. in a series of behavioral experiments using a visual stimuli two-interval forced-choice contrast discrimination task. When partners interact directly, they often employ a heuristic called "Confidence Matching" [42–44]. In this strategy, group members, through discussion, implicitly negotiate and align their communicated confidence levels to create a shared, common metric for expressing certainty.

While this strategy is effective when partners have comparable expertise, it becomes less so when an *expertise gap* exists. If confidence alignment is treated as perfectly symmetrical and driven solely by consensus, the group may fail to account for differences in competence once confidence values have been artificially matched. As a result, the more competent partner's opinion may be underweighted and the less competent partner's overweighted, leading to a degradation in group performance [45, 46]. This illustrates a key challenge: calibrating trust in a partner's

metacognitive signals is difficult when the mapping between confidence and accuracy (metacognitive sensitivity; see Section 2.1.2) is unknown [22, 27].

3.1.2 Weighted Confidence Sharing

An extension of the confidence-matching approach is the Weighted Confidence Sharing (WCS) model proposed by Bahrami et al. The WCS model posits that for an optimal joint decision, individuals in a dyad must communicate their confidence on a trial-by-trial basis. The group's final decision is then determined by *weighting* each individual's choice by their respective confidence level. This model captures the intuition that the opinion of a more confident individual should carry more weight in the final decision.

The key difference between confidence matching and WCS lies in their underlying mechanisms. Confidence matching is a descriptive heuristic based on the alignment of reported confidence, whereas WCS is a normative model that prescribes an explicit rule for combining individual inputs through weighted aggregation. In doing so, WCS shifts the focus from negotiating confidence as a shared metric to mathematically integrating signals to produce an optimal joint outcome.

While the model predicts improved collective performance under many circumstances, it also highlights important limitations. When members' competence levels are highly disparate, dyadic performance can fall below that of the more skilled individual acting alone. This highlights a crucial nuance: simply sharing confidence is not a universal solution for improving group performance. When a *large disparity* in competence exists between members, specifically, when the metacognitive sensitivity of the less able member is less than approximately 40% of the more able member's metacognitive sensitivity, the outcome of collaboration becomes suboptimal [48, 49].

Further research has refined our understanding of why performance heterogeneity is so damaging to team's joint decisions. The work of Massoni and Roux suggests that the core issue is not the difference in skill levels per se, but rather the group's failure to accurately represent this difference in their shared confidence signals. This is a problem of metacognitive calibration (described in Section 2.1.2) [51] and

metacognitive sensitivity (described in Section 2.1.2) [32] . Mahmoodi et al. also identified another perspective of social-cognitive heuristic that helps explain why groups fail to account for competence gaps: the equality bias [52]. This is the pervasive tendency for individuals in a group to weight their partner's opinion nearly equally to their own, effectively behaving as if their partner is just as competent (or incompetent) as they are, regardless of the objective reality.

Evidence for these mechanisms comes from experimental studies using collective perceptual decision-making tasks. In a typical paradigm, dyads engaged in a two-interval forced-choice visual search task. On each trial, both participants privately indicated in which interval they believed the target appeared and rated their confidence. Their individual responses were then exchanged, and, in cases of disagreement, one member (randomly assigned as arbitrator) made the joint decision on behalf of the pair. Feedback was provided after each trial, allowing participants to monitor performance over time.

This paradigm has also been modeled using reinforcement learning (RL) approaches. RL agents were designed to learn optimal policies for combining inputs based on confidence, updating their estimates of each partner's reliability through trial-by-trial prediction errors. The decision criterion took the form:

$$\mathrm{DC} = p_{s,i} \times c_{s,i} + \gamma \times p_{o,i} \times c_{o,i}$$

where $p_{s,i}$ and $p_{o,i}$ are the predicted probabilities of correctness for self and other. $c_{s,i}$ and $c_{o,i}$ are their respective confidence ratings. The parameter γ is a free parameter capturing the weight assigned to the partner's input (i.e., social influence).

By comparing the parameter value that maximized the fit between the model-derived decisions and the observed decisions (γ_{fit}) against the parameter value that maximized the accuracy of the decisions (γ_{opt}), researchers were able to quantify the extent of the equality bias shown by the human participants.

In Experiment 1, participants were separated by a screen and not allowed to communicate verbally. This constraint ensured that the only information exchanged between dyad members during a disagreement was their private decision (which

interval contained the target) and their expressed level of confidence. In Experiment 2, the same dyads were permitted to communicate verbally. Reinforcement learning (RL) agents were trained to model human behavior under both conditions. The agents successfully learned to collaborate effectively when communication was restricted to a non-verbal interface (Experiment 1). However, when verbal interaction was permitted (Experiment 2), the same models failed to achieve any collective benefit. The escalating, uncalibrated confidence signals in the verbal condition provided a noisy and unreliable training signal, preventing the agent from learning the true relationship between confidence and accuracy [53].

3.1.3 Maximum Confidence Slating

Research on dyadic joint inference has shown that teams decision can outperform individuals when guided by the Maximum Confidence Slating (MCS) approach [49, 53–56]. The Maximum Confidence Slating (MCS) heuristic (Section. 2.2) selects the decision of the participant with the higher confidence. The power of this approach was highlighted by the contribution of Koriat, who demonstrated that this collaborative benefit could be achieved without any direct interaction at all. By forming "virtual dyads" from the responses of non-interacting participants and applying the MCS rule, the resulting aggregated performance was superior to that of the individuals [57]. Later Bahrami et al. and Bang et al. confirmed that for individuals with nearly equal reliability, decisions advised by the confidence heuristic were just as accurate as those reached through interaction [25, 58, 59]. These finding were significant because it suggested that much of the benefit of group decision-making could be distilled into a simple, non-interactive algorithm. In this context, "non-interactive" refers to an experimental and analytical setting where participants perform tasks in isolation without any real-time communication or awareness of a partner. The "dyadic" benefit is explored through virtual dyads, where the responses of two independent individuals are aggregated post-hoc using the Maximum Confidence Slating (MCS) rule. This allows us to isolate the mathematical

efficacy of confidence-sharing from the social biases and "equality bias" often found in live verbal discussions.

Summary

In essence, the progression from Confidence Matching to WCS and MCS reveals a core tension between social integration and efficient heuristics. While collaborative strategies like Confidence Matching and WCS aim to synthesize opinions, they are fragile and susceptible to social biases and skill imbalances that can degrade group performance. The Maximum Confidence Slating heuristic offers a radically different, non-interactive solution that leverages the power of individual metacognitive signals in its simplest form: trusting the most confident agent. The key insight from this literature is that the apparent benefit of having "two heads" is not guaranteed; it is highly dependent on the mechanism used to combine their inputs. This landscape of effective but flawed human strategies underscores the profound challenge of building an AI partner that can navigate these same complexities.

3.2 Human-Agent Collaboration

The challenge of combining human and machine intelligence to make better inferences is not confined to a single field but is a cross-cutting theme in the development of applied AI and robotics. In this section, we review the methods by which joint inference is implemented across different domains, including AI, robotics, and cognitive science. This broad survey demonstrates the widespread relevance of the problem and provides a richer context for our deep dive into the robotics domain.

3.2.1 Human-AI Joint Inference

The most prevalent model for human-AI joint inference is the AI-assisted paradigm, where an AI Decision Support System (AI-DSS) provides recommendations or analyses, but a human operator makes the final inference for the team. This model is widely deployed in high-stakes fields like medicine and finance [60–62]. AI systems, particularly those based on deep learning, have demonstrated remarkable capabilities

in analysing medical images (e.g., radiographs, pathology slides) to detect signs of disease, such as cancer [63]. The dominant paradigm is AI-assisted, where the AI provides a preliminary analysis or highlights regions of interest, but the human clinician makes the final diagnosis. Recent research highlights the potential for more synergistic models [64, 65]. The study involving over 2,100 clinical vignettes found that hybrid diagnostic collectives - groups comprising both human experts and AI models (including large language models) - were significantly more accurate than collectives of only humans or only AIs. The financial industry has also been an early adopter of human-AI collaboration for tasks like fraud detection, credit scoring, and risk management. Research by Nagarakanti shows that a collaborative approach has yielded significant, quantifiable benefits, with financial institutions reporting improvements in operational efficiency, decision-making time, and the accuracy of fraud detection [67, 68]. In intelligence and security, AI systems are increasingly used to support this process by automating the analysis of unstructured data, such as satellite imagery, video feeds, and text documents. However, for an analyst to trust an AI's output, they must understand its capabilities, limitations, and potential biases [69]. An AI model trained on biased data will produce biased outputs, which, if accepted uncritically, can lead to flawed intelligence assessments. A key area of research in this domain is on communicating AI uncertainty and confidence in a way that allows human analysts to appropriately calibrate their trust and integrate the AI's output into their broader analytical framework [70].

Overall, the collaboration effectiveness is often undermined by a fundamental challenge: humans exhibit both over-reliance (uncritically accepting incorrect AI advice) and under-reliance (dismissing correct AI advice). This suggests that the primary bottleneck in human-AI teams is not the capability of the individual agents, but the quality of the "cognitive interface" between them [71–73].

Recent research by Steyvers and Peters directly investigated this interface failure. The researchers aimed to quantify the miscommunication of uncertainty between Large Language Models (LLMs) and human users, defining this miscommunication as a "calibration gap" (the difference between the AI's internal confidence and

the user's perception of it) and a "discrimination gap" (the difference in their abilities to distinguish correct from incorrect answers) [75–78]. They conducted behavioral experiments where participants read an LLM's answer and its default text explanation, then estimated the probability that the AI's answer was correct. These human confidence ratings were then compared against the LLM's actual accuracy and its internal confidence, which was calculated from the model's token likelihoods. The study found a significant calibration gap. Relying on the AI's default explanations, users consistently overestimated the LLM's accuracy, confirming a strong tendency toward over-reliance. The results also revealed a powerful "length bias"; longer AI explanations increased user confidence even when the additional content did not improve the answer's accuracy. This demonstrates that users are influenced by shallow textual cues and that default explanations are a flawed tool for calibrating trust [31, 79–81].

These findings empirically support earlier work by Bansal et al., which first highlighted the counter-intuitive failure of AI explanations. That work found that providing users with state-of-the-art explanations did not significantly improve the complementary performance of human-AI teams when compared to simpler strategies. Most notably, the explanations were found to increase the likelihood that a human would accept an AI's recommendation regardless of its correctness, thereby actively reinforcing over-reliance. They concluded that simply showing the AI's confidence score was a more effective strategy than providing a complex explanation [82–87].

Summary

These studies converge on a critical point: the bottleneck at the cognitive interface is not a lack of information from the AI, but a failure to communicate its internal state of uncertainty effectively. This shifts the focus of the problem away from explanation design and toward the core cognitive ability that governs judgments of certainty - metacognition.

The limitations of the fixed-role, AI-assisted paradigm naturally lead to the exploration of more integrated peer-to-peer, or mix-initiative models, where the human and AI act as equal partners. This is the pinnacle of our research that enable toward true joint inference where both human and AI-DSS independently assess the evidence and produce an inference. The final team decision can then be based on either of the individual inferences, creating a symbiotic relationship. This model directly confronts the shortcomings of the AI-assisted paradigm by removing the default assumption of human superiority in final judgment. However, it introduces a new and critical challenge: the arbitration problem. When the human and AI disagree, whose inference should the team adopt? The challenge of deciding whose judgment to trust in a disagreement - the "arbitration problem" does not simply depend on who is better calibrated, but also on who possesses higher metacognitive sensitivity. The development of a principled arbitration mechanism is the key to unlocking the potential of peer-to-peer collaboration. Without such a mechanism, the team may devolve into conflict or suboptimal compromises. Therefore, a bidirectional, dynamic arbitration mechanism is the logical next step, which is what peer-to-peer joint inference aims to achieve.

3.2.2 Human-Robot Joint Inference

While the principles of joint inference are universal, their application in the domain of robotics introduces a unique set of challenges rooted in embodiment. In this section, the “robot” is an embodied AI agent with perception, reasoning, and actuation in the physical world. The inference is grounded in the robot’s sensory inputs (camera, lidar, tactile sensors) plus the human’s contextual or task knowledge. So the human is interacting with physical objects either in simulation or in real world and making joint inference with the embodied-AI robots.

When the AI agent is a robot operating in the physical world, the process of inference is no longer a purely abstract, cognitive task based on static data. Instead, it becomes a dynamic, spatiotemporal problem that is inextricably linked

to perception, action, and the physical laws of the environment. This embodiment fundamentally changes the nature of the inference problem, demanding new approaches that can handle the complexities of real-world interaction [3, 88–93].

Inference tasks in non-robotic domains, such as classifying an image or detecting fake news, are often passive and atemporal. The agent is presented with a fixed set of evidence and must produce a conclusion. In contrast, human-robot inference tasks are characterized by three key distinguishing factors:

- Inference of Latent Variables: Many critical properties of the robot and its environment are not directly observable and must be inferred.
- Temporal Processes: The evidence for inference is not static but unfolds over time, often in the form of continuous video streams or sensor data.
- Inference Based on Action: To gather the necessary information to make an inference, the human operator often must actively interact with or control the robot, turning inference into an active, exploratory process.

The challenge of embodiment is most acutely felt in the pervasive problem of latent, or unobservable, states. An example of a critical latent system variable is control delay. In the task of inferring which of two teleoperated robots has a lower control delay, the human operator cannot simply look at a static image. They must actively command the robots to move and observe their responses over time through a video feed to perceive the subtle differences in maneuverability. This tight coupling of perception, action, and reasoning grounds the inference problem in the physical world and makes it significantly more complex than its disembodied counterparts [94–98]. As detailed by Andersen et al., delays in a robot's control loop are inherent and difficult to measure directly [16]. These delays include both actuation delay (the time from command to execution) and response delay (the time from execution to sensor feedback).

Simultaneously, for a robot to be an effective collaborator, it must infer the latent cognitive and emotional states of its human partner. These include the human's

intent (what they are trying to achieve), their preferences (how they want to achieve it), their adaptability (their willingness to change strategy), and their trust in the robot [99, 100]. These variables are not directly accessible to the robot's sensors and must be inferred from the human's actions, language, and even physiological cues. Misinterpreting these latent human states can lead to breakdowns in coordination, frustration, and a loss of trust, ultimately undermining team performance [101]. Given the uncertainty inherent in inferring latent states, probabilistic methods have become a cornerstone. The Partially Observable Markov Decision Process (POMDP) has emerged as a particularly powerful and principled framework for robot decision-making under uncertainty about a human's hidden state. A POMDP allows a robot to maintain a belief - a probability distribution - over the possible latent states of its human partner and to update this belief based on observations of the human's actions. A key example of this approach is the "trust-POMDP" developed by Chen et al. In their research, the Trust-POMDP model includes a trust dynamics model (how human trust evolves based on robot performance) and a human decision model (how human actions depend on trust). These models are learned from data collected in human subject experiments on a collaborative table-clearing task, both in simulation (201 participants) and with a real robot (20 participants). The robot then uses the Trust-POMDP to infer human trust and choose actions that maximize long-term team performance, which may involve exploiting current trust or exploring to build/reduce trust for future benefits. The results showed that robots using the Trust-POMDP could make effective decisions, such as starting with low-risk objects to build trust before handling high-risk ones, thereby minimizing interventions and maximizing team performance. Simulations also demonstrated that maximizing trust alone does not always lead to the best team performance, as strategic trust modulation (including intentional failures) can be necessary for optimal long-term outcomes in scenarios with potential robot failures [103].

Costen et al. propose modeling share autonomy (SA) systems as Mixed Observability Markov Decision Processes (MOMDPs). In this MOMDP framework, the state space is explicitly divided into observable environment states and hidden human

operator states. The SA system then maintains a belief distribution over these unobservable human states, updating it based on observed actions and outcomes, and uses this refined belief to make informed decisions about which operator (human or AI) should control the system. They illustrated their approach with a simple example where an agent learns about an unknown human operator (novice or experienced) through interaction. For empirical evaluation, they applied their MOMDP model to two scenarios: a surveillance task involving a UAV with uncertain human fatigue and the game Angry Birds, where the identity of an unknown AI player (random or greedy) was hidden. The results empirically showed that modeling SA systems as MOMDPs yielded "better performance" compared to using simplistic human models based on MDPs. It moves beyond simply acknowledging uncertainty to actively modeling and learning about human performance [104–106]. It is also noted that, to the best of our knowledge, these studies did not evaluate their methods with real teams in live settings.

On a broader level, the concept of a bidirectional understanding within a human-robot team is formalized in cognitive science as a Shared Mental Model (SMM), which represents a shared understanding of the task, environment, and each other's roles and intentions. [107]. A Shared Mental Model is a distributed knowledge structure that allows teammates to maintain awareness of each other's states and anticipate each other's actions and needs, facilitating fluent coordination with minimal explicit communication. The development of computational frameworks to implement SMMs in robots is an active area of research, with a central challenge being the prevention of mental model mismatch, which occurs when the human's internal model of the robot diverges from the robot's actual state [108, 109]. Scheutz, DeLoach, and Adams propose a comprehensive computational framework for developing and using SMMs in human-agent teams. They hypothesize that providing artificial agents with SMM capabilities will improve team efficiency and productivity. Their framework specifies the necessary data structures (e.g., representations of agent capabilities, goals, plans, and workload) and the computational processes required to build, maintain, and utilize these models in real-time. Preliminary implementations of the

SMM framework on robotic platforms have demonstrated significant improvements in team performance, enabling proactive actions, reduced human workload, and efficient coordination, even when communication is limited. For instance, a robot using an SMM can infer a human's cognitive workload and adjust its behavior to prevent overload [110–116].

Summary

In the existing literature, a pronounced asymmetry - where the robot is a modeler of the human while the human is left to decipher an opaque partner - reveals a deeper challenge for achieving true peer-to-peer collaboration. Current models allow a robot to predict a human's actions based on inferred states like trust or fatigue, but they stop short of capturing the human's own metacognitive state, such as the human's confidence in their assessment of the robot or their degree of certainty in a joint decision. For collaboration to improve, interaction mechanisms should move beyond one-way behavioural inference and incorporate mutual exchange of confidence and uncertainty.

The critical research gap lies in the lack of formal metacognitive principles applied to the design and evaluation of the AI agent itself also the understanding of human's metacognitive model while collaborating with robots. A significant void remains in understanding how humans and embodied AI models can work as peers and how such metacognitive signals can be utilized to facilitate dynamic joint inference. This gap is particularly acute for the dynamic, embodied tasks characteristic of robotics. In these settings, where inference is based on temporal processes and active information gathering, a reliable, moment-to-moment signal of the AI's certainty is crucial for a human partner to make effective decisions about when to rely on the AI's judgment.

To our knowledge, a significant void remains in understanding how humans and embodied AI models can work as peers and how such metacognitive signals can be utilized to facilitate dynamic joint inference. Our research begins to address this gap by examining how variations in an AI's metacognitive properties - specifically,

its sensitivity and calibration - shape the performance of human-AI teams in joint inference tasks.

3.3 Engineering Metacognition in AI Systems

In this section we review methods for engineering functional-not conscious-metacognition into AI systems. While the field has recently begun to assess the metacognitive properties of AI models [65, 117–119], it has done so primarily for diagnostic purposes: to evaluate a model's reliability after training. The literature reviewed here marks a conceptual shift, treating metacognition not as a passive property to be diagnosed, but as a functional component to be actively operationalized. This reframes the engineering goal, moving from measuring a model's capabilities post-hoc to building a dynamic signal of uncertainty that can be leveraged in real-time joint inference. The primary toolkit for this challenge comes from the field of Uncertainty Quantification (UQ) in deep learning, which provides principled methods for generating such data or signals [120–124].

Metacognitive Calibration

Metacognitive Calibration refers to the correspondence between a model's expressed confidence and its actual accuracy (describe in Section 2.1.2). For a model that outputs a probability distribution over possible outcomes for a given input x, its expressed confidence $\hat{p}$ is the probability assigned to its chosen prediction $\hat{y}$:

$$\hat{p} = \max_k p(y = k \mid x)$$

A model is considered to have *good functional metacognitive calibration* if this expressed confidence directly corresponds to the likelihood of being correct. Formally, this is stated as:

$$P(\hat{Y} = Y \mid \hat{P} = \hat{p}) = \hat{p}$$

In other words, across a large number of predictions, a well-calibrated model that states it is 90% confident will be correct on 90% of them. Achieving this from modern deep learning models is a non-trivial engineering challenge. The following methods represent prominent approaches to generating the robust probabilistic outputs from which calibrated confidence signals can be derived.

One prominent approach is *Bayesian Deep Learning (BDL)*. The BDL framework aims to infer a full posterior probability distribution over a model's weights given the training data [125–127]. Instead of a single point prediction, BDL performs Bayesian model averaging to produce a full predictive distribution, $p(y \mid x, D_{\text{train}})$. The probabilities derived from this distribution can then be evaluated for calibration using the formalism above, but their quality depends on how well the approximation technique estimates the true, intractable posterior distribution.

A powerful and practical alternative is the use of *Deep Ensembles*. This non-Bayesian method involves training an ensemble of several identical neural networks independently, differing only by their random weight initialization and data shuffling [128–133]. At inference time, the predictions from all models in the ensemble are aggregated. The mean of the individual model predictions is used as the final probabilistic estimate, which has been shown to produce well-calibrated outputs and achieve strong performance in UQ benchmarks.

The core idea for methods like Bayesian Deep Learning (BDL) and Deep Ensembles is that they don't produce a single answer but a predictive distribution over possible outcomes. *Functional metacognitive calibration* is a measure of how reliable the probabilities from this distribution are.

Metacognitive Sensitivity

Metacognitive sensitivity is a measure of a model's ability to discriminate between its own correct and incorrect predictions (describe in Section 2.1.2). Conceptually, a sensitive model is one that reliably expresses higher confidence when it is correct and lower confidence when it is wrong. To engineer this capability, we need a robust, quantifiable signal of the model's internal uncertainty. Evaluating this

requires a method that can provide such a signal, for which *Monte Carlo (MC) Dropout* is a common and efficient technique.

MC Dropout generates an uncertainty estimate by performing T stochastic forward passes of the same input x through the network with dropout layers active at test time [134–138]. This process creates a distribution of differing predictions, $\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T\}$. The variance of this distribution serves as a measure of the model's epistemic uncertainty (i.e., uncertainty in the model's parameters):

$$\text{Uncertainty}(x) \propto \text{Var}(y) \approx \frac{1}{T} \sum_{t=1}^{T} (\hat{y}_t - \mathbb{E}[\hat{y}])^2$$

This variance serves as a direct, quantitative signal of the model's certainty: a lower variance indicates higher confidence in its prediction, while a higher variance indicates uncertainty. For this signal to be meaningful in a metacognitive sense, it must correlate with the model's actual performance. A model demonstrates high functional metacognitive sensitivity if it reliably assigns lower variance to its correct predictions than to its incorrect ones. In essence, this means that the distribution of variance values for correct predictions is statistically separable from the distribution for incorrect predictions, providing concrete evidence of the model's ability to discriminate between its own successes and failures.

3.4 Metacognition Experimental Paradigms

A comprehensive understanding of joint metacognition requires robust empirical methods capable of measuring internal cognitive states and dynamic social interactions, which are not directly observable. Researchers in human-computer interaction (HCI), human-robot interaction (HRI), and cognitive science have developed a diverse toolkit of experimental paradigms and measurement techniques to address this challenge. Given the multifaceted nature of metacognition, a multi-method approach is often necessary. In this section we provide a review of these methodological approaches and the experimental paradigms in which they are commonly employed.

Experimental Paradigms

These are overarching structures or scenarios designed to elicit and study metacognitive behaviors. A classic approach involves *Dyadic Decision-Making Tasks*, where two participants first make individual decisions and rate their confidence. They are then shown their partner's decision and confidence before reaching a joint decision together [35, 46, 47, 50, 139]. To study social influence, researchers often use *Conformity Paradigms*, which investigate how an individual's judgment is shaped by the opinions of others–including AI agents that may provide deliberately incorrect advice [140–142]. For more realism validity, *Human-in-the-Loop Simulations* (e.g., robot teleoperation or medical screening) are employed to study joint inference under realistic conditions of time pressure, uncertainty, and high cognitive load [119, 143–146]. Finally, many studies use *Adapted Cognitive Psychology Paradigms*, where well-established paradigms are modified to target metacognition. For example, false memory tasks can be used to examine confidence in errors, and psychophysical tasks can precisely measure metacognitive sensitivity [147–150].

Key Measurement Techniques

Within these paradigms, researchers employ several specific methods to collect data on a participant's cognitive state. These can be broadly grouped into subjective, behavioral, and physiological measures.

1. Subjective Measures: These methods rely on participants' self-reports of their internal states and are the most direct way to probe conscious metacognitive judgments.

    - Confidence Ratings: Participants rate their belief in the correctness of a decision on a numerical or categorical scale [27, 35, 36, 151].
    - Standardized Questionnaires: Validated surveys are used to assess constructs like perceived mental workload (e.g., NASA-TLX) [152], trust (e.g., Trust Perception Scale-HRI), or situation awareness [153, 154].

- Think-Aloud Protocols: Participants continuously verbalize their thought processes as they perform a task, providing a rich qualitative stream of their ongoing reasoning and monitoring [145, 155].

2. Objective Behavioral Measures: These methods focus on observable, quantifiable actions and performance outcomes.

   - Performance Metrics: Core measures include decision accuracy (percentage of correct choices), reaction time, and task completion time [156].
   - Reliance Behavior: This is a key measure in human-AI collaboration, typically operationalized as the rate at which a human accepts or rejects an AI's advice [156].

3. Physiological Measures: These methods involve recording the body's physiological responses, which are correlated with cognitive and affective states, offering an implicit and continuous window into a user's state [157, 158].

   - Electroencephalography (EEG): Records brain-based event-related potentials (ERPs) associated with specific cognitive events, such as the conscious detection of an error [159, 160].
   - Functional Near-Infrared Spectroscopy (fNIRS): A non-invasive neuroimaging technique measuring changes in blood oxygenation in the prefrontal cortex that are correlated with cognitive workload [161, 162].
   - Galvanic Skin Response (GSR): Also known as Skin Conductance (SC), this method measures changes in the electrical properties of the skin that reflect sympathetic nervous system arousal [163, 164].

3.5 Discussion

In this chapter, we reviewed the literature on how humans collaborate in joint inference and decision-making, with a particular emphasis on how confidence is shared between partners. We then examined related work in human–agent

collaboration, highlighting gaps in how confidence and uncertainty are represented and communicated.

A recurring theme in this literature is that effective human–AI collaboration depends on the quality of the "cognitive interface": the mechanisms through which humans and AI exchange information about their decision processes. Current AI-assisted models often face issues of over-reliance or under-reliance, in part because AI systems provide limited or poorly calibrated signals about their own uncertainty. This motivates the need for more structured approaches to how confidence is expressed and integrated in joint decision-making.

For human–robot collaboration, additional challenges arise from embodiment and dynamic, spatiotemporal tasks. Robots must often infer hidden human states such as trust or intent, yet humans are rarely given comparable access to the robot's internal confidence or reliability. This asymmetry suggests an open research problem: how to design confidence-signaling mechanisms that are both calibrated (matching actual accuracy) and sensitive (able to discriminate between correct and incorrect inferences) in interactive settings.

Finally, we surveyed frameworks in which metacognition can be engineered into operational AI models and examined experimental paradigms for measuring metacognition in humans. This provides the foundation for our research, which investigates how structured exchange of confidence can improve joint human–agent inference.

4

Confidence-Based Heuristics for Human-Human Joint Inference

Contents

4.1 Introduction 49
4.2 Methodology 53
4.2.1 Experiment Setup 53
4.2.2 Procedure 54
4.2.3 Data Collection 55
4.2.4 Metrics 55
4.3 Results 56
4.3.1 Impact of Confidence on Accuracy 57
4.3.2 Effect of Performance Discrepancy 58
4.3.3 Influence of Metacognitive Sensitivity 61
4.3.4 Dyadic Metacognitive Sensitivity Correlation 65
4.4 Discussions 67

4.1 Introduction

Building on the foundational challenges of human-robot teaming introduced in Chapter 3.2.2, in this chapter we first investigate the role of confidence as a heuristic in human joint decision in the teleoperation task. The teleoperation mode of interaction is critical in demanding domains ranging from industrial inspection and nuclear decommissioning to dynamic search and rescue missions. In these

environments, an operator's success hinges on making cognitively demanding choices, such as selecting the most suitable robot from a fleet, choosing the appropriate controller for a delicate task, or determining the optimal level of autonomy for a robot to operate at under pressure [2, 165–169]. The difficulty of these decisions is often compounded by real-world operational factors, including communication latencies, incomplete information, and the unpredictable nature of the environment itself. Given these pressures, a single operator can become overwhelmed, leading to cognitive overload, stress, and decision-making biases that can compromise mission success [94–98]. This vulnerability of the individual operator raises a foundational question central to this thesis: **Could two decision-makers working together make better decisions than one individual acting alone?**

In response to the performance limitations of a single operator, the body of research suggests that collaborative decision-making offers a promising solution. The underlying principle is that while any individual has inherent weaknesses, a dyad of operators can leverage complementary skills to offset each other's shortcomings and improve overall performance. This concept is well-established; numerous studies in human-human interaction (HHI) confirm that collaboration leads to superior decision outcomes [33–37], and there is a strong case for transferring these findings into the domain of human-robot interaction (HRI) [170–174]. Indeed, this principle is already applied in practice, where many teleoperation scenarios involve multiple humans operating a single robot. For instance, mobile robot search and rescue teams often utilize an operator-to-robot ratio of 2-to-1 or higher, while Unmanned Aerial Vehicle (UAV) missions frequently assign distinct roles like 'pilot' and 'sensor operator' to different team members. As discussed in Section 3.2.2, the success of these multi-operator teams hinges on factors like team dynamics and communication, reinforcing the importance of understanding the fundamental mechanisms of effective collaboration.

While the benefits of multi-operator teams are evident, a central question concerns the specific strategies that enable dyads to outperform individuals. A particularly promising approach, emerging from research in HHI collaborative

cognition, is Maximum Confidence Slating (MCS) (Section 2.2) [49, 53–56]. This strategy dictates that when partners disagree, the joint decision should default to the choice of the individual who expresses a higher degree of confidence, operating on the assumption that individuals can reliably monitor their own performance. Although MCS has been proven to enhance joint performance in static domains like visual perception or knowledge-based tasks, its applicability to active, dynamic decision-making in robotic control remains inadequately studied–a research gap previously identified in Section 3.2.2. This chapter addresses that gap directly by investigating the performance of MCS in a spatio-temporal task. Unlike previous work centered on passive information processing, our study requires human operators to actively control simulated robots, thereby gathering and interpreting dynamic information to make real-time decisions.

To investigate this, an online user study was conducted with 100 participants. The experiment required each participant to control two simulated robots with different, unstated control latencies and subsequently identify which robot they believed was more responsive, as well as their confidence in that decision. The collected data was used to form virtual dyads, allowing for the application of the MCS approach to compare the accuracy of joint decisions with individual performance using an *accuracy gain*[1] metric. The analysis of the results is structured around four primary areas. First, it examines the direct impact of confidence on decision accuracy. Second, it assesses how performance discrepancies between members of a dyad affect the accuracy gains derived from MCS. Third, it investigates the influence of individual metacognitive sensitivity (see Section 2.1.2) on these outcomes. Finally, it explores the correlation between a dyad's collective metacognitive sensitivity and the resulting accuracy of its joint decisions.

Our findings reveal the following key insights into the effectiveness of the MCS-based decision-making process in joint decision making within robotics tele-operation domains:

[1]Accuracy gain is defined as the accuracy increase/decrease difference between the accuracy (%) of decision made by the MCS-based joint decision-making agent as compared with the higher performing individual in the dyad.

Table 4.1: A description of Chapter 4's contribution to joint inference research.

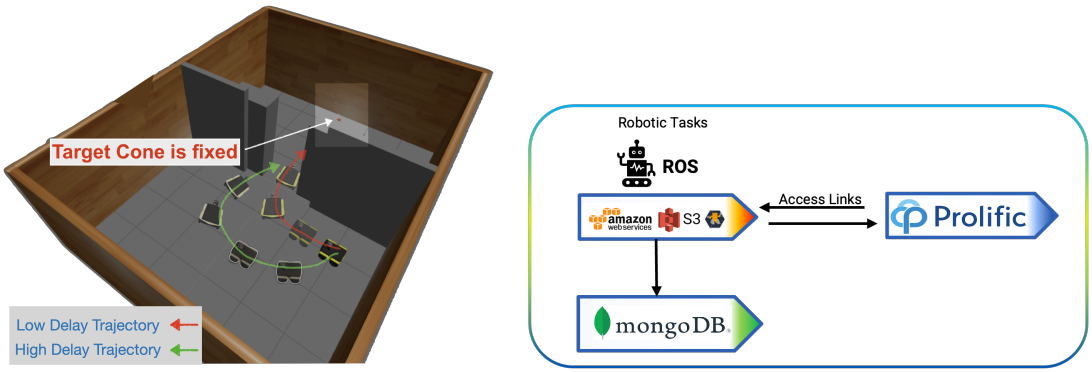
Dimension	Chapter 4	Chapter 5	Chapter 6
Collaboration Type	Human-Human dyads	Human-AI dyads	Human-AI dyads
Heuristic?	Yes	Yes	No
Adaptive?	No	Partially	Yes
Bandit Approach?	None	Thompson Sampling (MAB)	Contextual & Latent Bandit
Focus of Study	Metacognitive sensitivity & heuristics in joint inference	AI-assisted decision making under uncertainty	Unified framework with context awareness

1. **Impact of Confidence on Accuracy:** Joint decisions determined by high confidence via MCS were more accurate than those made by the highest-performing individual in the dyad. Conversely, joint decisions based on low-confidence inputs were less reliable than random choices, emphasizing the need to avoid low-confidence responses. (Section 4.3.1)

2. **Effect of Performance Discrepancy:** To the best of our knowledge, this is the first time we have observed that larger performance gaps between dyad members result in smaller accuracy gains from MCS. Pairing participants with similar performance levels yielded significantly higher accuracy, highlighting that MCS is more effective when participants have comparable competence. (Section 4.3.2)

3. **Influence of Metacognitive Sensitivity:** Metacognitive sensitivity [27] significantly impacted MCS accuracy gains. For individuals with above-average metacognitive sensitivity, MCS-based accuracy gains were stable regardless of sensitivity differences. For those with below-average sensitivity, pairing similarly calibrated individuals led to negative accuracy gains, while teaming individuals with diverse sensitivity levels improved the accuracy of the dyad. This marks the first time such analyses have been conducted on joint decision-making in dynamic tasks. (Section 4.3.3)

4. **Correlation with Dyadic Metacognitive Sensitivity:** Accuracy gains from MCS were positively correlated with dyadic confidence sensitivity. Higher overall metacognitive sensitivity of the dyad led to better decision accuracy,



(a) Robotic Task **(b)** Software Framework

Figure 4.1: Software Framework and User Study Scenario.
Fig. 4.1a presents the scenario with the red and green curves represent trajectories driven by a participant from our user study who was experiencing low and high delay in control of the mobile robot.
Fig. 4.1b presents the software framework for the web-based robot navigation simulator. The simulator is hosted on AWS, and data is collected and stored in a MongoDB database. The participants for the user study were recruited via the Prolific Platform.

further underscoring the importance of metacognitive sensitivity in joint decision-making. (Section 4.3.4)

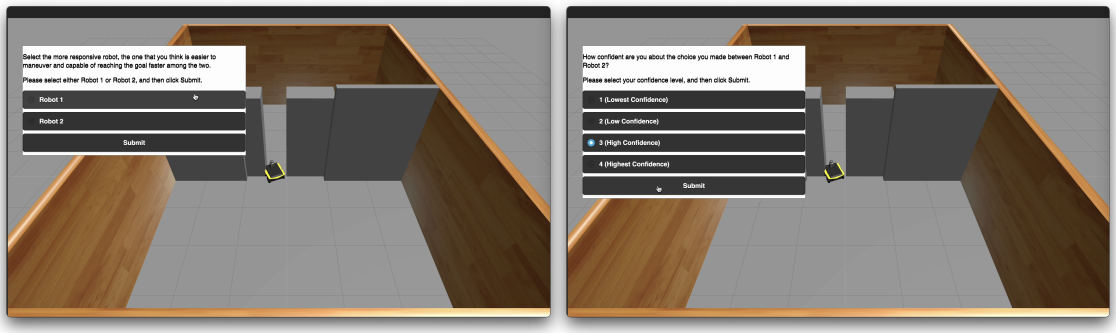
This chapter describes the work published in the Proceedings of ACM/IEEE International Conference on Human-Robot Interaction (HRI) 2025.

4.2 Methodology

We conducted an online study, approved by the University Research Ethics Committee, to empirically investigate how human make joint decisions when selecting robot controllers.

4.2.1 Experiment Setup

The experiment tasked participants with navigating two visually identical simulated robots through a narrow gap using an interface shown in Figure 4.1a. The environment consisted of 24 different conditions, combining 6 possible initial robot poses with 4 doorway configurations. A fixed goal location was marked by a traffic cone, with the initial robot pose and door configuration randomly sampled for each trial. Participants were recruited via the Prolific Platform, which provided



(a) Robot Choices **(b)** User Confidence Levels

Figure 4.2: Study Interface: After teleoperating the two robots, participants choose which robot they deemed to have lower control delay (Fig. 4.2a), and provide their associated confidence (Fig. 4.2b).

an overview of the study and ethics approval information. Those who agreed to participate and were at least 18 years old were shown a detailed video explaining the interface components and demonstrating example trials. Before the main experiment, participants completed 5 practice trials to familiarize themselves with the keyboard controls: W for forward, A for anticlockwise rotation, S for backward, and D for clockwise rotation.

4.2.2 Procedure

The main experiment consisted of 100 trials. In each trial, participants controlled each robot for 6 seconds in a random order. This duration was determined through a separate pilot study (Appendice A) as the minimum time needed for decision-making. To create a perceptible difference between the robots, we injected time delays into their controls. One robot was randomly assigned a fixed 50 ms delay between key press and action, while the other had a variable delay ranging from 70 ms to 150 ms. We implemented a two-down one-up staircase[2] procedure to adjust task difficulty. This procedure ensured that the task was not so easy as to enable ceiling performance and not so hard to make participant performance close to random [175]. This precluded participants from being overly confident or underconfident on the

[2]The two-down one-up staircase is procedure where the task was made harder after two consecutive successful choices by decreasing the delay by one level step, and the task was made easier after one failure by increasing the delay by one level step.

task. During each trial, participants first controlled one robot, then the other, with a brief pause between. The trial number and elapsed time were displayed throughout. After controlling both robots, participants selected which they believed had the lower delay and rated their confidence in this choice on a Likert four-point scale (1:Lowest, 2:Low, 3:High, 4:Highest). This experimental design allowed us to assess both decision-making accuracy and metacognitive sensitivity in human-robot interaction tasks under varying degrees of difficulty. Figure 4.2 shows the trial procedure.

4.2.3 Data Collection

To ensure data quality, we implemented exclusion criteria based on task performance and engagement. Participants with accuracy below 65% or those who gave the same confidence rating for over 95 out of 100 trials were excluded. This 65% threshold, set above the 50% chance level, allowed us to retain participants actively engaging with the task while excluding those who might have been inattentive or unable to perform effectively. The exclusion criteria and performance metrics are consistent with standard procedures in human self-confidence research [176–178], ensuring our results are comparable with existing literature in the field. After applying these criteria, our final analysis included data from 80 participants: 47 males and 33 females. The mean age was 37 years (SD = 11 years). On average, participants took 1.2 hours (SD = 0.25 hours) to complete the experiment. To mitigate and account for learning effects during the extended experimental sessions, we implemented a structured familiarization phase prior to data collection. Participants performed a minimum of 10 practice trials until their performance reached a stable plateau, ensuring that the subsequent measures of metacognitive sensitivity and calibration reflected true task competence rather than initial skill acquisition.

4.2.4 Metrics

We quantified task performance using two primary metrics:

1. Accuracy: The proportion (%) of correct selections of the lower-delay robot.

2. Metacognitive Sensitivity: Using AUROC2 computation, metacognitive sensitivity measures how well participants' confidence ratings align with their actual performance. We have discussed a detailed computation of AUROC2 in Section 2.1.2.

These metrics allowed us to evaluate both the participants' ability to discriminate between the robots based on delay and their metacognitive awareness of their performance. By combining accuracy and metacognitive sensitivity measures, we aimed to gain comprehensive insights into human decision-making and self-assessment in this human-robot interaction task. This enables us to understand not just how well participants performed, but also how accurately they judged their own performance across varying levels of task difficulty.

4.3 Results

From the data obtained from 80 participants, we combine pairwise participant data to form virtual dyads [57]. We obtained 3160 virtual dyads $\binom{80}{2}$. For each virtual dyad, we selected those trials where both the participants faced the same delay pair. From these, we eliminated those trials where both participants agreed on their choice of robot, and further analysed those trials where the participants differed in their choice such that a joint decision was required.

The pairing process was repeated for all delay-pair conditions and 4,148 virtual joint decision-making trials were generated from real data. Of these, 1,932 trials involved disagreements, requiring joint decisions for this dyad.

Within each dyad, the member with a higher percentage of correct responses was designated as high-performing (HP), while the other was designated as low-performing (LP). Additionally, three dummy participants were created: dummy high-confidence (D-HC), dummy low-confidence (D-LC), and dummy random (D-Random). For each trial, the response of the participant who indicated higher confidence was selected as D-HC participant, and the other as the D-LC participant. **D-HC corresponds to the Maximum Confidence Slating (MCS) approach**

of joint decision-making. D-Random represented a joint decision-maker who picks randomly between the choices made by both the participants in the dyad. Accuracy was then calculated for these five participants.

We then present our analysis formulated from the four backbone research questions. Formally, the independent variables are (1) individual confidence, (2) performance discrepancy, and (3) metacognitive sensitivity. The dependent variables include joint decision accuracy, accuracy gains from MCS, and the dyadic AUROC2 value measuring the dyad's metacognitive sensitivity.

4.3.1 Impact of Confidence on Accuracy

Figure 4.3 illustrates the accuracy of robot selection across all five participant types, based on 3,160 virtual dyads. We assessed the performance of D-HC against other participants using the paired t-test. The results show that D-HC exhibited significantly higher accuracy compared to HP (Highest Performing individual), $(t(3158) = 8.36, p < .0001)$. This confirms that decisions made using the Maximum Confidence Slating (MCS) approach provide a significant advantage over those made by the highest-performing individual within the dyad. The accuracy of D-HC was also significantly greater than that of D-Random, $(t(3158) = 33.48, p < 0.0001)$. Interestingly, D-Random performed better than D-LC, $(t(3158) = 33.48, p < 0.0001)$. This suggests that, in any given scenario, a low-confidence response should not be favored. Additionally, we confirmed that D-HC had significantly higher accuracy than D-LC, $(t(3158) = 62.06, p < 0.0001)$, reinforcing that high-confidence decisions lead to better outcomes than low-confidence ones.

Through this analysis, we validated **H1: MCS is effective in robot teleoperation control selection task. Picking the higher confidence decisions within a dyad result in more accurate joint decisions than those made by the highest-performing individual alone. Conversely, low-confidence decisions yield accuracy levels comparable to random choices.**

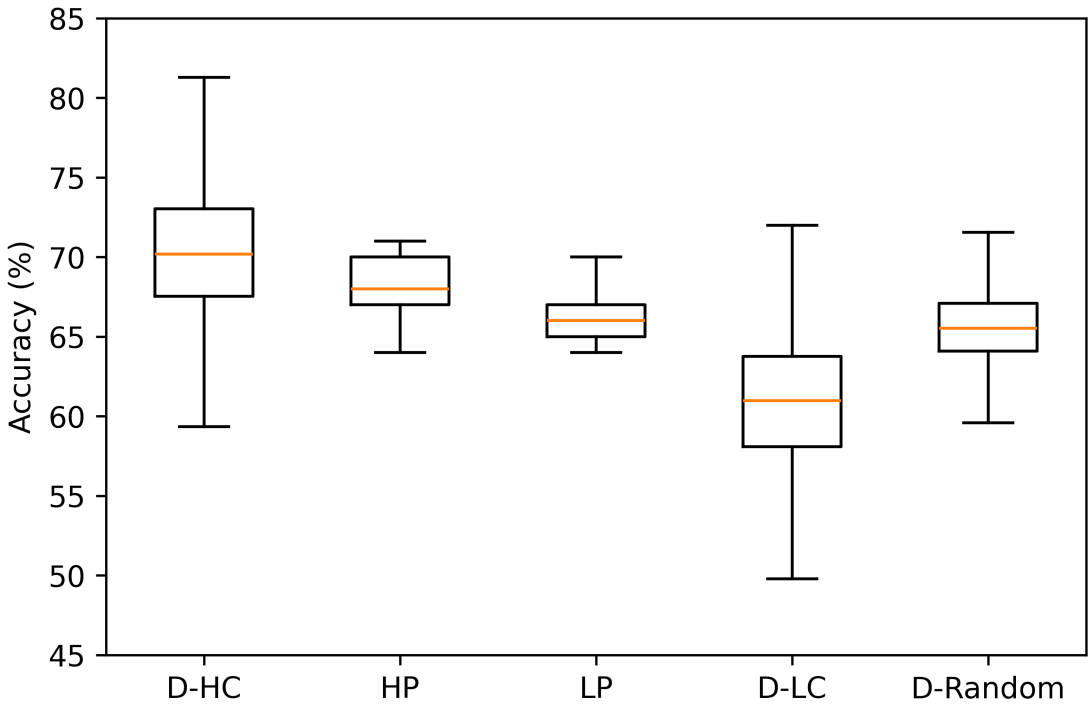


Figure 4.3: Accuracy in robot selection for high-confidence, higher-performing, low-confidence, lower-performing, and random participants from the paired dyads.

4.3.2 Effect of Performance Discrepancy

While MCS-based joint decision-making delivers a benefit overall, specific ways of pairing participants provides further insights. In this section, we analyze the impact of pairing participants according to their performance level. We sorted the 3160 virtual dyads according to the **difference** in the performance of both individuals in each dyad, as measured using two factors:

1. Success rate (%) over the 100 trials.

2. Settling difficulty level of task via the two-down one-up procedure.

For every virtual dyad, we extracted the corresponding *accuracy gain* [55]. Figure 4.4 shows the impact of difference in performance level on the resulting accuracy gain of the dyad. The difference in success percentage between the individuals in each virtual dyad ranged from 0 to 11%. As the dissimilarity in success rate increases, the resulting accuracy gain decreases (regression coefficient : $r = -0.48, p < 0.0001$).

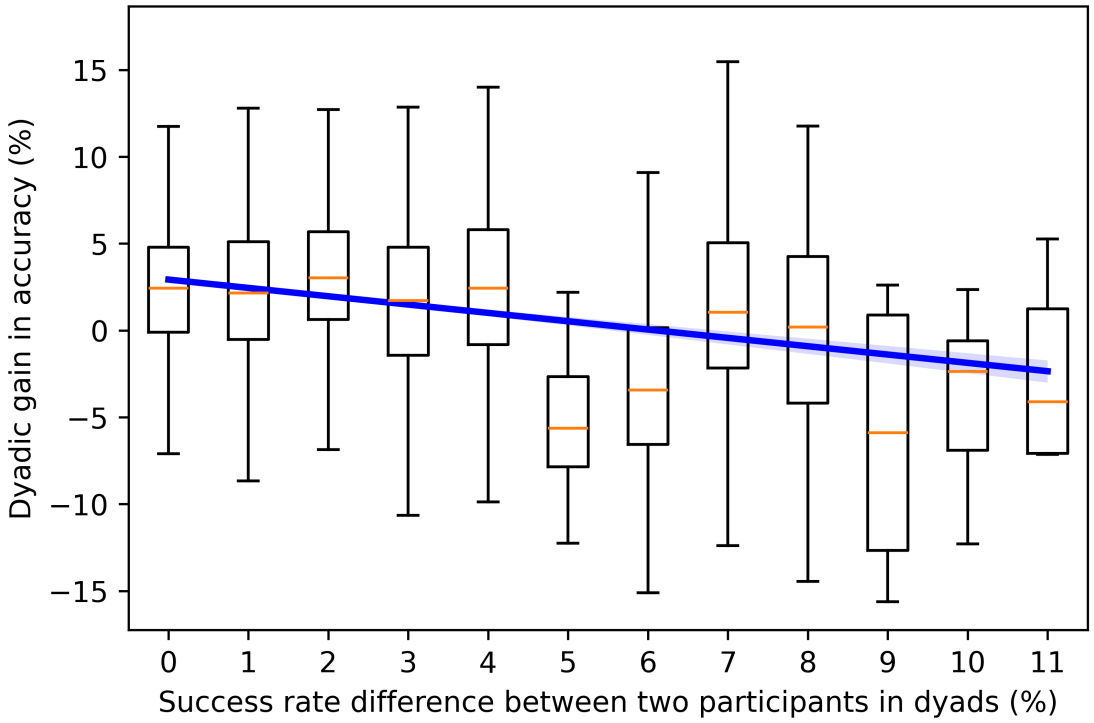


Figure 4.4: Variation in accuracy gain as participants of different success rates are paired together. Pairing similar performing participants leads to higher benefit.

The accuracy gain corresponding to the dyads with no difference in success rate (i.e., identical skill level) was significantly higher than both dyads with a 5% difference in success rate ($t(1008) = 7.1, p < 0.0001$), as well as dyads with a 10% difference in success rate ($t(694) = 8.35, p < 0.0001$). Pairing participants with a success rate difference of more than 8%, leads to a negative accuracy gain, i.e., MCS causes poorer joint performance as compared to the better performing individual.

For a more granular understanding of the impact of skill difference (similarly skilled can correspond to both being highly skilled or both being poorly skilled), we divided the participant data into two groups: above mean skill level (success percentage $>$ 67%), and below mean skill level (success percentage $<$ 67%). Figure 4.5 shows the variation in accuracy gain with difference in performance of the two individuals for the two groups of participants. While the trend is similar for both groups (blue: both participants being below the mean, and red: both participants being above the mean), the absolute values of the accuracy gains show that when both participants are below mean performance the benefit provided

by MCS-based joint decision is higher. When both participants are above mean performance, the benefit is more observable for participants of similar performance level (lower success rate difference).

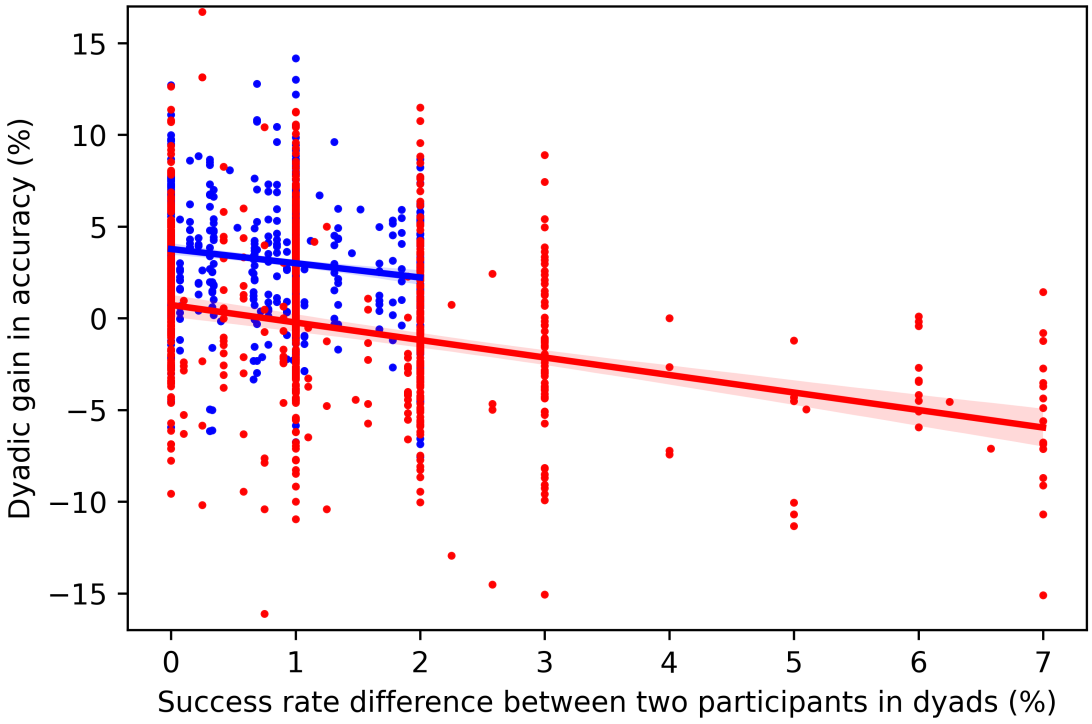


Figure 4.5: Variation in accuracy gain as participants of different success rates are paired together. Pairing similar performing participants leads to higher accuracy gain.

Figure 4.6 shows the impact of pairing based on performance, measured by the task's settled difficulty level. Virtual dyads are grouped into three categories based on the delay difference faced by the two participants when task difficulty plateaus. A 0 ms value on the x-axis indicates similar performance between participants, while a 40 ms value indicates differing performance levels. We see that the resulting accuracy gain is again higher for participants with similar performance level. The accuracy gain is higher for the virtual dyad with similarly performing participants 0 ms as compared to both 20 ms ($t(5752) = 9.67, p < 0.0001$) as well as 40 ms ($t(3280) = 11.04, p < 0.0001$).

This analysis supports **H2: Similar task performance leads to higher accuracy gains. Specifically, the accuracy gains from MCS are greater in**

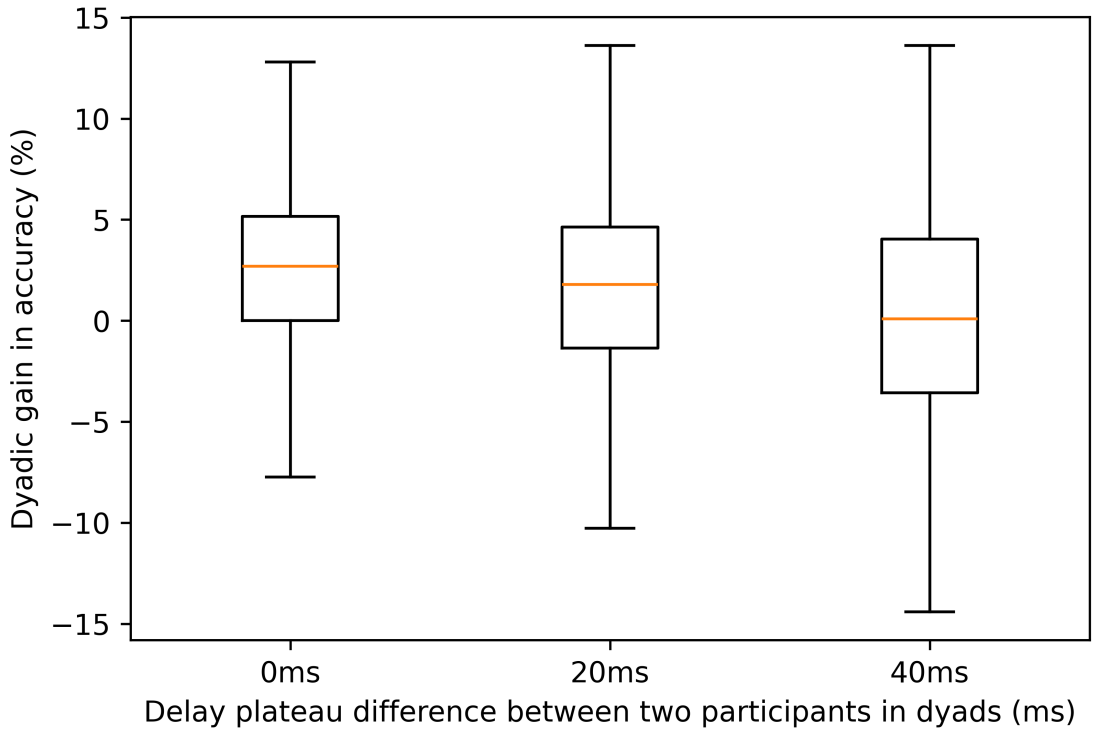


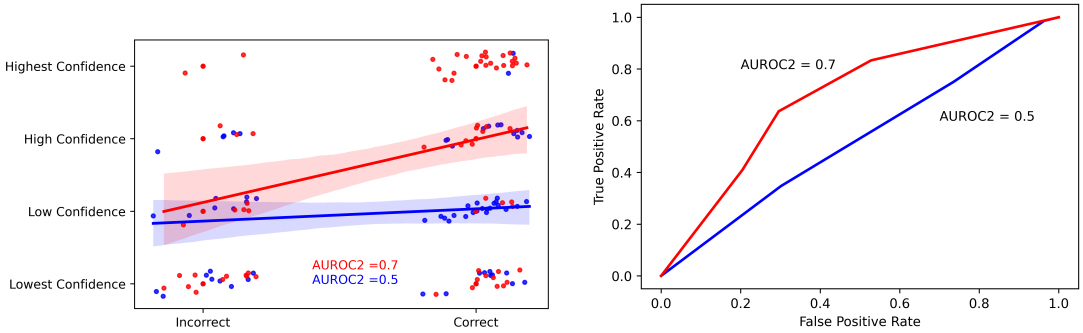
Figure 4.6: Variation in accuracy gain as participants of different performance levels are paired together. Performance level is assessed by the difficulty of the task as established by delay plateau difference. Pairing similar performing participants leads to higher accuracy gain.

dyads where the performance levels of the members are similar, compared to dyads with significant performance discrepancies. This is the first time it is observed that larger performance gaps between dyad members result in smaller accuracy gains.

4.3.3 Influence of Metacognitive Sensitivity

In this section, we paired participants based on their metacognitive sensitivity and analyzed the resulting dyads to gain further insights into the performance of the MCS approach in joint decision-making.

Figure 4.7 illustrates the variation of confidence rating with the correctness of response for two different participants at either end of the spectrum of AUROC2 values. The (red) participant has better sensitivity (AUROC2 = 0.7) than the (blue) participant (AUROC2 = 0.5). Figure 4.7a shows a plot of trial result



(a) Correlation between confidence and outcome. **(b)** Corresponding AUROC2 curves.

Figure 4.7: Metacognitive sensitivity of two different participants.
4.7a: Plot of trial result vs. confidence rating. The line is the regression fit with 95% confidence interval.
4.7b: ROC curves for both participants. The participant represented by red demonstrates higher metacognitive sensitivity compared to the participant represented in blue.

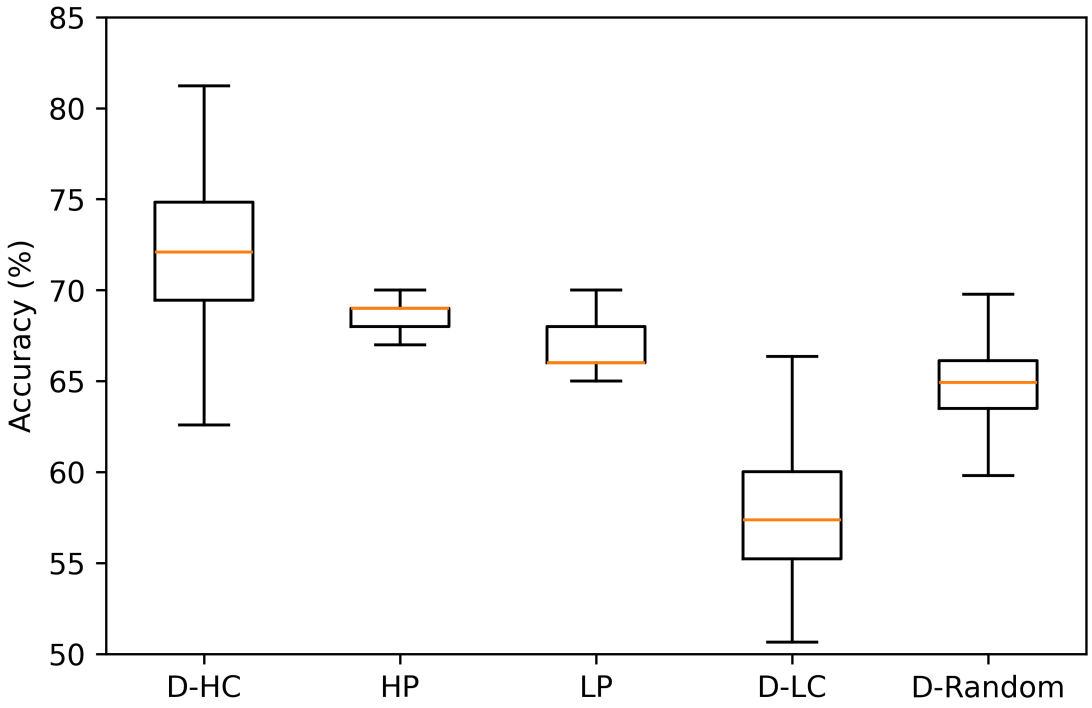


Figure 4.8: Choice accuracy (%) of dyads with both participants of good metacognitive sensitivity (288 pairs).

versus associated confidence value and Figure 4.7b shows the corresponding ROC curves for both the participants.

To analyze the impact of metacognitive sensitivity similarity on accuracy gain, participants were grouped based on their AUROC2 values: above the mean

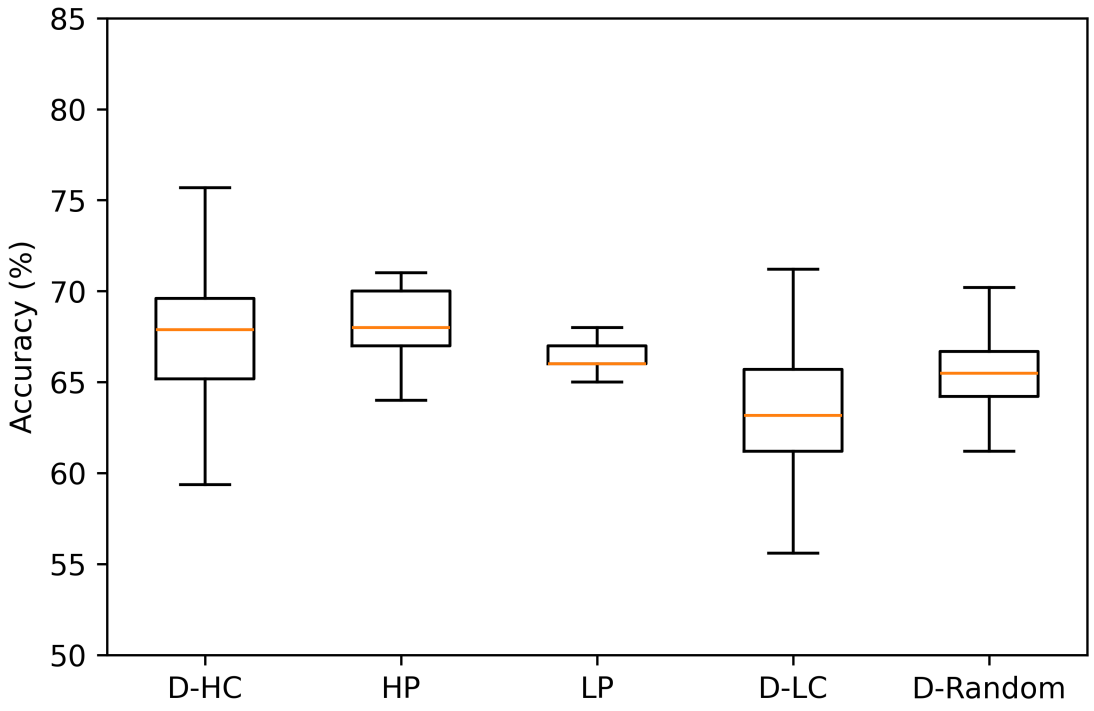


Figure 4.9: Choice accuracy (%) of dyads with both participants of poor metacognitive sensitivity (242 pairs).

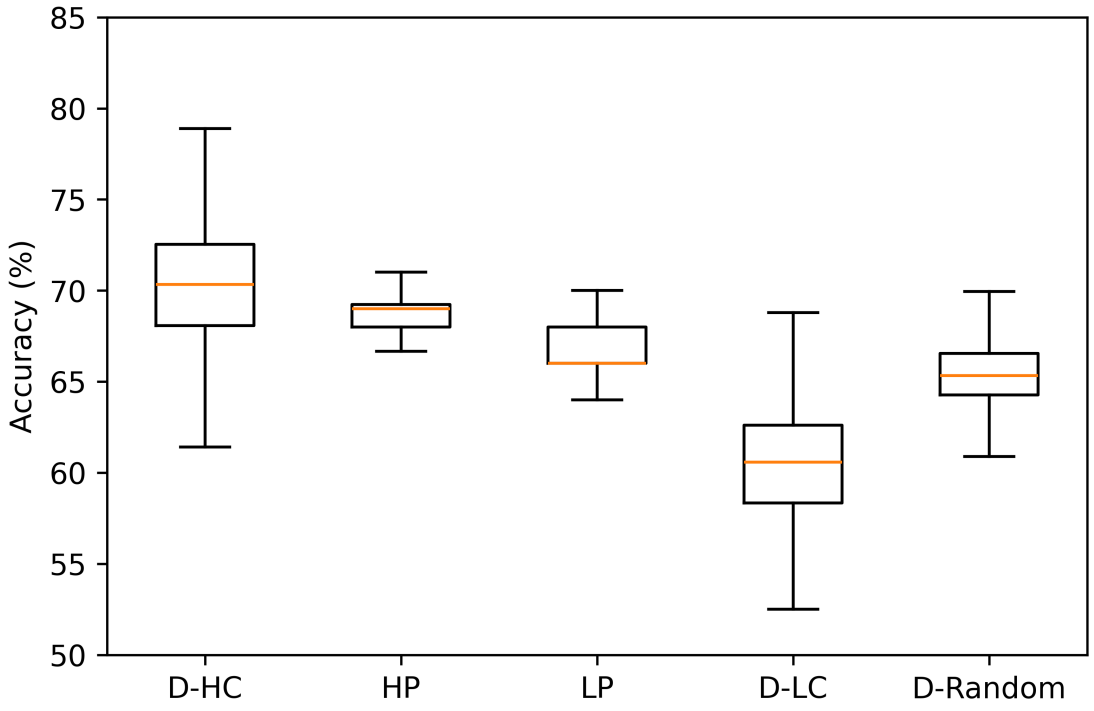


Figure 4.10: Choice accuracy (%) of dyads with one participant of good metacognitive sensitivity paired with one of poor metacognitive sensitivity (264 pairs).

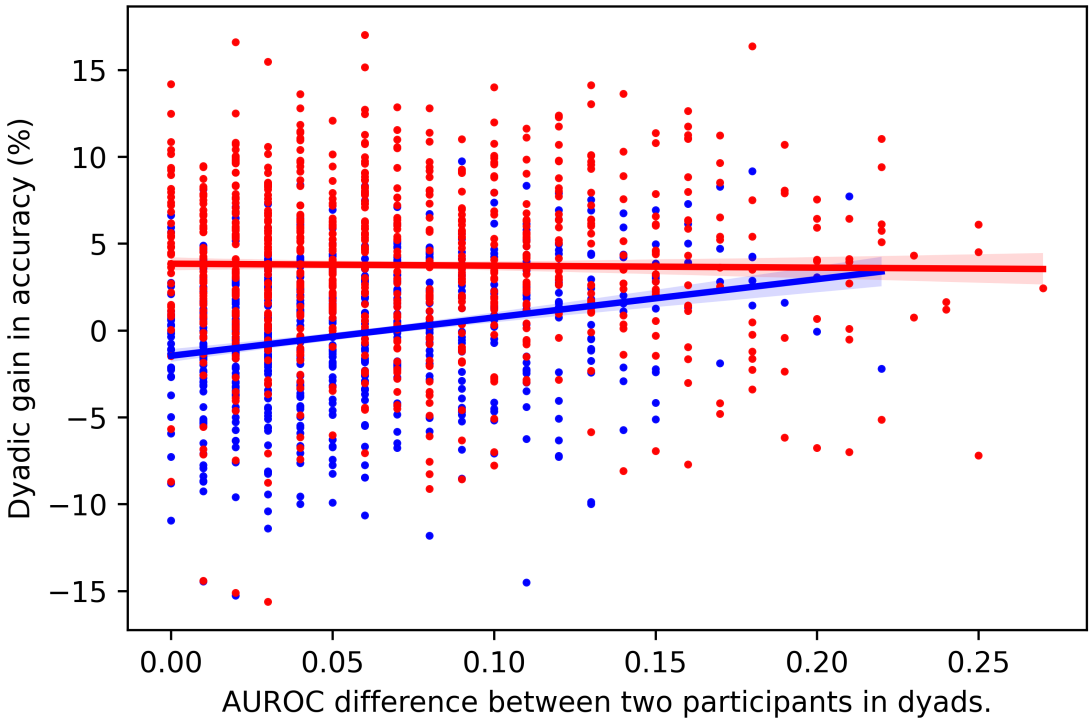


Figure 4.11: Variation in accuracy gain as participants of different metacognitive sensitivity are paired together. For good metacognitive sensitivity (red), pairing participants of similar metacognitive sensitivity leads to higher benefit. Pairing participants with similarly poor metacognitive sensitivity (blue) does not yield a benefit.

(AUROC2 > 0.6) and below the mean (AUROC2 < 0.6). We then paired participants within the same group and analyzed the impact on accuracy gain. Figure 4.11 shows that for participants with AUROC2 above the mean, MCS-based joint decisions with pairs of similar metacognitive sensitivity improved accuracy. However, for those below the mean, pairing similar participants led to negative accuracy gain, with MCS performing worse than the better individual.

We further investigated the impact of pairing participants according to metacognitive sensitivity by selecting the top participants from the group with AUROC2 values above the mean (AUROC2 > 0.65: good metacognitive sensitivity, 25 participants), and the bottom participants from the group with AUROC2 values below the mean (AUROC2 < 0.55: poor metacognitive sensitivity, 26 participants).

We formed virtual dyads by pairing participants based on their metacognitive sensitivity: dyads with two participants of good sensitivity (288 virtual dyads), dyads

with two participants of poor sensitivity (242 virtual dyads), and dyads with one participant of good sensitivity paired with one of poor sensitivity (264 virtual dyads).

Figure 4.8 shows the accuracy when both individuals in the dyad are of good sensitivity. The performance of D-HC was significantly better than HP, $(t(574) = 1.39, p < 0.0001)$. This demonstrates that pairing two individuals results in an increased choice's accuracy compared to the higher performance individual. However, this is not the case when both individuals are of poor metacognitive sensitivity. Figure 4.9 illustrates the accuracy for dyads with two individuals with poor metacognitive sensitivity, where HP outperformed D-HC, $(t(482) = 6.08, p < 0.0001)$. This suggests that in such cases, it is preferable to rely on the more performant individual's choice. Figure 4.10 presents the results when a participant with good metacognitive sensitivity is paired with one with poor metacognitive sensitivity. In this case, D-HC outperformed HP, $(t(526) = 10.86, p < 0.0001)$, indicating that the joint decision was more accurate when at least one individual had good metacognitive sensitivity.

Through this analysis, we confirmed **H3: Participants with above-average confidence calibration achieve stable accuracy gains from MCS, regardless of the calibration differences between dyad members. Conversely, participants with below-average calibration show improved accuracy when paired with individuals who have diverse calibration levels.**

4.3.4 Dyadic Metacognitive Sensitivity Correlation

We further investigated the behavior of resulting dyads formed from pairing individuals together. We computed dyadic metacognitive sensitivity using Koriat's method [57], applying the AUROC2 computation (see Section 2.1.2) to the the concatenated decision-confidence pairs of both participants. For instance, Participant 1 completed 20 trials and Participant 2 completed 15 trials under the same delay condition (50 ms for Robot 1, 70 ms for Robot 2). Of the combined 35 trials, 6 with identical robot choices were excluded, leaving 29 trials for dyadic metacognitive

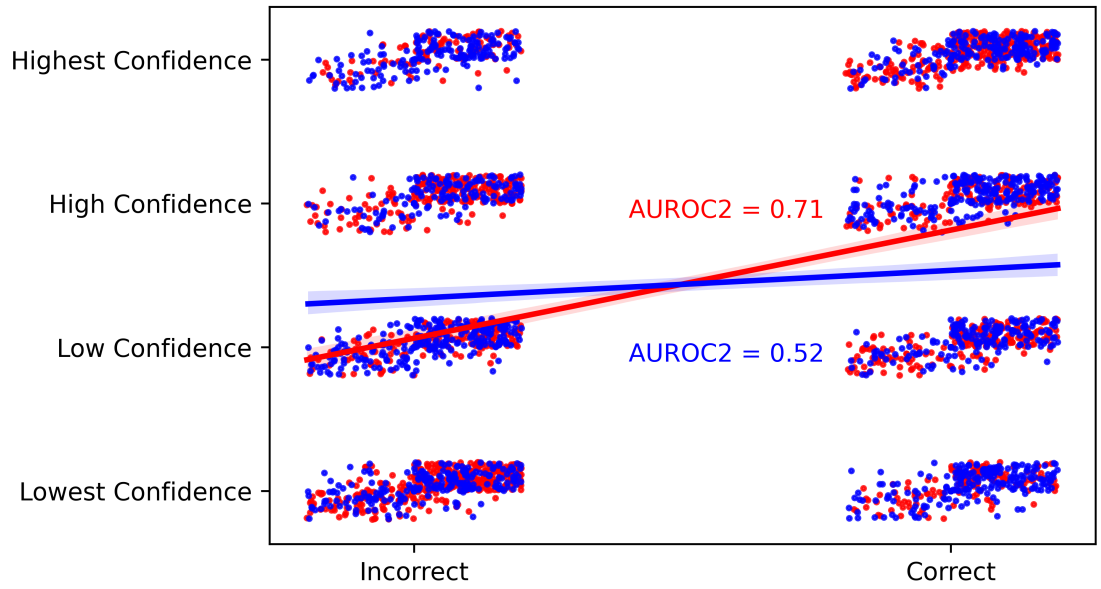


Figure 4.12: Trial results versus confidence ratings for an example virtual dyad: Red indicates a dyad formed by pairing participants of good metacognitive sensitivity, while blue represents a dyad formed by pairing participants of poor metacognitive sensitivity.

sensitivity computation. Figure 4.12 shows the trial result vs confidence rating for an example participant dyad, along with the associated AUROC2 value.

The matecognitive sensitivity of individuals in a dyad affects the resulting dyad's metacognitive sensitivity. Figure 4.13 shows dyadic metacognitive sensitivity based on individual sensitivity. Statistical analysis reveals that dyads of two individuals of good metacognitive sensitivity have higher AUROC2 than those of individuals with poor metacognitive sensitivity ($t(1012) = 28.74, p < 0.001$) or mixed-sensitivity dyads ($t(1078) = 13.28, p < 0.001$).

Finally, our findings highlight the significant impact of dyadic AUROC2 on joint decision-making performance. Figure 4.14 shows a strong correlation between accuracy gain and dyadic metacognitive sensitivity, with higher sensitivity leading to higher joint decision accuracy. This underscored the importance of good metacognitive sensitivity within the dyad for the effectiveness of the Maximum Confidence Slating (MCS) approach.

This analysis supports our fourth hypothesis **H4: Dyads with higher overall confidence calibration demonstrate better decision accuracy when using**

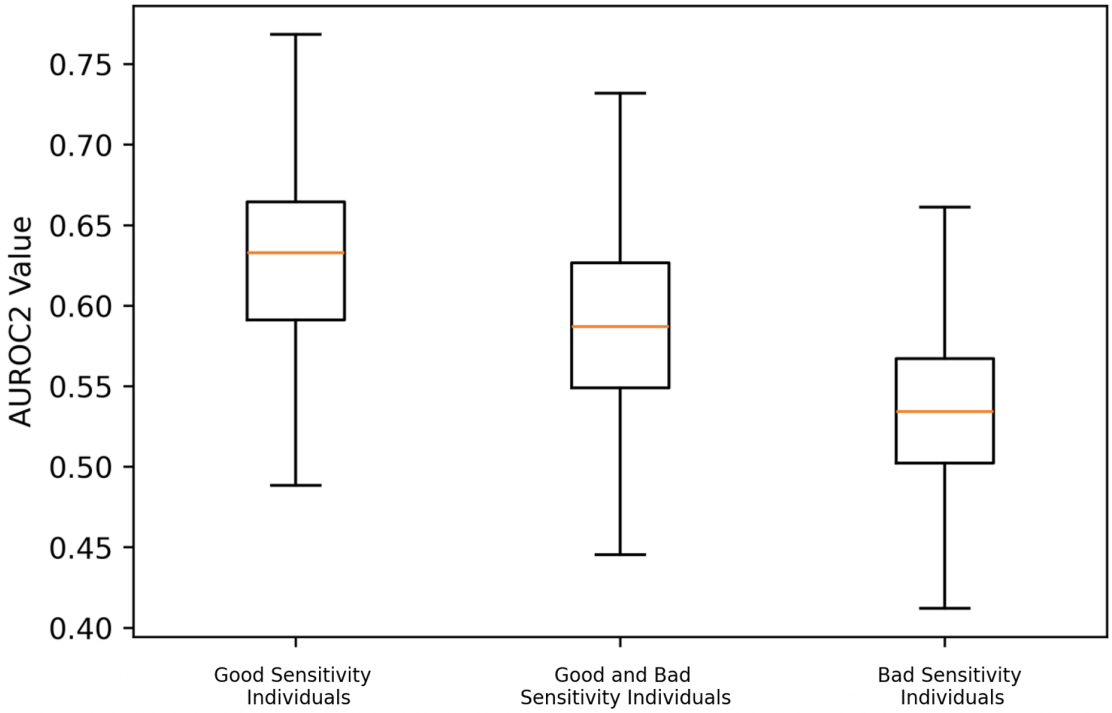


Figure 4.13: Variation in dyadic metacognitive sensitivity based on how participants are paired according to their individual sensitivity.

MCS-based joint decision, compared to dyads with lower calibration. This finding reinforces the value of pairing individuals with well confidence calibration for optimal team performance.

4.4 Discussions

Summary

In Chapter 4.4, we investigated human-human dyad joint decision-making in a robot teleoperation task, focusing on how heuristic Maximum Confidence Slating (MCS) choice selection impacts decision accuracy. To our knowledge, this is the first study applying MCS-based joint decision-making in a dynamic, spatiotemporal task involving active robot control. Our results showed that the accuracy of dyad joint decisions was significantly higher than that of the more skilled individual in the pair. Our findings emphasise the importance of skill similarity and metacognitive sensitivity in achieving better outcomes for human collaboration in robotic tasks, establishing a foundational understanding of MCS's role in the human-robot interaction domain.

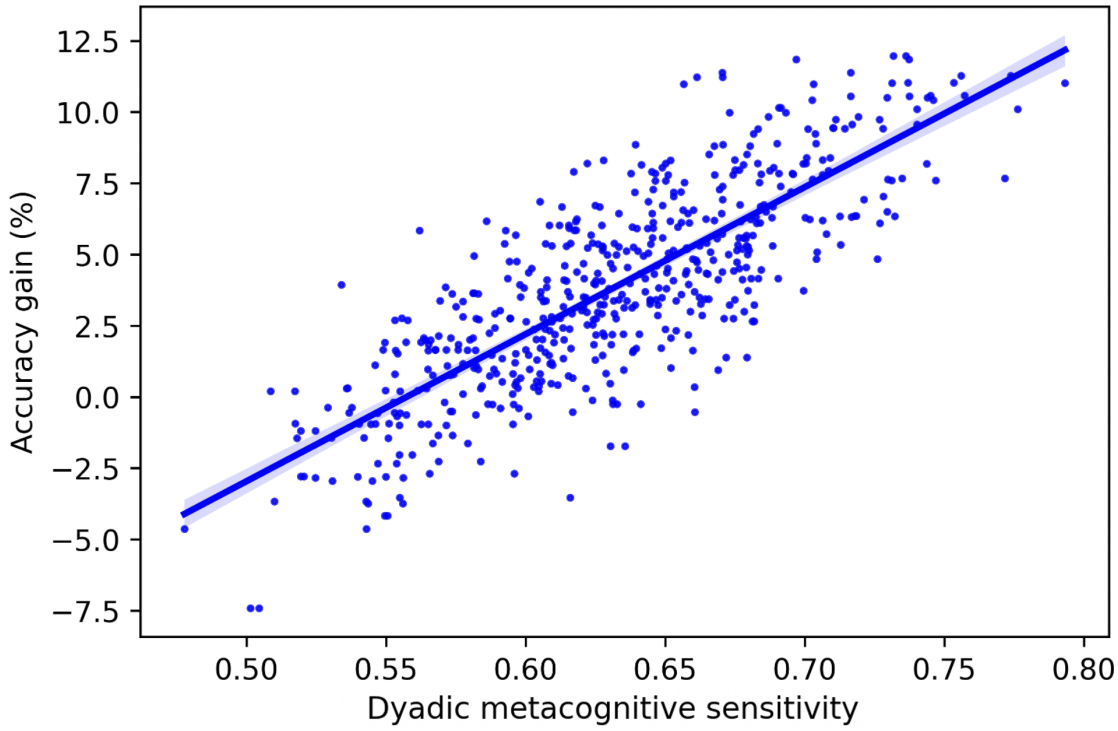


Figure 4.14: Dyadic metacognitive sensitivity versus accuracy gain. Higher dyadic metacognitive sensitivity leads to an increased accuracy gain.

The effectiveness of MCS in this context highlights its potential for improving joint decision-making in human-IDS (Intelligent Decision Support) system dyads.

This work also reveal the potential of MCS for real-world applications requiring critical, time-sensitive decisions. By leveraging confidence as a low-cost, accessible metric, MCS can enhance task efficiency and accuracy by combining confidence levels from two operators to reduce deliberation time, minimize second-guessing, and enable faster and more reliable decisions.

Limitations and Research Directions

The scope of our study was constrained to a specific robot teleoperation and controller selection task, where the primary decision involved selecting between two robots with different control delays. While this research looked into virtual dyads scenarios, to isolate metacognitive signals, real-world joint decision-making between two people involves significant social effects. For instance, the perceived competence of oneself versus a partner can lead to the 'equality bias', a tendency to weight a

partner's input equally regardless of objective skill gaps. Future research can expand to encompass a broader range of tasks and more complex decision-making scenarios.

In Chapter 5.4, we extend this work by specifically applying these findings to the study of human-AI dyads. Additionally, we compare MCS with other *meta*-decision-making strategies to identify the most effective approaches for enhancing joint decision accuracy in human-AI collaboration in robotic mission. Through these research directions, we hope to contribute to the integration of AI systems and development of more effective collaborative Human-Robot Interaction (HRI) settings.

5

Confidence-Based Heuristics for Human-AI Joint Inference

Contents

5.1 Introduction 71
5.2 Methodology 75
5.2.1 Problem Statement 75
5.2.2 Environment and Task 75
5.2.3 AI Decision Support System (AI-DSS) 77
5.2.4 Experimental Procedure 79
5.2.5 Human-AI Joint Inference 80
5.2.6 Performance Metrics 82
5.3 Results 82
5.3.1 Human Inference Dynamics 82
5.3.2 Joint Inference Performance 87
5.4 Discussions 92

5.1 Introduction

In this chapter, with the results obtained from Chapter 4.4, we extend the investigation of confidence-based heuristics from human-human collaboration to the domain of human-AI dyads. The focus here narrows to the specific cognitive challenge of inference, which is the process by which human supervisors must draw conclusions from limited or ambiguous data during complex robot missions. In time-critical

scenarios such as search and rescue, defense, and firefighting, these inferences are often made under adverse conditions, including significant communication latencies, incomplete information, and severe time pressure [179–181]. To mitigate these difficulties, AI-based Decision Support Systems (AI-DSS) have emerged as a promising technology to augment human inference and, consequently, improve the quality of decision-making in human-supervised missions.

While AI-DSS have been used in robot missions such as urban search and rescue [182], their use has largely been explored in *AI-assisted* configurations, i.e., where the final decision to accept or reject the AI-DSS recommendation lies with the human operator. Similarly to non-robotic tasks such as recidivism risk prediction [183] and medical diagnosis [184], the AI-DSS is often used as a "second opinion" system [167, 185]. As detailed in the literature review in Chapter 3.5, this approach has a significant limitation: humans often struggle to determine when to accept or reject the AI's input, which can hinder team performance and prevent the achievement of true human-AI complementarity. An operator might, for instance, be swayed by an AI's suggestion to abandon an initially correct inference for an incorrect one [83, 186, 187]. In response to this challenge, a joint human-AI decision-making paradigm has been proposed as an alternative.

In this collaborative framework, the final team inference can be based on the input of either the human or the AI, which raises the central question of how to best arbitrate between the two when they disagree. However, whose inference should finally be chosen remains an open problem [186, 188].

A promising answer to this arbitration problem comes from the field of human-human joint decision-making, which proposes a heuristic known as Maximum Confidence Slating (MCS). As previously introduced in Chapter 4.4, the MCS approach resolves disagreements by accepting the inference made by the agent with the higher self-reported confidence. This method has been shown to improve team accuracy on tasks involving static, visual inputs, raising the prospect of its use in joint human-AI inference. However, the inference challenges in robotic missions are distinct and more complex, typically involving a combination of three

factors: 1) latent variables, 2) temporal processes, and 3) inference based on actions. Specifically, human supervisors have to make inferences about unobserved variables such as robot state [189]. These inferences often have to be made on the basis of video, which is an unfolding temporal process as opposed to static, visual representations of information. Finally, the human supervisor has to often take actions, such as operating the robot, to be able to gather information to make inferences. This is in contrast to passive inference where the images are provided to the human. While prior studies have examined joint human-AI inference on tasks that include one of these factors, a gap remains in understanding performance on tasks that integrate all three.

To address this research gap, this chapter presents a behavioral study involving 100 human participants paired with an AI-DSS in a simulated robotic task designed to encompass all three inferential challenges. In the experiment, participants teleoperate two robots with differing control delays (a latent variable) by observing video feedback (a temporal process) and taking direct action, before inferring which robot is more responsive. The participants are paired with an AI-DSS which also provides its inference. The AI-DSS inference is based on human subject data gathered separately. Both the human and the AI-DSS provide a confidence value associated to their inference and the joint inference is decided via MCS, i.e., inference made with higher confidence. The study is structured around two primary research questions.

I) How is human inference impacted by the AI-DSS inference and its confidence characteristics?

II) Does joint human-AI inference using MCS lead to better performance as compared to the AI-assisted inference setting?

To answer these questions, the analysis firstly investigates the impact of confidence alignment[1] between humans and AI on the accuracy of joint inference. We then compare the accuracy of MCS-based joint inference against individual human

[1]Confidence alignment refers to the distributional distance between the confidence levels of the human and the AI-DSS in the inference process.

Table 5.1: A description of Chapter 5's contribution to joint inference research.

Dimension	Chapter 4	Chapter 5	Chapter 6
Collaboration Type	Human-Human dyads	Human-AI dyads	Human-AI dyads
Heuristic?	Yes	Yes	No
Adaptive?	No	Partially	Yes
Bandit Approach?	None	Thompson Sampling (MAB)	Contextual & Latent Bandit
Focus of Study	Metacognitive sensitivity & heuristics in joint inference	AI-assisted decision making under uncertainty	Unified framework with context awareness

performance and AI-assisted outcomes with an baseline adaptive bandit algorithm - Thompson Sampling. In a between-subject study, we randomly vary whether participants are paired with a good metacognitive sensitivity AI-DSS or poor metacognitive sensitivity AI-DSS. We measure the impact of these experimental conditions on factors such as the participants' accuracy, whether they change their original inference in response to the AI-DSS suggestion, and the correctness of their changed inferences based on metacognitive sensitivity characteristic of the AI-DSS. Finally, we investigate how the relative metacognitive sensitivity between the human and AI-DSS impacts joint team inference performance.

This study is, to the best of our knowledge, the first to examine the MCS algorithm for human-AI joint inference in a representative robotic task which poses the challenges of latent variables, temporal processes and actions for inference. Our study yields the following findings:

I) Humans change from incorrect to correct inferences more when paired with AI-DSS that is having good metacognitive sensitivity. (Section 5.3.1)

II) The MCS-based joint human–AI inference achieved higher accuracy than both human-initiated inference after observing AI-DSS ($p < 0.001$) and inference based on a Thompson Sampling bandit ($p < 0.001$). (Section 5.3.2)

III) MCS-based joint human-AI inference accuracy with a AI-DSS with good metacognitive sensitivity was significantly higher than that achieved with a poorl metacognitive sensitivity AI-DSS ($p < 0.001$). (Section 5.3.2)

This chapter describes the work published in the International Conference on Learning Representations (ICLR) 2025, Workshop on Human-AI Coevolution.

5.2 Methodology

5.2.1 Problem Statement

Consider two agents, a human and an AI-DSS who, given an input x, produce probabilistic predictions $P_{human}(x), P_{AI}(x) \in [0, 1]$.

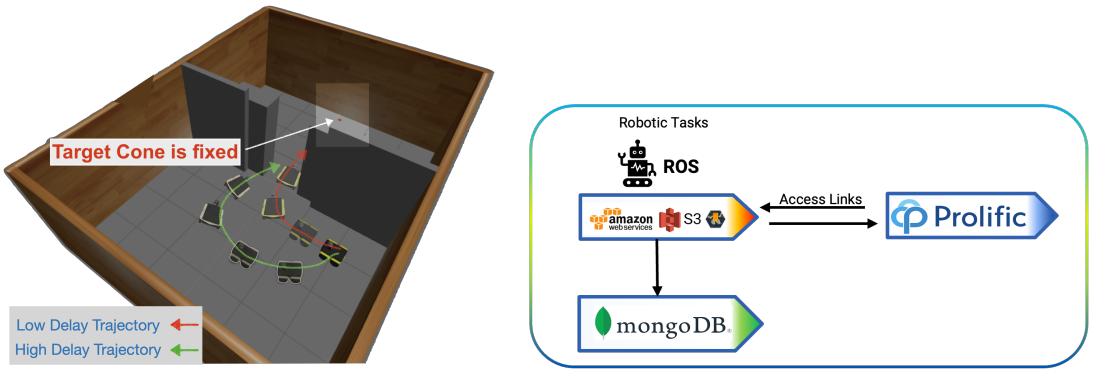
Each individual agent, i.e., the human and the AI-DSS can achieve some level of accuracy on their own. For the agent M, given their uncertainty quantification capability (metacognitive sensitivity) (Section 2.1.2) $P_M(x)$, the optimal binary classification to maximize accuracy is obtained using a threshold rule: predict 0 if $P_M(x) < 0.5$ and 1 if $P_M(x) > 0.5$.

A joint inference strategy is a way to combine the predictions $P_{human}(x), P_{AI}(x)$ into a binary classification. The MCS heuristic uses the following collaboration strategy: accept the classification of the most confident agent, i.e., the agent whose probabilistic prediction is furthest from 0.5. This collaboration strategy also achieves an accuracy, which can be compared to the accuracy of either individual agent. Human-AI complementarity is achieved when the joint inference strategy produces binary classifications that are at least as accurate than either individual agent [187].

On a representative joint inference task (detailed next), the human-AI-DSS team makes a joint inference using the MCS approach. We seek to find whether the team inference accuracy is higher than: 1) The human making the inference alone, and 2) The human making the inference in the AI-assisted setting. Further, does good metacognitive sensitivity AI-DSS lead to better team inference accuracy?

5.2.2 Environment and Task

We consider the following joint human-AI inference task. Human participants complete a set of N_{trials} binary robot selection tasks in varying environments with the aid of an AI-DSS. In each trial, human participants teleoperate two identical robots. The only difference between the two robots is the delay in the



(a) Robotic Task **(b)** Software Framework

Figure 5.1: Software Framework and User Study Scenario.
Fig. 5.1a presents the scenario with The red and green curves represent trajectories driven by a participant from our user study who was experiencing low and high delay in control of the mobile robot.
Fig. 5.1b presents the software framework for the web-based robot navigation simulator. The simulator is hosted on AWS, and data is collected and stored in a MongoDB database. The participants for the user study were recruited via the Prolific Platform.

controller, which is an unobserved latent variable. They control the robot from a start location to the goal through a narrow doorway. At the end of each trial, participants must select the robot that is more maneuverable, i.e., the one with the lower controller delay (see Fig 5.1a).

At environment reset the teleoperated robot starts in one of six possible initial positions, and the doorway has four different gap configurations, resulting in diversified 24 environment settings. The goal position, to which the human must teleoperate the robot within the time limit, remains fixed and marked by a traffic cone. Two example trajectories of the robot teleoperation task are shown in Fig. 5.1a.

The delay characteristics of each robot follows a staircase procedure. In this procedure, the task difficulty is adjusted adaptively: after two consecutive successful inferences, i.e., selecting the robot with lesser delay, the delay difference between the two robots is decreased by 20 ms, making the behaviour difference between the two robots harder to distinguish. Conversely, after an incorrect choice, the delay is increased by $\Delta_d = 20ms$, making the task easier. This adaptive adjustment is standard in human behavioral studies on confidence and ensures that the accuracy achieved by any participant is around 70%, preventing the task from being too easy

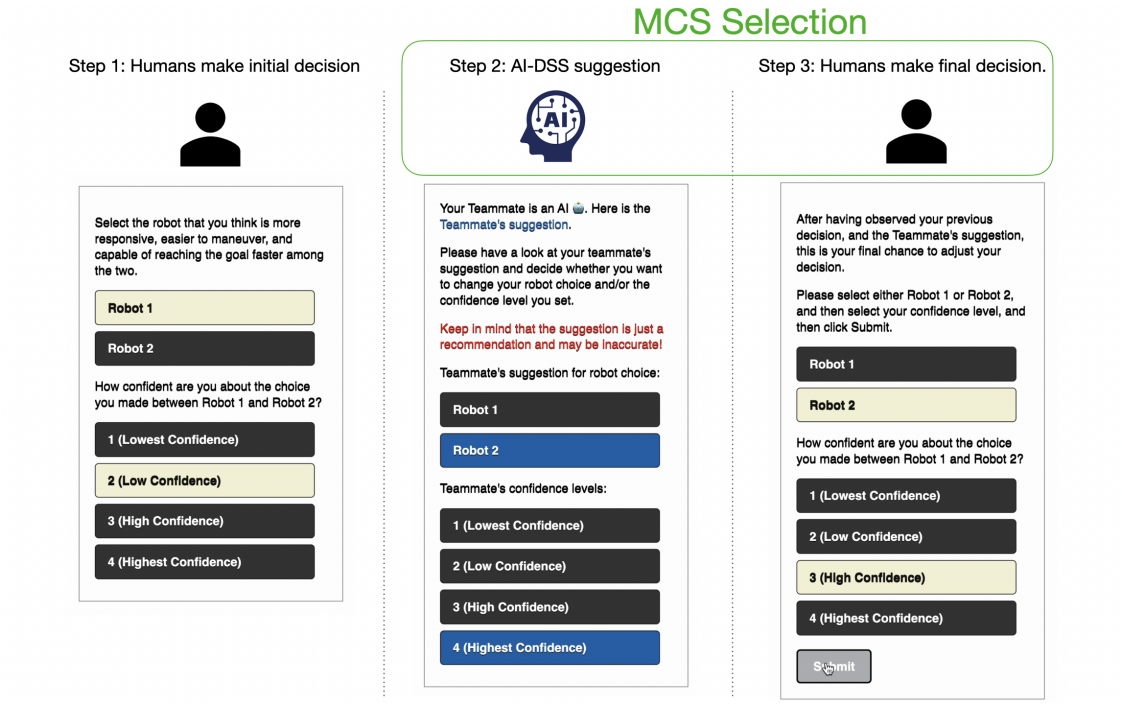


Figure 5.2: Study Interface: After teleoperating the two robots, participants make an initial inference and provide associated confidence (left panel). Then, the AI-DSS provides its inference and confidence (middle panel). If the participant decides to change their initial inference, they provide an updated inference and associated confidence (right panel). The joint human-AI inference is then decided using MCS (top right).

(resulting in ceiling performance) or too difficult (leading to performance close to random) [25, 27]. The control delay differential between the robots starts at 35 ms and progresses through five levels of task difficulty: 20 ms, 40 ms, 60 ms, 80 ms, and 100 ms differences in the robots' control delays. On average, it takes a participant 60 minutes to complete 100 trials. Gazebo was used as the simulation environment, with the Jackal robot model.

5.2.3 AI Decision Support System (AI-DSS)

To construct the AI-DSS, we collected data from 100 participants who performed the same task without the presence of an AI-DSS. To ensure a controlled comparison, the AI-DSS confidence distributions used in this stage were sampled from the empirical human data collected in the experiment described in Chapter 5.4. This allows the AI to act as a "metacognitively grounded" partner, mimicking human-like uncertainty profiles while allowing us to control its sensitivity systematically.

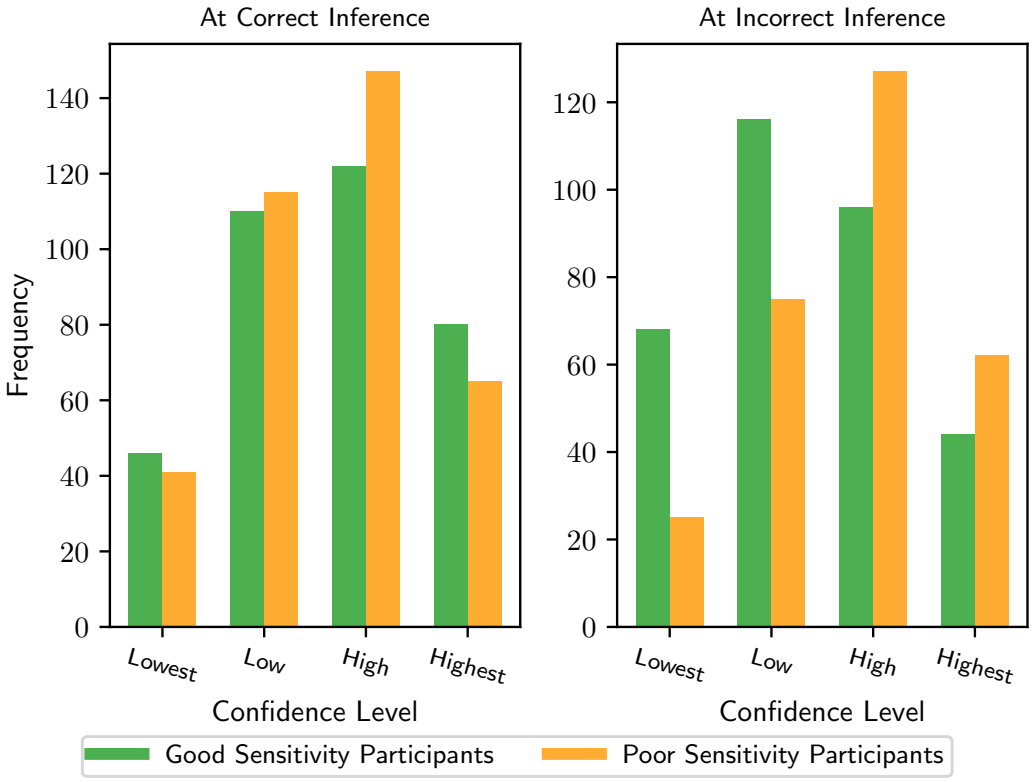


Figure 5.3: Histograms of human confidence levels associated with correct and incorrect inferences and separated by metacognitive sensitivity of the participants. The delay difference between the two robots was 60 ms in this case.

For each set of robot delay characteristics, which determine the difficulty of the task we separated the data into two sets of trials where the humans made correct and incorrect inferences, respectively. Fig. 5.3 shows an example of the discrete distribution over the human participants' confidence level associated with them making correct and incorrect inferences for a delay difference of 60 ms. During the human-AI-DSS joint inference tasks, the AI-DSS samples its confidence from these discrete distributions conditioned on the task difficulty level and accordingly provides an associated confidence score.

To study the impact of metacognitive sensitivity, we constructed both good metacognitive sensitivity and poor metacognitive sensitivity AI-DSS. We use AUROC2 as a measure of metacognitive sensitivity (Section 2.1.2). Unlike other measures, AUROC2 is a robust and bias-free metric - it remains largely unaffected by whether someone tends to give generally high or low confidence ratings [149, 190].

In our previous study, which involves teleoperating robots and inferring control differences from partial information, human participants showed a wide range of AUROC2 scores - from 0.45 to 0.85, with an average of 0.65. Based on this, we divided the dataset into two groups for AI-DSS learning: the good metacognitive sensitivity AI-DSS samples from the data of humans with metacognitive sensitivity (AUROC2 ≥ 0.65). The poor metacognitive sensitivity AI-DSS sample from the data of participants with poor metacognitive sensitivity (AUROC2 ≤ 0.55). We fixed the AI-DSS accuracy at 70% to simulate a realistic AI-based decision support system that may be imperfect.

5.2.4 Experimental Procedure

We conducted an online study, approved by the University Research Ethics Committee for data collection. Responses from 100 participants in total were collected via the Prolific online platform[2]. Each participant performed 100 trials.

First, the participants received 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 the keyboard controller. Next, they performed the 100 trials. In each trial, they teleoperated a pair of robots, denoted $\mathcal{R} = \{R_A, R_B\}$. After each trial, they made an initial inference, i.e., they selected the robot they deemed having the lower control delay $I_{\mathrm{H}} \in \{R_A, R_B\}$, and provide their confidence scores $c_H \in$ {Lowest, Low, High, Highest}. The AI-DSS, designed as described in Section 5.2.3, then provided the human participant with its inference $I_{AI} \in \{R_A, R_B\}$ along with its confidence $c_{AI} \in$ {Lowest, Low, High, Highest}. The participants were then asked to make a final inference after having observed the AI-DSS inference and associated confidence. The experimental flow and study interface is shown in Fig. 5.2.

After data collection, we excluded participants who were either unable to perform the task or were inattentive, based on the following criteria: (i) those with an accuracy of less than 65%, and (ii) those who gave the same confidence

[2]https://www.prolific.com/

rating response for over 95 of 100 trials. Inattentive participants could achieve above-chance performance by guessing, so setting the threshold at 65% excluded them while retaining engaged participants. This is standard procedure in studies on human self-confidence [25]. Out of the 100 participants, 80 met the above criteria.

5.2.5 Human-AI Joint Inference

In this work, the central methodological contribution is the systematic evaluation of different strategies for arbitrating between the human's inference (I_H) and the AI's inference (I_{AI}) to produce the final team decision, I_{Team}.

Maximum Confidence Slating (MCS)

The first strategy is a non-adaptive heuristic. MCS is a well-established and powerful baseline from the decision-making literature (see Section 2.1.2) [191]. The rule is simple: the team's final decision is the decision of the agent who expressed higher confidence.

$$I_{\text{Team}} = \begin{cases} I_H & \text{if } c_H > c_{AI} \\ I_{AI} & \text{if } c_{AI} > c_H \\ \text{random}(I_H, I_{AI}) & \text{if } c_H = c_{AI} \end{cases} \tag{5.1}$$

Thompson Sampling (TS)

In the comparison between heuristic with algorithmic joint inference, we implement adaptive learning strategy using a bandit approach: Thompson Sampling (TS), also known as posterior sampling. The TS bandit makes decisions by sampling from a posterior distribution. The core principle is to choose an action according to its posterior probability of being the optimal action, given the data observed thus far. This is achieved through a two-step process at each time step:

1. **Sample from the Posterior:** For each arm, draw one random sample of its unknown reward parameter from its current posterior distribution.

2. **Act Greedily on the Sample:** Select the arm whose sampled parameter is the highest.

The posterior distribution for each arm's (agent) parameter reflects the current belief, incorporating both the estimated value (the mean of the posterior) and the uncertainty (the variance of the posterior). The reward of making correct, incorrect inference in our setting is binary ($R_t \in \{0, 1\}$).

A canonical application of Thompson Sampling for our binary reward setting is the Beta-Bernoulli Model. The Beta distribution is used as the conjugate prior for the Bernoulli likelihood. The Beta distribution, $\text{Beta}(\alpha, \beta)$, is defined on $(0, 1)$, and its PDF is:

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} x^{\alpha-1}(1 - x)^{\beta-1}$$

In this context, the parameters α and β have a clear meaning for each agent:

- α is intuitively interpreted as the total number of times the agent's inference was correct (successes) + 1.
- β is intuitively interpreted as the total number of times the agent's inference was incorrect (failures) + 1.

The complete Thompson Sampling algorithm using this model is summarized below, adapted to use S_i (Successes) and F_i (Failures) for $i \in \{H, AI\}$.

Algorithm 4 Thompson Sampling (Beta-Bernoulli) for Joint Inference

Input: Agent Set $A = \{A_H, A_{AI}\}$, Time horizon T

for $i \in \{H, AI\}$ **do**
    $S_i \leftarrow 0$
    $F_i \leftarrow 0$
for $t = 1$ **to** T **do**
    for $i \in \{H, AI\}$ **do**
        Sample $\theta_i(t) \sim \text{Beta}(S_i + 1, F_i + 1)$
    $a_t \leftarrow \arg\max_{i \in \{H,AI\}} \theta_i(t)$
    $I_{\text{Team}} \leftarrow I_{a_t}$ ▷ *The team decision is the selected agent's inference*
    Observe binary reward $R_t \in \{0, 1\}$ from the chosen agent a_t's inference
    if $R_t = 1$ **then**
        $S_{a_t} \leftarrow S_{a_t} + 1$
    else
        $F_{a_t} \leftarrow F_{a_t} + 1$

5.2.6 Performance Metrics

The primary evaluation metric is Joint Team Accuracy, defined as the percentage of trials where the final team inference, I_{Team}, correctly identified the robot with the lower control delay. Performance is reported both overall and by task difficulty level. Paired t-tests are used to assess the statistical significance of accuracy differences between strategies. To evaluate arbitration, our analysis focuses on the most challenging cases: trials where the human and AI provided conflicting inferences.

5.3 Results

We report our findings on: a) the influence of AI-DSS performance and confidence on human inference and confidence, and b) joint inference accuracy when MCS is used to determine the final inference by selecting between the human and AI-DSS inference according to higher confidence.

5.3.1 Human Inference Dynamics

We first investigate the frequency with which humans change their initial inference based on the AI-DSS confidence level. In other words, does the confidence score associated with the AI-DSS inference in each trial impact humans' propensity to change their original inference of which robot they thought had a lesser delay?

Fig. 5.4 shows that humans exhibited the strongest tendency to revise their initial inferences when presented with AI-DSS recommendations accompanied by the *highest confidence levels* (14.9% of trials). On the other hand, when the AI-DSS confidence level was low, the percentage of trials where the human participants changed their inference dropped to 3.8%.

However, did the humans change their inferences to correct ones or did they end up changing from an initial correct inferences to an incorrect one? We defined ***positive changes*** as instances where humans initially made incorrect inferences but, after observing the AI-DSS inference, revised their original choice to a correct inference. Conversely, ***negative changes*** occurred when humans initially made

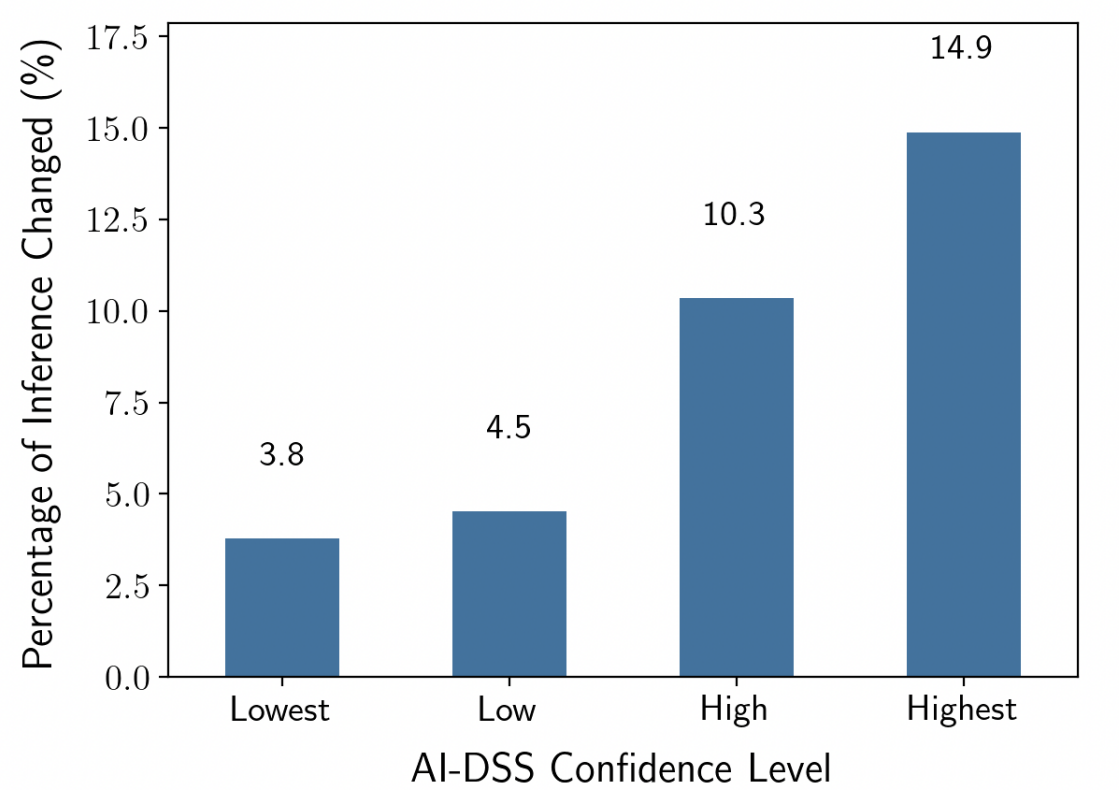


Figure 5.4: Higher confidence scores from the AI-DSS result in a higher number of inferences changed by humans.

correct inferences but changed to incorrect inferences after observing the AI-DSS inference. The number of each type of change was counted against the total number of changes to calculate the percentage of each type. We analysed how the percentage of *positive* and *negative* inferences changes were influenced by: 1) the difference in confidence level between humans and AI-DSS, and 2) the metacognitive sensitivity of the AI-DSS.

Fig. 5.5 shows the percentage of positive and negative changes versus the difference in confidence level between humans and AI-DSS. We denote 'Human less confident' as the case where the human's confidence level is 2 or more levels lower than the AI-DSS's. Similarly, we denote 'Human more confident' as the case where the human's confidence level is 2 or more levels higher than the AI-DSS's. The analysis revealed a statistically significant effect of confidence level disparity on the percentage of inference changes. Specifically, in the 'Human less confident' case, humans exhibited a higher percentage of positive inference

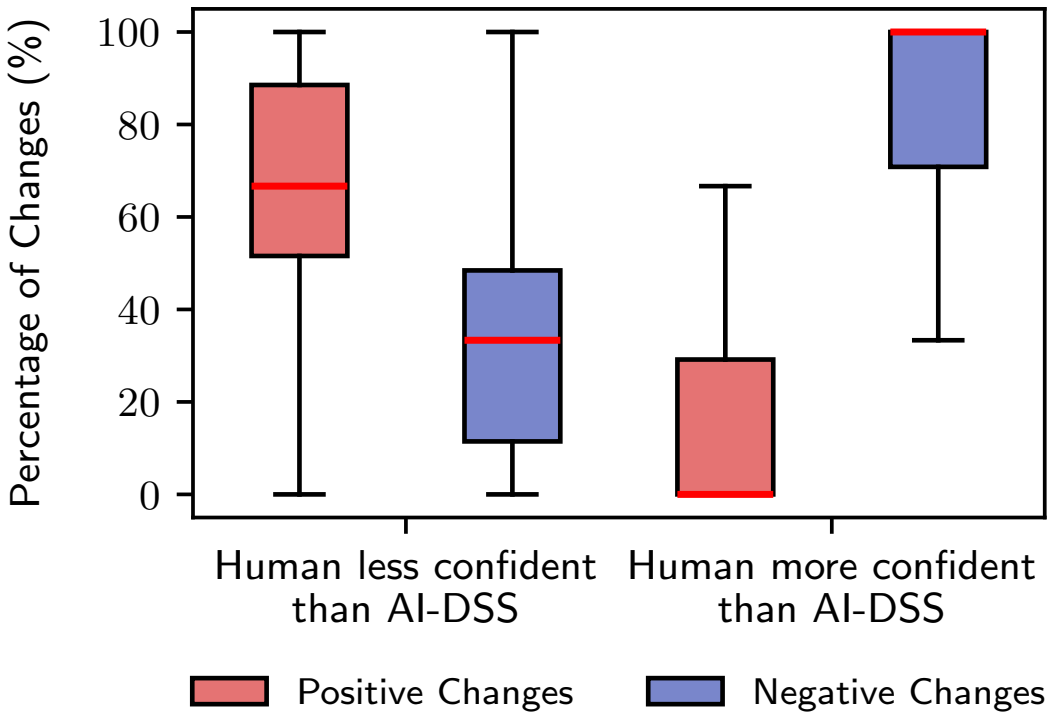


Figure 5.5: Dynamics of human inferences: Difference in *confidence levels* between humans and AI-DSS influences whether humans change from incorrect to correct inferences.

changes ($t(66) = 7.21, p < 0.001$). And in the 'Human more confident' case, humans exhibited a higher percentage of negative inference changes ($t(18) = 5.81, p < 0.001$).

Fig. 5.6 shows the mean percentage of positive and negative inference changes according to the metacognitive sensitivity of the AI-DSS. The analysis reveals that good metacognitive sensitivity AI-DSS systems facilitated a significantly higher percentage of positive inference changes ($42.53\% \pm 6.19\%$) vs poor metacognitive sensitivity AI-DSS ($28.41\% \pm 4.64\%$), whereas poor metacognitive sensitivity AI systems induced negative inference changes at a significantly higher percentage ($25.58\% \pm 4.46\%$) vs good metacognitive sensitivity AI-DSS ($10.95\% \pm 2.50\%$).

Next, we investigated how human confidence is impacted by the metacognitive sensitivity of the AI-DSS. Fig. 5.7 summarizes the dynamics between human confidence shifts and inference correctness across different AI metacognitive sensitivity types. Recall that both human and AI-DSS confidence were provided on a 4-level Likert scale: Highest Confidence (4), High Confidence (3), Low Confidence

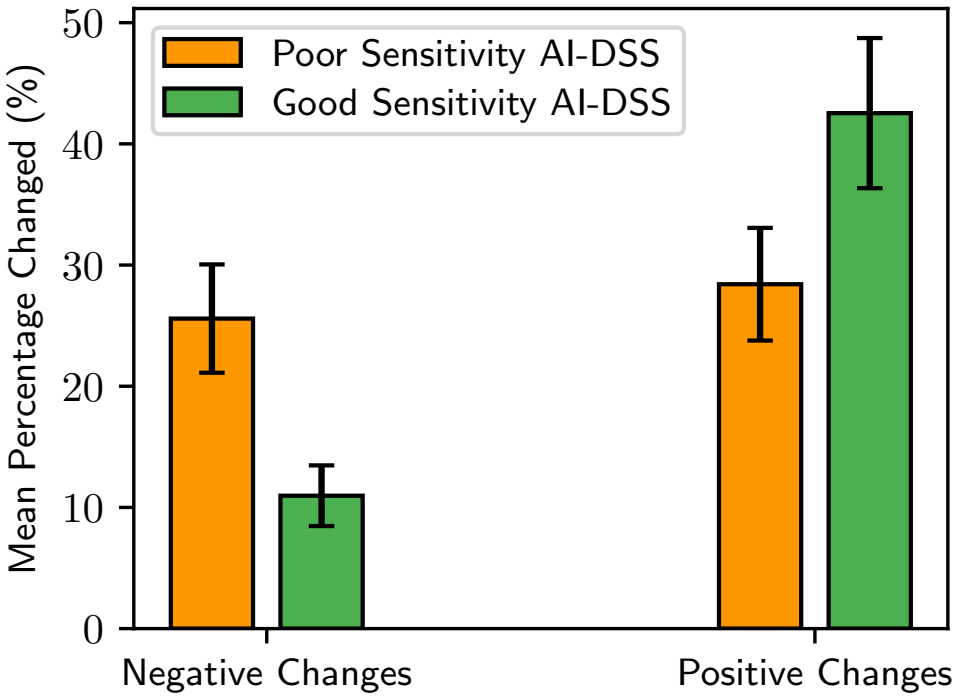


Figure 5.6: Dynamics of human inferences: AI-DSS metacognitive sensitivity influences whether humans change from incorrect to correct inferences.

(2), and Lowest Confidence (1). Confidence shift was quantified as the difference between human confidence levels after and before observing the AI-DSS inference. The results show the following: interaction with AI-DSS of good metacognitive sensitivity is associated with a greater increase in human confidence following correct inferences (green vs orange: Right side boxplots $t(104) = 3.53, p < 0.001$), whereas interaction with poor metacognitive sensitivity AI-DSS is associated with a greater increase in human confidence following incorrect inferences (green vs orange: Left side boxplots $t(71) = -2.01, p < 0.05$).

We further plotted the regression results in Figure 5.8 showing trends in human confidence changes when interacting with good sensitivity and poor sensitivity AI-DSS systems. Left panel – *Good Sensitivity AI*: Participants' confidence increased significantly after making a correct choice. The upward slope of the regression line indicates that confidence boosts were appropriately stronger for correct outcomes than for incorrect ones. This suggests that the good sensitivity AI's feedback helped calibrate users' confidence in a way that reflected their actual performance.

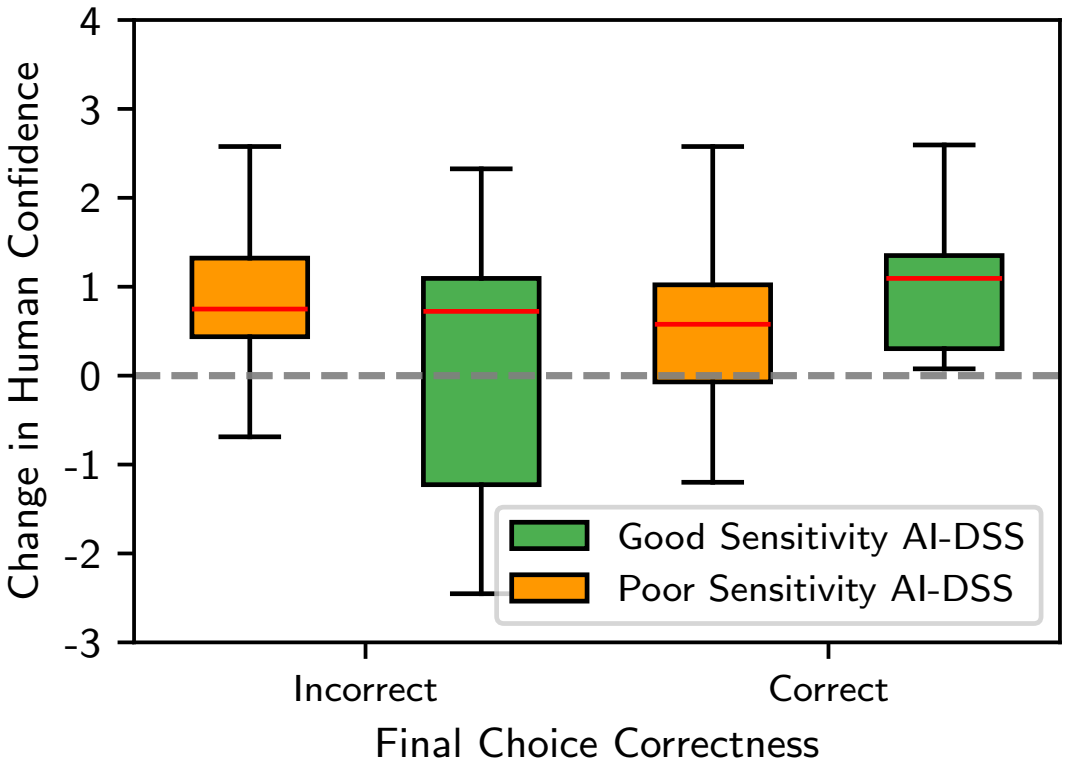


Figure 5.7: Dynamics of human confidence: AI *metacognitive sensitivity* influence on human confidence changes and joint inference correctness.

Right panel – *Poor Sensitivity AI*: Participants' confidence increased on average even after incorrect choices, and this increase was slightly greater than for correct ones. The slight downward slope of the regression line shows that there was no positive reinforcement for being correct. In fact, participants experienced a slightly larger average confidence boost when they were wrong. This suggests that the poor sensitivity AI's feedback may have been misleading, resulting in poorly calibrated or even misplaced user confidence.

We also analyze the relationship between confidence alignment and human-initiative inference performance. The Jensen-Shannon divergence (JS) was used to quantify distributional differences between human and AI-DSS confidence levels. Higher JS divergence values indicate greater dissimilarity and, consequently, increased misalignment between human and AI-DSS confidence. For each of the 80 participants, we calculated JS divergence of human confidence's distribution

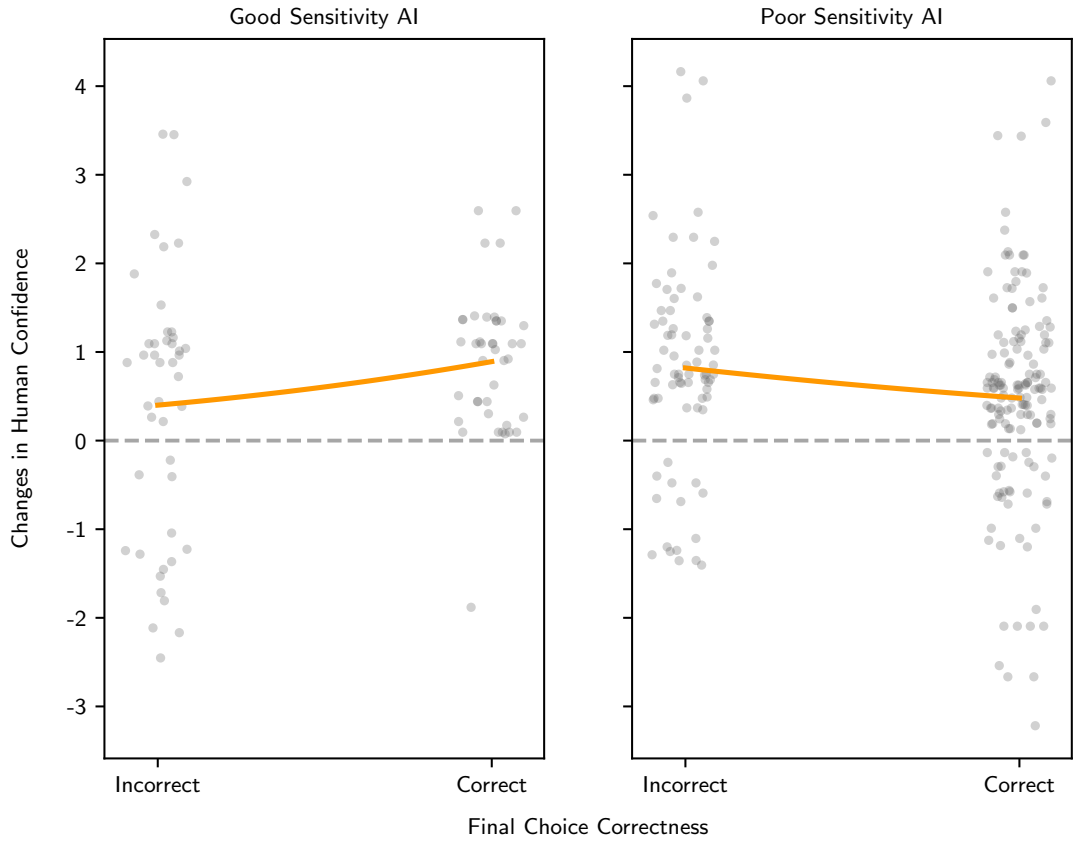


Figure 5.8: Dynamics of human confidence: AI *metacognitive sensitivity* influence on human confidence changes and joint inference correctness.

from AI-DSS confidence distribution across all task difficulty levels (defined by the delay differential between the two robot controllers as mentioned previously) and then averaged the value for the five levels of difficulty to compute the averaged JS divergence. This average distributional differences were then plotted against joint human-initiative inference accuracy, with the aggregate data fitted using linear regression Fig. 5.9 revealed statistically significant relationships ($r^2 = 0.183, p < 0.001$) demonstrating that greater divergence in confidence alignment between humans and AI-DSS predicts lower joint human-initiative inference accuracy.

5.3.2 Joint Inference Performance

We present our benchmark results on joint human-AI-DSS inference performance. We compare the performance of Maximum Confidence Slating (MCS) against human-initiative inference with AI-DSS support as well as against individual human

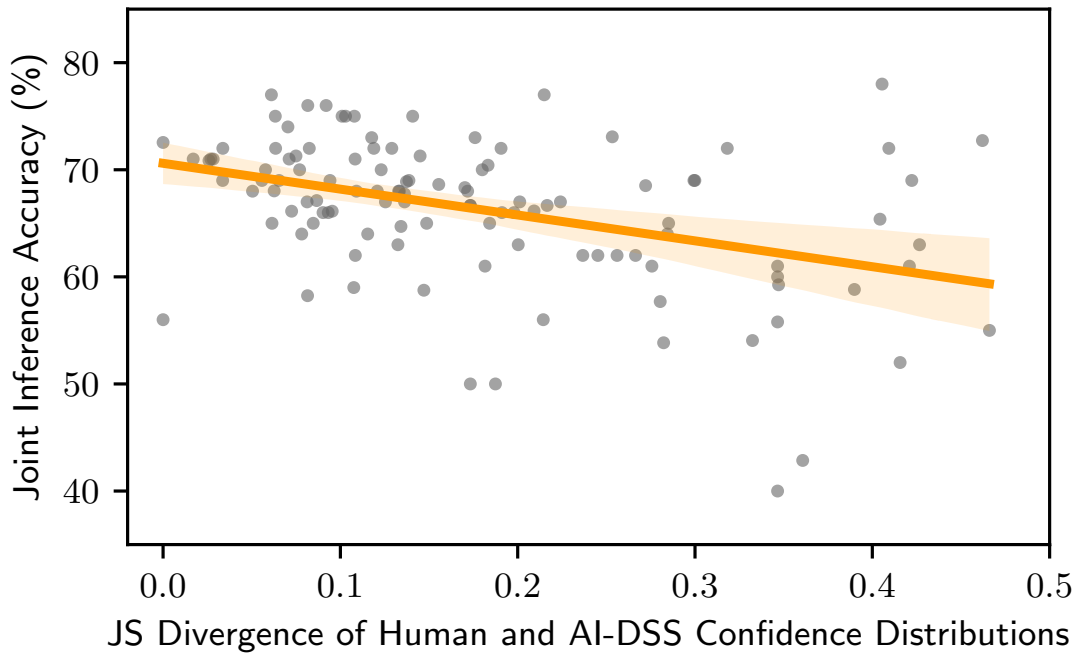


Figure 5.9: Misalignment in human and AI confidence ratings leads to degradation in accuracy of human-initiative joint inference.

inference without AI-DSS support. The performance was measured using accuracy, i.e., the proportion of correct inferences made relative to the total number of inference trials.

Fig. 5.10 presents results from aggregated data of 80 human participants interacting with AI-DSS of both good and poor metacognitive sensitivity. These results focus on cases where the human and AI-DSS provided opposing inferences that required resolution to produce a final joint team decision (e.g. Human chose Robot 1, while AI-DSS suggested Robot 2). '*Joint Inference by Human*' represents cases the final inference is made by the human participant (i.e., human-initiative). '*Joint Inference by MCS*' refers to cases where the final inference is determined by the MCS. '*Joint Inference by TS Bandits*' refers to the case where a Thompson Sampling bandit algorithm chooses the final inference, selecting between the human's or AI-DSS's choice. We observe that human participants improved their inference accuracy after receiving AI-DSS recommendations compared to their initial individual decisions. However the MCS-based joint inference significantly

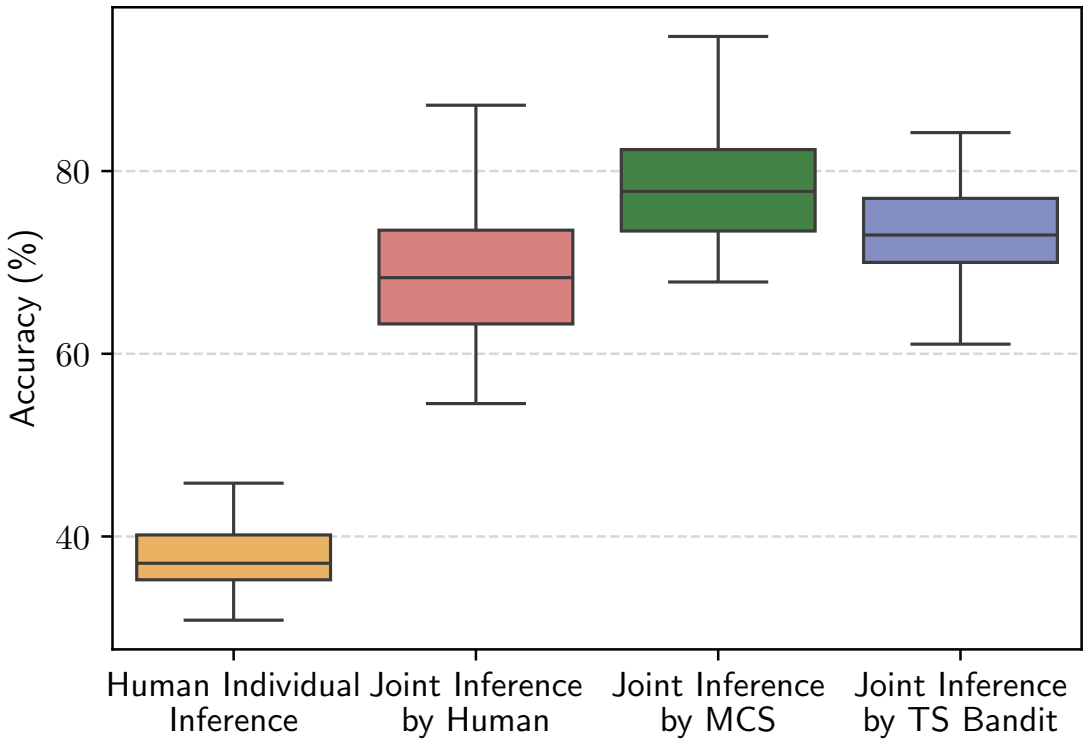


Figure 5.10: Joint human-AI inference accuracy is highest with MCS, outperforming purely human individual inference, joint inference dictated by human, and approaches using Thompson Sampling (TS) bandit.

outperformed both human-initiated joint inference ($t(4828) = 8.52$, $p < 0.001$) and Thompson Sampling bandit approach ($t(4828) = 4.74$, $p < 0.001$). All t-tests in our results are 2-sided paired.

This finding highlights that relying solely on human inference is suboptimal. Instead, MCS provides a simple yet robust method for enhancing team inference by leveraging a readily available metric - confidence.

We further analyzed the effect of AI-DSS both good and poor metacognitive sensitivity on the performance of MCS-based joint human-AI inference. Participants were paired with either a good both good and poor metacognitive sensitivity AI-DSS (AUROC2 ≥ 0.65; 42 participants) or a poor metacognitive sensitivity AI-DSS (AUROC2 ≤ 0.55; 38 participants).

Fig. 5.11 shows joint inference accuracy for both the MCS-based and human-initiated approaches, separated by the AI-DSS metacognitive sensitivity condition. When paired with a good metacognitive sensitivity AI-DSS, the MCS-based joint

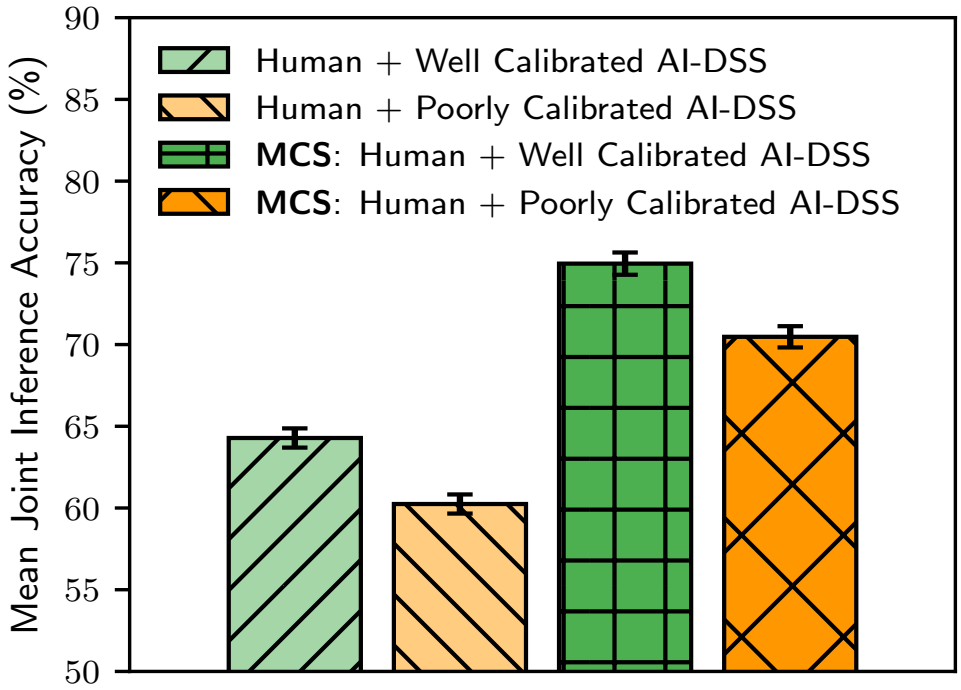


Figure 5.11: Influence of varying AI type on team inference accuracy.

inference achieved an accuracy of $(64.28\% \pm 0.59\%)$, significantly outperforming the human-initiated joint inference, which reached $(60.24\% \pm 0.49\%)$. Moreover, MCS performance was also sensitive to the metacognitive sensitivity quality of the AI-DSS. MCS-based joint inference with a good metacognitive sensitivity AI-DSS yielded $(74.95\% \pm 0.68\%)$ accuracy, which was significantly higher than the $(70.47\% \pm 0.65\%)$ achieved when paired with a poor metacognitive sensitivity AI-DSS. These findings demonstrate two key insights:

1. MCS consistently outperforms human-initiated resolution, confirming the value of confidence-based joint inference.

2. The effectiveness of MCS is amplified when paired with a good metacognitive sensitivity AI-DSS, highlighting the critical role of AI-DSS's metacognitive sensitivity in human–AI team performance.

To better understand why MCS did not perform well when paired with poor metacognitive sensitivity AI-DSS, we identified all trials where the AI-DSS are of

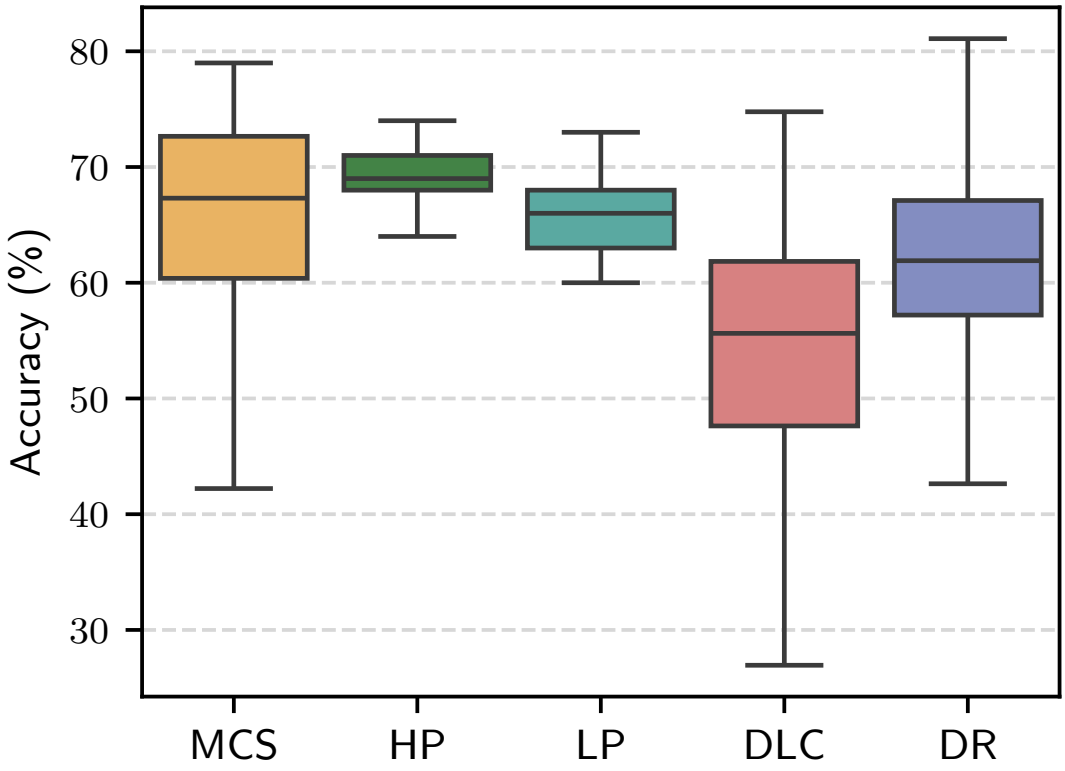


Figure 5.12: Inference accuracy obtained by pairing human participants with the poor metacognitive sensitivity AI-DSS.

poor metacognitive sensitivity. We created a virtual pairing of the poor metacognitive sensitivity AI-DSS with data from all human participants who operated under the same robot control delay differential. This virtual pairing allowed us to compute joint inference accuracy across consistent task difficulty levels.

Within each virtual human-AI dyad, we investigated three different joint inference strategies: MCS, Dummy Low Confidence (DLC), which selects the low-confidence inference, and Dummy Random (DR), which randomly selects between the human and AI-DSS inferences. We also included two baselines: High-Performing (HP), the individual in the dyad with a higher percentage of correct responses across trials, and Low-Performing (LP), the less accurate individual. Fig. 5.12 presents the accuracy of these joint inference strategies. HP slightly outperformed MCS ($t(1716) = -1.15, p < 0.001$). This outcome reflects an important limitation of MCS. When MCS is applied with poor metacognitive sensitivity AI-DSS, it selects

the AI-DSS inference due to its inappropriately high confidence, even when the AI-DSS is incorrect. Thus, MCS overrides the human's correct judgment. This result underscores the importance of metacognitive sensitivity for the effective functioning of confidence-based methods such as MCS.

5.4 Discussions

Summary

In Chapter 5.4, we investigate how AI-DSS inference and confidence impacted human inference and confidence and whether joint human-AI inference using MCS lead to better performance as compared to the AI-assisted inference setting. Our results give important insights into human making joint inference with an AI-DSS.

First, the metacognitive sensitivity of the AI-DSS impacted the nature of the human inference change (Fig. 5.6). A AI-DSS of good metacognitive sensitivity led to humans making more positive changes, i.e., changing from an initially incorrect inference to a correct one (Fig. 5.7). Second, misalignment in confidence between the human and the AI-DSS led to reduced inference accuracy when the human made the final inference after having observed the AI-DSS inference (Fig. 5.9) This finding underscores that while the provision of confidence scores alongside AI-DSS inferences can be beneficial, the degree to which this information improves joint human-AI inference is critically dependent on the underlying metacognitive sensitivity of AI-DSS.

Second, joint human-AI inference accuracy was highest with MCS, outperforming both AI-assisted inference, purely human inference and the Thompson sampling bandit (Fig. 5.10). MCS with good metacognitive sensitivity AI-DSS yielded better joint inference outcomes as compared to both MCS with poor metacognitive sensitivity AI-DSS and human final inference with poor metacognitive sensitivity AI-DSS (Fig. 5.11). Third, we identified the limitation (poor metacognitive sensitivity AI-DSS) that prevents the heuristic method such as MCS from achieving fully complementary performance (Fig. 5.12). These findings show the potential and dependency of MCS-based joint human-AI inference towards achieving complementarity.

Limitations and Research Directions

While the proposed approach and designed user study provides a valuable first step for confidence-guided human-AI joint inference in robot missions, there are several limitations. First, our study assumes that both the AI-DSS and the human receive the same observable information. However, in real-world situations, they may have different sources and amounts of information leading to the dependence of confidence on context. Second, we highlight the limitations of a static arbitration rule like MCS, particularly when agent metacognitive sensitivity varies. This motivates a move towards a more dynamic and context-aware approach.

In Chapter 6.4, we address this by introducing an adaptive framework for joint inference. We will investigate whether an adaptive system that learns to arbitrate between agents can outperform static heuristics. This system considers not just confidence, but also task difficulty and other metacognitive traits. We further consider the scenarios where key contextual variables, such as task difficulty, are unobservable.

6

A Contextual Bandit Framework for Adaptive Joint Inference

Contents

6.1 Introduction 95
6.2 Methodology 98
6.2.1 The Human &AI-DSS Dyad 99
6.2.2 Human-Subject Experiment 100
6.2.3 Virtual Dyad Pairings 101
6.2.4 Performance Metrics 102
6.2.5 Methods for Human-AI Joint Inference 102
6.3 Results 109
6.3.1 Adaptive Joint Inference Outperforms Heuristics 109
6.3.2 Performance Across Various Difficulty Conditions 110
6.3.3 Latent Variable Modeling of Unobservable Context . . . 112
6.4 Discussion 112

6.1 Introduction

Building on the analysis of static joint inference heuristics in Chapters 5.4 in this chapter we introduce an adaptive framework designed to handle the dynamic uncertainties inherent in human-robot collaboration. As established in Section 3.2.1 and Section 3.2.2, robots and humans are increasingly deployed together in high-stakes environments like search and rescue or space exploration, where system

failures are not always immediately visible [2–9]. In such settings, delays, errors, and other latent faults can disrupt coordination, making effective human-robot collaboration difficult. In search and rescue operations or nuclear surveillance, environmental changes can degrade sensor reliability without obvious indicators [10, 11]. Medical robotic systems face patient-specific variations that create diagnostic uncertainty [12–14]. These scenarios share a common thread: the need for real-time decision-making under uncertainty, where neither human nor agent has perfect information. Unlike static AI applications, robotic systems operate in physical environments with temporal constraints and safety implications that amplify the consequences of poor decisions. The spatial and temporal dynamics inherent to robotic tasks create compound uncertainty that traditional human-AI collaboration methods struggle to address effectively [15]. Addressing these challenges requires methods that adaptively combine human and AI decision-making. A key step toward this goal is joint inference, where human and AI inference are combined.

Existing approaches to this problem are often insufficient as they typically rely on static heuristics. Methods such as simple voting, weighted averaging, or the Maximum Confidence Slating (MCS) approach discussed in previous chapters, operate on fixed rules that assume stable agent capabilities. A confidence-only rule like MCS, for instance, overlooks crucial context, such as the immediate task difficulty or the distinct metacognitive traits of the agents. One partner might be confident yet poorly calibrated, while another might be highly sensitive but cautious - nuances to which fixed heuristics cannot adapt. This inflexibility renders such methods brittle, particularly when faced with latent faults that degrade performance in non-obvious ways [15, 81, 102, 192]. To address this, we introduce an adaptive framework that learns to dynamically select the more reliable partner by considering not just their confidence, but also the task context and their historical metacognitive performance. This is achieved by leveraging contextual bandits, a reinforcement learning approach well-suited to this challenge because it can balance exploration and exploitation in non-stationary environments.

Table 6.1: A description of Chapter 6's contribution to joint inference research.

Dimension	Chapter 4	Chapter 5	Chapter 6
Collaboration Type	Human-Human dyads	Human-AI dyads	Human-AI dyads
Heuristic?	Yes	Yes	No
Adaptive?	No	Partially	Yes
Bandit Approach?	None	Thompson Sampling (MAB)	Contextual & Latent Bandit
Focus of Study	Metacognitive sensitivity & heuristics in joint inference	AI-assisted decision making under uncertainty	Unified framework with context awareness

To ground this investigation in a tangible problem, we focus on a representative latent fault: differential control delay in robot teleoperation. This fault, which manifests as a subtle difference in responsiveness between two robots, presents a diagnostic challenge that mirrors real-world issues in distributed systems where network conditions or processing loads create hidden performance asymmetries [16, 17]. We formulate the challenge of arbitrating between the human and AI partners as a bandit problem, where the action is to select either the human's inference or the AI's inference. The reward is a binary signal based on the correctness of the resulting team decision. This framework allows for a systematic evaluation of different arbitration policies by progressively enriching the context vector used for decision-making. The analysis begins with a policy that considers only agent confidence, then incorporates agent-specific metacognitive traits including metacognitive sensitivity (AUROC2) (see Section 2.1.2) and calibration (ECE) (see Section 2.1.2), and finally adds the inferred-unobservable task features (difficulty).

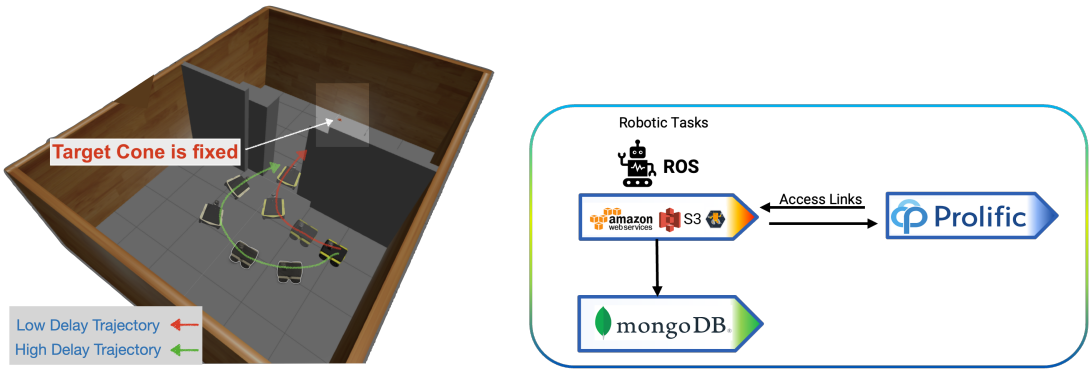
This framework allows us to investigate the following research questions:

RQ0: Does adaptive joint inference via contextual bandits outperform static confidence-based heuristics such as Maximum Confidence Slating?

RQ1: Does augmenting a confidence-based policy with observable task difficulty improve joint inference accuracy?

RQ2: Does adding agent-specific metacognitive signals (medium-term traits) provide gains to joint inference accuracy?

RQ3: How robust are these gains when external task features are unobservable and be modelled as latent variables?



(a) Robotic Task **(b)** Software Framework

Figure 6.1: Software Framework and User Study Scenario.
Fig. 6.1a presents the scenario with The red and green curves represent trajectories driven by a participant from our user study who was experiencing low and high delay in control of the mobile robot.
Fig. 6.1b presents the software framework for the web-based robot navigation simulator. The simulator is hosted on AWS, and data is collected and stored in a MongoDB database. The participants for the user study were recruited via the Prolific Platform.

The contributions of this work include:

- We present the first application of contextual bandits as an adaptive joint inference policy for a dynamic, spatiotemporal human-robot diagnostic task, with accuracy improvements up-to 20% over static heuristics. (Section 6.3.1)
- We demonstrate empirically that an adaptive policy incorporating metacognitive features yields robust significant accuracy gains across diverse difficulty conditions. (Section 6.3.2)
- We introduce and validate a Latent Variable Bandit as potential method for learning an effective joint inference policy when task difficulty is an unobservable, causally confounding variable, providing a framework applicable to broader classes of robotic deployment scenarios. (Section 6.3.3)

6.2 Methodology

The core task is formulated as a sequential joint human-AI inference problem. In each trial, a human operator teleoperates a pair of robots, denoted $\mathcal{R} = \{R_A, R_B\}$, through a standardized task environment (e.g., navigating a narrow doorway). The

robots are kinematically identical, but each robot $R_i \in \mathcal{R}$ is subject to a control delay, δ_i, drawn from a predefined set of possible delay values. These delays, δ_A and δ_B, are latent variables, unobservable to the operator or the AI system. The objective of the human-AI team is to make a joint inference, $I_{\text{Team}} \in \{R_A, R_B\}$, that correctly classifies which robot has the lower control delay. This experimental framework provides the basis for the post-hoc analysis of joint inference performance. The simulation scenario is depicted in Fig. 6.1a.

6.2.1 The Human & AI-DSS Dyad

The collaborative team consists of two primary agents whose inputs must be integrated.

The Human Operator (H)

The human operator interacts directly with the robots via a teleoperation interface. After performing the task, the operator provides two pieces of information:

1. An initial inference, $I_H \in \{R_A, R_B\}$, representing their judgment about which robot had lower control delay.

2. A confidence rating corresponding to their subjective certainty in that judgment, provided on a four-point Likert scale, $c_H \in \{\text{Lowest, Low, High, Highest}\}$.

The AI-Decision Support System (AI-DSS)

The AI-DSS is designed to act as a collaborative partner, providing its own inference and confidence. Its construction is data-driven, based on a preliminary study with 100 human participants performing the same task without AI assistance. For each level of task difficulty (defined by the difference in delay between the robots), the system learns two discrete probability distributions over the four confidence levels: one for trials where humans made a correct inference and one for trials where they made an incorrect inference. During a joint task, the AI-DSS is provided with the ground-truth task difficulty and samples its inference ($I_{AI} \in \{R_A, R_B\}$) and

confidence ($c_{AI} \in$ {Lowest, Low, High, Highest}) from these learned distributions to simulate a realistic, imperfect partner.

6.2.2 Human-Subject Experiment

The human-AI interaction data collection was conducted through a human-subject experiment approved by the University Research Ethics Committee involving 100 participants. The experiment was performed in a simulated environment where each participant was tasked with teleoperating two identical mobile robots. In each trial, the participant's goal was to navigate the pair of robots from a starting location through a narrow doorway to a goal location.

The core experimental manipulation was the introduction of different control delays for the two robots. The difficulty of the task - discerning which robot had lower control delay - was systematically varied across four levels (Low:1, Medium:2, High:3, Extreme:4), defined by the magnitude of the difference in control delay. These levels were established based on baseline human performance data collected prior to the main experiment. The trial structure proceeded as follows:

1. The participant completed the teleoperation task for a single trial.

2. The participant made an initial inference (I_H) and provided a confidence rating (c_H).

3. The AI-DSS presented its inference (I_{AI}) and confidence (c_{AI}) to human participant.

4. The participant was then asked to make a final inference based on both sets of inferences and confidence ratings.

On average, it takes a participant 60 minutes to complete 100 trials. Gazebo was used as the simulation environment, with the Jackal robot as the model.

6.2.3 Virtual Dyad Pairings

To obtain a large and varied dataset to train and robustly evaluate our adaptive algorithms, we employed a *virtual dyad pairing* methodology, a well-established post-hoc data augmentation strategy adapted from research in human-human collaboration [57]. This allow us to expand our dataset of 80 qualified participants into a large-scale simulated interaction set for algorithmic evaluation.

The augmentation process was executed as follows:

- Decoupling of Data Streams: First, the collected data was segregated into two distinct pools: a human-only data pool containing the inference (I_H) and confidence (c_H) from each of the 80 participants across their 100 trials, and a corresponding AI-DSS data pool with its inferences (I_{AI}) and confidence ratings (c_{AI}). This resulted in 80 independent sets of human data and 80 sets of AI-DSS data.

- Combinatorial Pairing: We then systematically created virtual dyads by pairing the trial data from each human participant with the data from every AI-DSS agent, resulting in 3,160 unique virtual pairs (80 * 79 / 2).

- Task-Contingent Matching: to ensure the validity of these pairings, trials were matched based on the ground-truth differential control delay (δ_i). A human's judgment on a trial with a 20ms delay difference was only paired with an AI's judgment on a 20ms trial. This critical step ensures that each virtual pair's decisions are grounded in an identical, well-defined stimulus.

This task-contingent matching procedure generated a dataset where each virtual dyad contained approximately 1,000 paired interaction instances, distributed across the four predefined difficulty levels. The resulting augmented dataset was sufficiently large and varied to robustly train and evaluate the performance of both the baseline MCS heuristic and the proposed adaptive bandit algorithms, which are detailed in the subsequent section.

6.2.4 Performance Metrics

The primary evaluation metric is Joint Team Accuracy, defined as the percentage of trials where the final team inference, I_{Team}, correctly identified the robot with the lower control delay. Performance is reported both overall and by task difficulty level. Paired t-tests are used to assess the statistical significance of accuracy differences between strategies. To evaluate arbitration, our analysis focuses on the most challenging cases: trials where the human and AI provided conflicting inferences.

6.2.5 Methods for Human-AI Joint Inference

In this work, the central methodological contribution is the systematic evaluation of different strategies for arbitrating between the human's inference (I_H) and the AI's inference (I_{AI}) to produce the final team decision, I_{Team}. The strategies are presented in order of increasing sophistication.

Maximum Confidence Slating (MCS) - A Static Baseline

The first strategy is a non-adaptive heuristic. MCS is a well-established and powerful baseline from the decision-making literature [19, 54, 57, 191]. The rule is simple: the team's final decision is the decision of the agent who expressed higher confidence.

$$I_{\text{Team}} = \begin{cases} I_H & \text{if } c_H > c_{AI} \\ I_{AI} & \text{if } c_{AI} > c_H \\ \text{random}(I_H, I_{AI}) & \text{if } c_H = c_{AI} \end{cases} \tag{6.1}$$

Contextual Bandit with Enriched Context

In the first extension, we introduce adaptive learning using a contextual bandit approach [193]. The problem of deciding whether to select the human or the AI inference is formulated as learning an optimal policy, $\pi : \mathcal{X} \rightarrow \mathcal{A}$, which maps a context vector $\mathcal{X}$ to an action $\mathcal{A}$ with the objective of maximizing cumulative reward, i.e., improving joint inference accuracy.

We begin by examining a contextual bandit that uses only the agent's confidence as input. Next, we extend the context to include information about the external task environment (task difficulty) in addition to the internal agents' state (confidence).

Algorithm 5 Contextual Bandit for Joint Inference

Require: Agents set $\mathcal{K} = \{H, AI\}$.
Exploration parameter ϵ, number of trials T.
Sliding window update interval $W = 20$ for metacognitive stats
Burn-in trials for initial metacognitive measure $B = 50$.
Initialize: Empty buffers of past performance $\mathcal{H}_H$, $\mathcal{H}_{AI}$.
Predictive models: $\mathcal{M}_H$ and $\mathcal{M}_{AI}$ (e.g, SGDC).
Total Reward $R_{total} \leftarrow 0$

for trial $t = 1, 2, \ldots, T$ **do**

▷ *Observe immediate, trial-specific information*

Observe agents' confidence, reward $c_{k,t}, r_{k,t} \forall k \in \mathcal{K}$
Task difficulty z_t.

▷ *Update & retrieve metacognitive traits*

if $(t - B) \pmod W = 1$ **then**
  $\mathcal{H}_k \leftarrow \{(c_{k,i}, r_{k,i})\}_{i=0}^{t-1}$.
  $\mu_{k,t} \leftarrow \text{auroc_2}(\mathcal{H}_k)$
  $\tau_{k,t} \leftarrow \text{ece}(\mathcal{H}_k)$.
else
  $\mu_{k,t} \leftarrow \mu_{k,t-1}$
  $\tau_{k,t} \leftarrow \tau_{k,t-1}$

▷ *Construct context vector (X)*

// With Only Confidence $X_{k,t} \leftarrow [1,\ c_{k,t}]$

// With External Context $X_{k,t}^{ext} \leftarrow [1,\ c_{k,t},\ z_t]$

// Metacognitive Context $X_{k,t}^{metacog} \leftarrow [1,\ c_{k,t},\ \mu_{k,t},\ \tau_{k,t}]$

▷ *Choose human or AI*

if random() $< \epsilon$ **then**
  I_{Team} : ChosenArm$_t \leftarrow$ randomArm(k)
else
  $\hat{r}_{H,t} \leftarrow \mathcal{M}_k$.predict(context_vector$(X)$ $\forall X \in \{X_{k,t}, X_{k,t}^{ext}, X_{k,t}^{metacog}\}$
  I_{Team} : ChosenArm$_t \leftarrow \arg\max_{k \in \mathcal{K}} \hat{r}_{k,t}$
Observe reward $R_t \in \{0, 1\}$.
Append performance data to $\mathcal{H}_k$.
Update the model $\mathcal{M}_k$ of the chosen arm k for each algo with $(X_{k,t}, R_t)$, $(X_{k,t}^{ext}, R_t)$ and $(X_{k,t}^{metacog}, R_t)$

▷ *Calculate R_{total} obtained for each bandit algo*

$R_{total} \leftarrow R_{total} + R_t$.

return Accuracy of each joint inference: $R_{total}/(T - B)$.

Finally, we incorporate metacognitive features, including metacognitive sensitivity (measured using AUROC2) and metacognitive calibration (measured using ECE), into the context vector. The implementation code is illustrate in Algorithm. 5.

This process is summarized as follows:

- **Action Space ($\mathcal{A}$):** The bandit has two actions to select: $\mathcal{A} = \{\text{Human's Inference } (I_H), \text{AI's Inference } (I_{AI})\}$.
- **Reward (R):** The reward is binary: $r = 1$ if the selected inference is correct, and $r = 0$ otherwise.

Please note that the reward is modelled as a linear relationship to context, so an offset $[1 \dots]$ is needed in the context vector, and each of the context vectors (CV) below corresponds to an individual bandit algorithm. Since they can be executed within the same loop, we include them in a single representation within the Algorithm. 5.

- **CV with only agent intrinsic information**: Via Online Linear Regression Bandits (OLRB), the first algorithm uses a context vector containing only the confidence ratings of the two agents: $X_{k\in\{H,AI\}} = [1,\ c_k]$ (OLRB-Conf).
- **CV with extrinsic information**: The context vector is augmented with the task difficulty (z): $X^{ext}_{k\in\{H,AI\}} = [1,\ c_k,\ z]$. The difficulty z is represented as a one-hot encoded vector corresponding to the four levels (OLRB-ConfDiff).
- **CV with metacognitive information**: The context vector is augmented with medium-term traits including metacognitive sensitivity (μ_k) and metacognitive calibration (τ_k): $X^{metacog}_{k\in\{H,AI\}} = [1,\ c_k,\ \mu_k,\ \tau_k]$ (OLBR-Metacog).

Latent Variable Bandit for Unobservable Difficulty Level

While the contextual bandit leverages observable signals, its reliance on an explicit, ground-truth measure of task difficulty (z) is a practical limitation. The final model addresses the practical limitation that in many real-world scenarios, the true task difficulty is not an observable variable. But we can model a causal link that task difficulty (Z) directly influences agents' confidence in their choice (X), and together, they determine the final reward (R). To address this, we introduce an *Online Latent Variable Bandit (OLVB)* framework. This model operates under a causal assumption, represented by the graph $X \leftarrow Z$ and $(Z, X) \rightarrow R$, where:

- Z is the unobservable latent task difficulty.
- X represents the observable context, specifically the human and AI confidences (c_H, c_{AI}).
- R is the resulting reward.

Our model assumes a bidirectional causal relationship between task difficulty and confidence. While increased difficulty (e.g., smaller delay differences) naturally degrades confidence by reducing signal-to-noise ratios, the kind of relationship (e.g., linear vs. non-linear) is what the bandit algorithm seeks to learn as part of the context vector. By incorporating difficulty as an explicit feature, the algorithm can "re-calibrate" the weight it gives to human confidence when the task enters an objective high-difficulty regime.

This is an extension into testing a adaptive joint inference algorithm where the external variable (difficulty) is not measurable. We define our models, priors and its relation to latent variables as following:

Priors

We place prior distributions on all unknown parameters to encode initial beliefs and regularize the model.

Prior for Reward Weights $p(\Theta)$: A standard Gaussian prior is used for the weights: $p(\Theta) = \mathcal{N}(\mu_0, \Sigma_0)$

Prior for Difficulty $p(Z_t)$: Categorical distribution is used as the prior for the discrete difficulty: $p(Z_t) = \text{Categorical}(\pi_{z,0})$

Prior for Confidence Parameters $p(\Xi)$: The conjugate prior for the parameters of a Categorical distribution (the probability vectors π_z) is the Dirichlet distribution. We place an independent Dirichlet prior on each vector.

$$p(\Xi) = \prod_{z=1}^{4} p(\pi_z) = \prod_{z=1}^{4} \text{Dirichlet}(\pi_z|\alpha_0)$$

Models

Reward Model $p(R|\dots)$: The reward R_t is a binary outcome (success/failure) modeled by a Bernoulli distribution: $p(R_t = 1|X'''_{k_t,t}, \Theta) = \sigma(\Theta^T X'''_{k_t,t})$ The context vector $X'''_{k_t,t} = [1, c_{k_t,t}, z_{oh}]$ includes the one-hot encoded difficulty, z_{oh}.

Confidence Model $p(c|\dots)$: The confidence c_t is a discrete outcome from the set $\{1, 2, 3, 4\}$, modeled by a Categorical distribution. The probabilities depend on the true difficulty Z_t:$p(c_t = j|Z_t = z, \Xi) = (\pi_z)_j$. Here, $\Xi = \{\pi_1, \pi_2, \pi_3, \pi_4\}$ is the new set of learnable parameters, where each π_z is a 4-element probability vector.

Variational Distributions

We learn the posterior distribution using Stochastic Variational Inference (SVI) by optimising the parameters (weights) of approximate posterior distributions.

For Reward Weights $q(\Theta)$: The approximate posterior for the weights is a Gaussian distribution: $q(\Theta; \phi) = \mathcal{N}(\Theta|\mu_\theta, \Sigma_\theta)$.

For Difficulty at time t**,** $q(Z_t)$: The approximate posterior for the difficulty in a specific trial is a Categorical distribution: $q(Z_t; \psi_t) = \text{Categorical}(Z_t|\psi_{z,t})$.

For Confidence Parameters $q(\Xi)$: The approximate posterior for the confidence model parameters is a set of independent Dirichlet distributions.

$$q(\Xi; \xi) = \prod_{z=1}^{4} q(\pi_z; \xi_z) = \prod_{z=1}^{4} \text{Dirichlet}(\pi_z|\alpha_z)$$

These variational parameters: $\phi = \{\mu_\theta, \Sigma_\theta\}$, $\psi_t = \{\psi_{z,t}\}$, and $\xi = \{\alpha_1, \dots, \alpha_4\}$ is learned via Stochastic Variational Inference (SVI). The full implementation of OLVB is detailed in Algorithm 6 and explained below:

Stage I: Infer Current Difficulty

The newly observed confidence c_t is used to refine belief about the current difficulty Z_t.

(i) **Prior Belief:** The prior for the t difficulty comes from the $t-1$ posterior : $p(Z_t = z) = (\psi_{z,t-1})_z$.

Algorithm 6 Online Latent Variable Bandit (OLVB)

Require: Agents set $\mathcal{K} = \{H, AI\}$, learning rate ρ_t,
Priors:
Reward weights: $p(\Theta) = \mathcal{N}(\mu_0, \Sigma_0)$
Difficulty: $p(Z_t) = \text{Categorical}(\pi_{z,0})$
Confidence params: $p(\Xi) = \prod_{z=1}^{D} \text{Dirichlet}(\pi_z|\alpha_0)$
Initialize: Variational parameters:
$q(\Theta; \phi_0) = \mathcal{N}(\mu_{\theta,0}, \Sigma_{\theta,0})$
$q(Z_0; \psi_0) = \text{Categorical}(\psi_{z,0})$
$q(\Xi; \xi_0) = \prod_{z=1}^{D} \text{Dirichlet}(\pi_z|\alpha_{z,0})$
Total Reward $R_{total} \leftarrow 0$

for $t = 1, 2, ..., T$ **do**

▷ *Infer Current Difficulty*

Observe confidences $c_{k,t}$, $\forall k \in \mathcal{K}$
Set prior for current difficulty: $p(Z_t = z) \leftarrow (\psi_{z,t-1})_z$
for Each difficulty level $z \in \{1, ..., D\}$ **do**
Calculate expected confidence probabilities:
$\mathbb{E}_q[\pi_z] \leftarrow \alpha_{z,t-1} / \sum_j (\alpha_{z,t-1})_j$
Calculate likelihood: $L(z) \leftarrow \prod_{k \in \mathcal{K}} (\mathbb{E}_q[\pi_z])_{c_{k,t}}$
Compute posterior for difficulty:
$q(Z_t = z|c_t) \propto L(z) \cdot p(Z_t = z)$
Normalize $q(Z_t|c_t)$ to create the prob vector $\psi_{z,t|c}$

▷ *Construct context vector*

Sample difficulty from the new posterior:
$\tilde{z}_t \sim \text{Categorical}(\psi_{z,t|c})$
Sample reward weights: $\tilde{\Theta}_t \sim q(\Theta; \phi_{t-1})$
// With Latent Inferred Context $X_{k,t}^{latent} \leftarrow [1,\ c_{k,t},\ \tilde{z}_t]$

▷ *Choose Human or AI*

I_{Team} : ChosenArm$_t \leftarrow \arg\max_{k \in \mathcal{K}} \sigma(\tilde{\Theta}_t^T X_{k,t}''')$
Observe reward $R_t \in \{0, 1\}$.

▷ *Belief Update (Single SVI Step)*

Define the full ELBO for data point $\mathcal{D}_t = (c_{k_t,t}, R_t)$:

$$\mathcal{L}_t = \mathbb{E}_{q(\Theta)q(Z_t)q(\Xi)} \left[\log p(R_t|X_{k_t,t}''', \Theta, Z_t) + \log p(c_{k_t,t}|Z_t, \Xi) \right] - \text{KL}\left(q(\Theta)\|p(\Theta)\right) - \text{KL}\left(q(Z_t)\|p(Z_t)\right) - \sum_{z=1}^{4} \text{KL}\left(q(\pi_z)\|p(\pi_z)\right)$$

Estimate gradient $\nabla \mathcal{L}_t$ w.r.t. $\{\phi, \psi_t, \xi\}$
Update reward weight params: $\phi_t \leftarrow \phi_{t-1} + \rho_t \nabla_\phi \mathcal{L}_t$
Update confidence model params: $\xi_t \leftarrow \xi_{t-1} + \rho_t \nabla_\xi \mathcal{L}_t$
Update difficulty params: $\psi_t \leftarrow \nabla_{\psi_t} \mathcal{L}_t$
$R_{total} \leftarrow R_{total} + R_t$.

return Accuracy of each joint inference: $R_{total}/(T)$.

(ii) **Confidence Likelihood:** We need $p(c_t|Z_t = z)$. This requires the parameters π_z, which are themselves uncertain. We use their expected value from the current variational posterior $q(\Xi; \xi_{t-1})$. The expectation of a Dirichlet(α_z) is: $\mathbb{E}_q[\pi_z] = \alpha_z \ / \ \sum_j (\alpha_z)_j$

The likelihood is therefore:
$p(c_t = j|Z_t = z) \approx (\mathbb{E}_q[\pi_z])_j. \ \forall j \in \{1, 2, 3, 4\}.$

(iii) **Posterior Belief:** Using Bayes' rule for each difficulty level $z \in \{1, 2, 3, 4\}$:

$$q(Z_t = z|c_t) \propto p(c_t|Z_t = z, \mathbb{E}_q[\Xi]) \cdot p(Z_t = z)$$

After calculating this for all z, we normalize the results to get the full posterior probability vector for the current difficulty.

Stage II: Action Selection

(i) Sample difficulty from the new, informed posterior: $\tilde{z}_t \sim q(Z_t|c_t)$.

(ii) Sample reward weights from their posterior: $\tilde{\Theta}_t \sim q(\Theta; \phi_{t-1})$.

(iii) Construct the context vector $X^{latent}_{k,t} \leftarrow [1, \ c_{k,t}, \ \tilde{z}_t]$ using one-hot encoding of $\tilde{z}_t$.

(iv) Select the arm: $k_t = \arg\max_{k \in \{H, AI\}} \sigma(\tilde{\Theta}_t^T X'''_{k,t})$.

The SVI Objective

After observing the reward R_t, all beliefs is updated by taking a single stochastic gradient step on the expanded Evidence Lower Bound (ELBO). The objective function to be maximised is:

$$\begin{aligned} \mathcal{L}_t = \mathbb{E}_{q(\Theta)q(Z_t)q(\Xi)} \Big[& \log p(R_t|X'''_{k_t,t}, \Theta, Z_t) + \log p(c_{k_t,t}|Z_t, \Xi) \Big] \\ & - \mathrm{KL}\left(q(\Theta) \| p(\Theta)\right) - \mathrm{KL}\left(q(Z_t) \| p(Z_t)\right) \\ & - \sum_{z=1}^{4} \mathrm{KL}\left(q(\pi_z) \| p(\pi_z)\right) \end{aligned} \tag{6.2}$$

6.3 Results

This section presents the empirical evaluation of the proposed adaptive joint inference frameworks against static heuristics. We systematically analyse the performance of each model to address the research questions outlined previously (**RQ0-RQ3**), focusing on overall accuracy, performance across varying task difficulties, and robustness when context is unobservable.

6.3.1 Adaptive Joint Inference Outperforms Heuristics

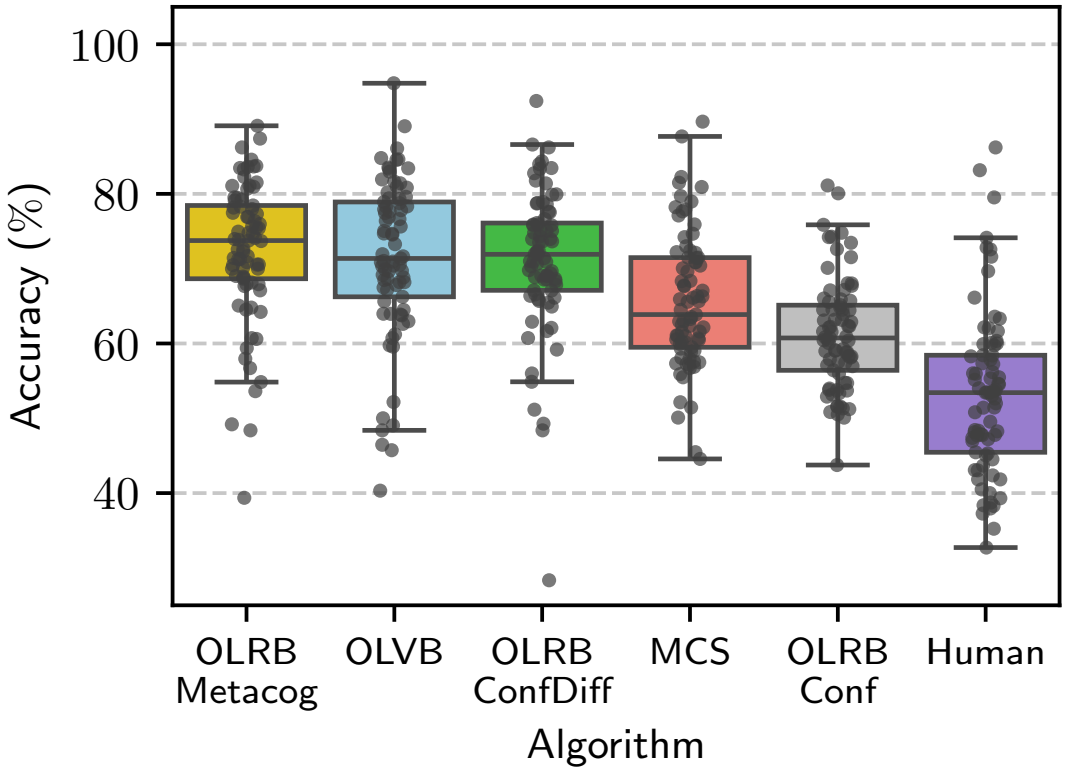


Figure 6.2: Overall Joint Inference Accuracy: The performance of all tested methods across all trials where the human and AI disagreed. The adaptive bandit models incorporating task context (OLRB-Metacog, OLRB-ConfDiff, OLVB) demonstrate higher median accuracies than the static MCS heuristic and the Human baseline.

Our first key result demonstrates that adaptive learning algorithms significantly outperform static baselines, directly addressing **RQ0**. As shown in Table 6.2 and visualized in Fig. 6.2, the top-performing adaptive models, OLRB-Metacog

(72.4%±9.3%) and OLRB-ConfDiff (71.4%±9.7%), achieved markedly higher overall accuracy compared to the static Maximum Confidence Slating (MCS) heuristic (65.6% ± 9.1%). A paired t-test confirmed that these improvements were statistically significant (OLRB-Metacog vs. MCS: $t(7796) = 7.91, p < 0.001$; OLRB-ConfDiff vs. MCS: $t(7796) = 6.48, p < 0.001$). These results establish that while MCS serves as a strong non-learning baseline, an algorithm that dynamically adapts its policy based on context achieves significantly superior team performance.

Crucially, this advantage is contingent on the quality of contextual information. Addressing **RQ1**, we find that enriching the context with task difficulty is not merely beneficial, but essential. The OLRB-ConfDiff model, which incorporated both confidence and task difficulty, achieved approximately 10% higher overall accuracy than OLRB-Conf (61.2% ± 7.3%), which used confidence alone. OLRB-Conf also performed worse than the simpler, non-adaptive MCS heuristic (65.6% ± 9.1%).

This underscores a core finding of our work: adaptive arbitration is powerful, but its success hinges on moving beyond simple agent-centric signals (confidence) to incorporate richer, world-centric features like task difficulty.

6.3.2 Performance Across Various Difficulty Conditions

While the overall accuracies establish a clear hierarchy, the true value of the adaptive frameworks is revealed when analyzing performance across different task difficulties, as illustrated in Fig. 3. The data shows a dramatic divergence in resilience between adaptive and static models. The performance of both the Human baseline and the static MCS heuristic collapses in challenging scenarios. For instance, MCS accuracy drops from a high of 80.0% in Low difficulty trials to just 53.4% in High difficulty trials. This demonstrates that while simple heuristics can be effective when the task is easy, they become brittle when task difficulty increases.

In contrast, the context-enriched bandit models exhibit resilience. The OLRB-Metacog, OLRB-ConfDiff, and OLVB models all maintain high accuracy (above 71%) even in High and Extreme difficulty conditions. At the Hard and Extreme difficulty levels, the accuracy gap between the adaptive OLRB-Metacog and the

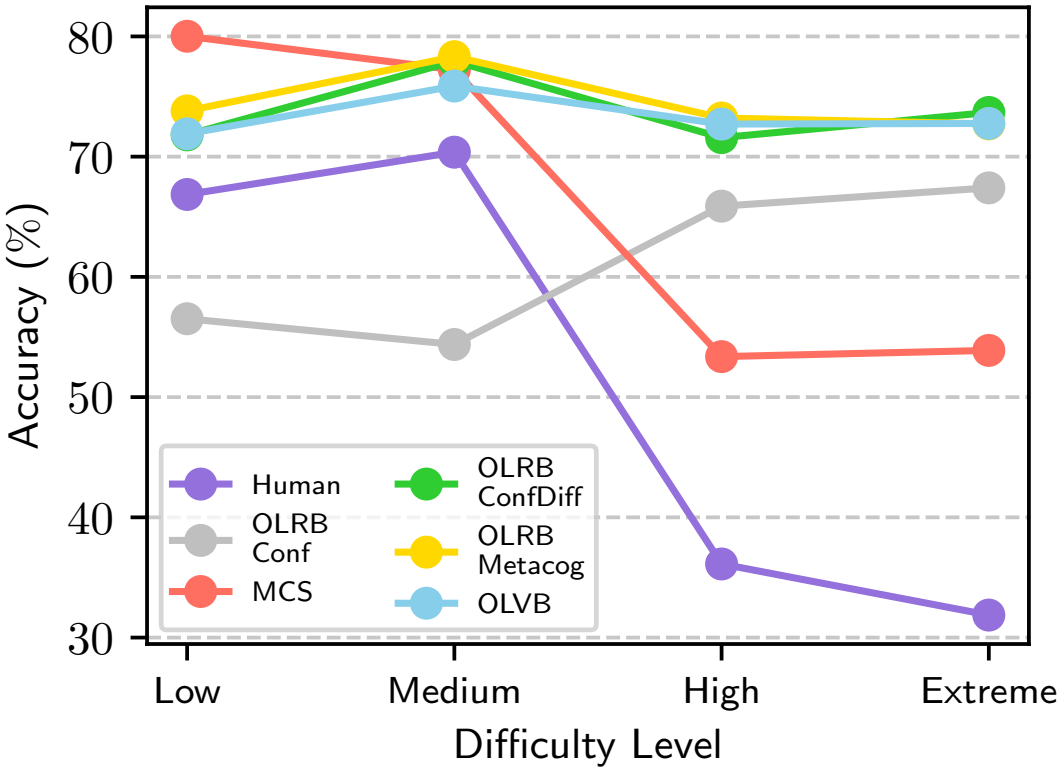


Figure 6.3: Joint Inference Accuracy Across Task Difficulty Levels: This figure illustrates the resilience of adaptive models compared to static baselines. While the performance of Human and MCS collapses on High and Extreme difficulty tasks, the top adaptive bandits (OLRB-Metacog, OLRB-ConfDiff, OLVB) maintain consistently high accuracy.

heuristic MCS is 19.8% and 18.8%, respectively. This supports the conclusion that the framework facilitates human-AI complementarity precisely where it is most needed. These findings directly address **RQ2**, showing that metacognitive signals provide an additional, consistent performance advantage. The OLRB-Metacog model slightly outperforms the OLRB-ConfDiff model across Low, Medium, and High difficulty settings (Table 6.2). This suggests that medium-term traits help refine the arbitration policy beyond what is possible with immediate context alone, and critically, it achieves the highest performance without relying on external environmental variables like task difficulty, which may not always be observable in real-world deployments.

Method	Overall	Low Difficulty	Medium Difficulty	High Difficulty	Extreme Difficulty
Human	$53.2\% \pm 11.0\%$	$66.9\% \pm 4.7\%$	$70.4\% \pm 4.6\%$	$36.1\% \pm 4.8\%$	$31.9\% \pm 4.5\%$
Max Confidence Slating (MCS)	$65.6\% \pm 9.1\%$	$80.0\% \pm 4.0\%$	$77.2\% \pm 4.2\%$	$53.4\% \pm 4.5\%$	$53.9\% \pm 4.8\%$
Contextual Bandit (OLRB-Conf)	$61.2\% \pm 7.3\%$	$56.5\% \pm 4.5\%$	$54.4\% \pm 4.2\%$	$65.9\% \pm 3.4\%$	$57.4\% \pm 4.3\%$
Contextual Bandit (OLRB-ConfDiff)	$71.4\% \pm 9.7\%$	$71.8\% \pm 4.5\%$	$77.9\% \pm 4.1\%$	$71.6\% \pm 4.5\%$	$73.7\% \pm 4.1\%$
Metacognitive Bandit (OLRB-Metacog)	$72.4\% \pm 9.3\%$	$73.8\% \pm 4.4\%$	$78.3\% \pm 5.1\%$	$73.2\% \pm 4.3\%$	$72.7\% \pm 4.4\%$
Latent Variable Bandit (OLVB)	$71.1\% \pm 10.6\%$	$71.9\% \pm 4.9\%$	$75.9\% \pm 4.9\%$	$72.7\% \pm 4.7\%$	$72.8\% \pm 4.6\%$

Table 6.2: Summary of Joint Inference Accuracy (Mean ± SD) Across All Conditions.

6.3.3 Latent Variable Modeling of Unobservable Context

A practical limitation of the OLRB-ConfDiff model is its reliance on an explicit, ground-truth measure of task difficulty, which is often unavailable in real-world settings. To address this, our final analysis evaluates the Online Latent Variable Bandit (OLVB), a model designed to infer task difficulty when it is unobservable. Answering **RQ3**, our results demonstrate the remarkable robustness of this approach. As shown in Table I and Fig. 3, the overall accuracy of the OLVB ($71.1\% \pm 10.6\%$) was statistically comparable to that of the OLRB-ConfDiff model ($71.4\% \pm 9.7\%$), which had access to the ground-truth difficulty.

This performance parity holds across all difficulty conditions, with the two models' accuracy lines on Fig. 3 tracking each other almost perfectly. This is a powerful demonstration that the causally-informed learning approach is effective. By explicitly modeling task difficulty as a latent confounder, the system was able to learn an effective arbitration policy even in a partially-observable environment.

6.4 Discussion

In this chapter, we presented and validated a framework for adaptive human-AI joint inference in a dynamic robotics task. We demonstrated that contextual bandits significantly outperform static confidence-based heuristics, particularly in challenging, high-difficulty scenarios where effective collaboration is most critical. By showing the value of incorporating task difficulty, agent-specific metacognitive traits, and latent variable modeling, we have established that the key to resilient teaming lies in learning a nuanced arbitration policy from a rich contextual landscape.

Our results highlight two practical paths toward deploying such systems in the real world. First, the superior performance of the OLRB-Metacog model is a critical

finding. It proves that high resilience can be achieved without relying on external environmental sensors, using only agent-internal, medium-term metacognitive traits. This is particularly valuable for unstructured environments where instrumenting the world to get a clean "difficulty" signal is infeasible. Second, the success of the OLVB model provides a robust alternative, confirming that even when external context is deemed crucial, it does not need to be perfectly observable. By successfully inferring latent task difficulty, the OLVB validates that the framework's benefits are not contingent on having perfect environmental information, making it a principled and practical solution for real-world deployment.

Limitations and Future Work

We acknowledge that our experiments were conducted in a simulated environment with post-hoc analysis, offering precise control and repeatability but not yet capturing the full richness of real-world physics and sensing. The task focused on a single diagnostic scenario, with difficulty modeled using four discrete levels rather than a continuous spectrum. These limitations suggest several directions for future work:

- Validation on physical systems: Deploy the framework on real robots to test robustness against noise.
- Broader fault types: Extend the approach to other latent faults such as sensor drift, actuator wear.
- Closing the loop: Move from diagnosis to adaptation by integrating inference with planning and control.

7 Conclusion

7.1 Summary of Contributions

In this thesis, we have addressed the problem of joint inference in robot teleoperation missions. Our research focused on the task of inferring robot's latent control delays and was structured around three-stage investigation, corresponding to a progressive development of a joint inference framework. First, we established a baseline for dyadic performance by examining collaboration in human-human teams using a static, confidence-based heuristic in enhancing human-human joint decision capability. Second, we extended this investigation to human-AI dyads, analyzing whether the same heuristic could achieve complementarity and how the AI's metacognitive profile influenced the team's dynamics. Finally, we moved beyond static rules to develop an adaptive joint inference framework, using contextual bandits to create a policy that learns to arbitrate between partners in a sequential, context-aware manner, advancing human-AI complementarity in robotic missions.

To guide this investigation, this thesis defined a set of research questions aimed at systematically exploring dyadic collaboration, from human-human teams to adaptive human-AI systems. The following sections summarize the progress made in addressing each of these questions, detailing the key findings and contributions presented in Chapters 4, 5, and 6.

RQ1: Human & Human Complementarity

To what extent can human dyads achieve joint inference complementarity by sharing confidence about robot control delays? How do individual differences in metacognitive sensitivity affect the performance of this collaboration?

Chapter 4.4 established that human-human dyads can achieve joint inference complementarity by leveraging a simple, confidence-based heuristic. The experiment and subsequent analysis demonstrated through a large user study (N=100) that applying the Maximum Confidence Slating (MCS) heuristic improves a dyad's joint inference accuracy beyond that of its own highest-performing member. This finding confirmed that confidence sharing is a viable mechanism for enhancing human collaborative performance in a dynamic robotics task.

Furthermore, the research identified the specific metacognitive factors that regulate the success of this approach. The analysis revealed that the performance gains from MCS are not uniform across all pairings. Specifically, smaller accuracy gains were observed in dyads with large discrepancies in skill level, while pairings of individuals with higher collective metacognitive sensitivity and calibration were correlated with greater joint inference accuracy. These findings provided an insightful understanding of when and why confidence-based heuristics are most effective, laying the groundwork for the subsequent investigation into human-AI teams.

RQ2: Human & AI-DSS Complementarity

Can the principles of confidence-sharing be extended to human & AI-DSS dyads to achieve similar complementarity in the same robotics task? How does the metacognitive profile of the AI and its alignment with the human partner affect team performance?

Having established the viability of confidence-sharing in human-human teams, Chapter 5.4 of this thesis extended this metacognitive framework to human-AI dyads. This research provided the first demonstration that these principles can be successfully applied to achieve complementarity in human-AI collaboration on a dynamic robotics task. A second user study (N=100) confirmed that using the

MCS heuristic again led to higher joint inference accuracy compared to both the performance of the individual human and the traditional AI-assisted paradigm.

This investigation revealed the critical role of the AI-DSS's own metacognitive profile in regulating team performance. The AI's metacognitive sensitivity was found to not only affect overall team accuracy but also to influence the human partner's own inference and confidence dynamics throughout the task. These findings underscored the importance of designing AI partners with well-calibrated metacognitive abilities. The work also highlighted the robustness of the MCS heuristic, demonstrating that it remained effective even in non-ideal pairings where both the human and AI partners exhibited poor metacognitive sensitivity. This validation of the framework in a human-AI context motivated the final stage of the research: moving beyond static heuristics to more adaptive methods.

RQ3: Adaptive Joint Inference

Can adaptive, sequential decision-making algorithms enhance joint inference performance beyond static heuristics? What constitutes the most effective contextual information for these algorithms to optimize team performance, particularly in challenging scenarios?

Finally, to overcome the limitations of static rules, the thesis introduced and validated an adaptive joint inference framework using contextual bandits, as presented in Chapter 6. This work directly answered the final research question, demonstrating that a sequential, learning-based approach consistently surpasses the performance of the static MCS heuristic. This finding established that an adaptive policy that learns how to arbitrate between partners can achieve a robust level of joint inference accuracy.

Our investigation also identified the specific contextual information that is critical for unlocking the potential of these adaptive algorithms. A key finding was that a bandit policy relying solely on agent confidence as its context did not yield a significant advantage over the simpler MCS heuristic. However, a substantial performance improvement was observed when the bandit's context was augmented

to include external information such as task difficulty alongside confidence. The most important contribution of this framework was its robust performance in high-difficulty tasks. In these challenging scenarios where the accuracy of static heuristics degraded, the adaptive contextual bandit method achieved significant accuracy gains, facilitating human-AI complementarity precisely when it was most needed.

By grounding our approach in established insights from neuroscience and psychology, we aim to develop more than a task-specific solution. The teleoperation delay case serves as a testbed, but the framework is designed to illustrate how confidence, metacognition, and adaptive decision processes can enhance teamwork in various human-AI settings. In this way, our work contributes to a broader foundation for analyzing complex teaming challenges.

7.1.1 Generalizability to Real-World Robotic Tasks

While this research utilized teleoperation delay as a representative latent fault, the proposed framework is designed to be domain-agnostic and generalizable across various robotic platforms and mission types. Control delay served as a high-fidelity proxy because it necessitates active, temporally extended inference rather than static observation . However, the core principles of metacognitive arbitration—specifically the use of Maximum Confidence Slating (MCS) and contextual bandits—can be extended to any scenario involving unobservable system degradation . In real-world deployments, this could include detecting and mitigating sensor drift in autonomous navigation, identifying mechanical wear in surgical robotics, or managing communication throughput fluctuations in multi-UAV swarms. Because the adaptive models rely on universal signals of sensitivity and calibration rather than task-specific features, the framework provides a scalable pathway for maintaining human-robot complementarity across diverse spatiotemporal challenges.

7.1.2 Impact on HRI Dynamics

The transition from a "simple assistance" model to a metacognitive, peer-to-peer architecture fundamentally reshapes the dynamics of Human-Robot Interaction

(HRI) . Our findings demonstrate that the AI's metacognitive profile is not a passive property; it actively influences the human partner's decision-making trajectory, confidence levels, and propensity to revise incorrect inferences. By establishing a bidirectional exchange of uncertainty, the system addresses the "cognitive interface" bottleneck that frequently leads to over-reliance or under-reliance on automated advice . The implementation of the contextual bandit framework further enhances system resilience by autonomously identifying high-difficulty scenarios where human-AI complementarity is most vital. This transforms the robotic agent from a mere tool into a collaborative partner capable of shaping a shared mental model for more robust joint performance.

In summary, within this thesis, we present a systematic progression of research, moving from an initial validation of confidence-based heuristics in human-human teams to the development of a sophisticated, adaptive framework for human-AI collaboration. Across three user studies and a progressive analytical framework, the work has demonstrated not only that complementarity can be achieved in joint inference tasks, but also identified the specific metacognitive and contextual factors required to do so effectively. The primary contribution of this research is a validated, context-aware framework that enables more resilient and effective human-robot teaming.

7.2 Limitations

While the specific limitations of each study were addressed in Chapters 4, 5, and 6, it is important to reflect on the broader limitations of the overall research framing and approach presented in our work. These constraints define the boundaries of the current contributions and provide a context for the directions of future work.

Post-Hoc Analysis and Social Dynamics

Our methodological post-hoc analysis choice was deliberate, as it enabled the creation of a large and varied dataset essential for the robust training and evaluation of the adaptive bandit algorithms. By systematically pairing human and AI data

streams after collection, we could control for variables like task difficulty and agent calibration across thousands of simulated interactions, which would be infeasible in a live study. However, this approach excludes the dynamic social factors inherent in social collaboration. For instance, a human operator's reported confidence and their willingness to accept an AI's suggestion might be influenced by their evolving trust in their partner over the course of a mission, a dynamic that can be extended in the current experimental design.

Abstraction of the AI-Decision Support System

The AI-DSS was intentionally designed as a data-driven system that samples its inferences and confidence ratings from distributions of prior human performance. This approach was a methodological necessity; it allowed for the precise experimental control over the AI's metacognitive characteristics needed to isolate its effects on the human partner and overall team performance. However, this means the AI was not a fully autonomous agent, and we acknowledge that high-stakes scenarios may introduce psychological stressors that distort confidence calibration beyond what can be captured by metacognitive arbitration in simulation.

7.3 Future Work

Our research presented here serves as a foundation upon which more complex and capable human-AI/robot teams can be built. The following sections outline three key directions for extending this research:

FW1: Advanced Capabilities AI-DSS

A primary direction for future work is to advance the AI-Decision Support System beyond its current form. While sampling from prior human data provided the necessary experimental control to isolate metacognitive effects, a real-world system would require an AI that can generate its own inferences and confidence estimates directly from raw sensor data. Future research should investigate the integration of learning-based models, such as deep neural networks or Bayesian models, capable

of diagnosing latent faults like control delay from the robot's state information and control inputs. This would not only create a more ecologically valid AI partner but also introduce new research questions. For instance, how does the source of the AI's confidence (e.g., model uncertainty, softmax outputs) affect human trust and team performance? Furthermore, an AI could be designed to learn and adapt its own metacognitive calibration over time, potentially leading to a system that can actively improve its collaborative fluency with a human partner.

FW2: Broader Collaboration Tasks

A second avenue for future research involves transitioning the adaptive joint inference framework to physical robotic systems. This transition from simulation to reality opens several promising research questions. For instance, can the contextual bandit models, particularly the Latent Variable Bandit, effectively distinguish between the target latent fault and the general stochasticity of a physical system? How do the pressures and rich sensory feedback of operating an embodied robot impact the human partner's own metacognition monitoring, and is the framework robust enough to adapt to these shifts in real-time? Another key step would be to evaluate the system's performance not only on control delay but also on a broader class of latent faults, such as sensor drift or actuator wear, which are common in physical hardware.

FW3: Integrating Richer Communication and Explainability

The communication between agents in this work was limited to their choice and a numerical confidence rating. Future research could explore the integration of richer communication modalities. For instance, the adaptive framework could be extended to allow the AI to provide a concise explanation for its inference (e.g., "High confidence due to smooth trajectory" or "Low confidence due to sensor anomaly"). This would connect the framework to the field of Explainable AI (XAI) and allow for investigating whether such explanations improve the human's trust calibration and overall team effectiveness.

Appendices

A

Pilot Study Experiments

We conducted a pilot study with five participants to find the shortest driving time that participants felt allowed them to make a choice between the two robots. Participants drove each robot for intervals of 4, 6, 8, and 10 seconds, each with 5 trials. Three fixed levels of difficulties were considered, with delay differences in the control of the two robots: 50ms - 70ms; 50ms - 90ms; 50ms - 110ms. Therefore, each participant experienced a total of 60 trials.

At the end of the experiment, each participant was asked to identify the shortest time they found sufficient for making a decision among 4, 6, 8, or 10 seconds. All the participants reported that 6 seconds was the shortest time that allowed them to identify differences between two robots and make a decision. Further, we compared the success rate among all difficulties for the 4, 6, 8, and 10 seconds driving scenarios. The results are shown in Table A.1. We find that 6 seconds also leads to our desired target success rate of 70%.

Time (seconds)	Success rate (%)
4	63.3
6	71.2
8	77.8
10	91.1

Table A.1: Success rate for each pilot scenario.

B

Software Design and Metadata of Human-Human Experiment

B.1 Experiment Data Labels

The following table provides an overview of the variables collected during the study, grouped into two main categories: Environment Setup and Participant Data. These variables capture both the experimental conditions and the participants' responses.

Group	Variable Label	Explanation
Environment Setup	`participant_id`	The unique identifier for the participant, anonymized for privacy.
	`delay_1`	The injected delay time recorded for the 1st Robot Controller.
	`delay_2`	The injected delay time recorded for the 2nd Robot Controller.
	`timestamp`	The exact date and time when the experiment data was recorded.
	`location`	The randomised initial location of the Robots.
	`orientation`	The randomised initial orientation of the Robots.
	`door_configuration`	The configuration of the door used in each trials.
Participant Data	`user_choice`	The choice made by the user at the end of each trials.
	`correct_choice`	The correct choice of the lesser delayed in control Robot.
	`if_correct`	Indicator (1 or 0) of whether the user's choice was correct.
	`user_confidence`	The level of confidence expressed by the user regarding their choice.

Table B.2: Description of variables used in our human-human joint inference study.

B.2 Metacognitive Sensitivity Data

In this section, we provide additional visual representations of participant confidence ratings and metacognitive sensitivity through AUROC2 analysis. Figure B.1 shows the histogram of confidence ratings across four levels, highlighting the range of confidence levels expressed by participants during the task. This distribution demonstrates that the task successfully elicited a broad spectrum of confidence responses, supporting robust metacognitive sensitivity analysis.

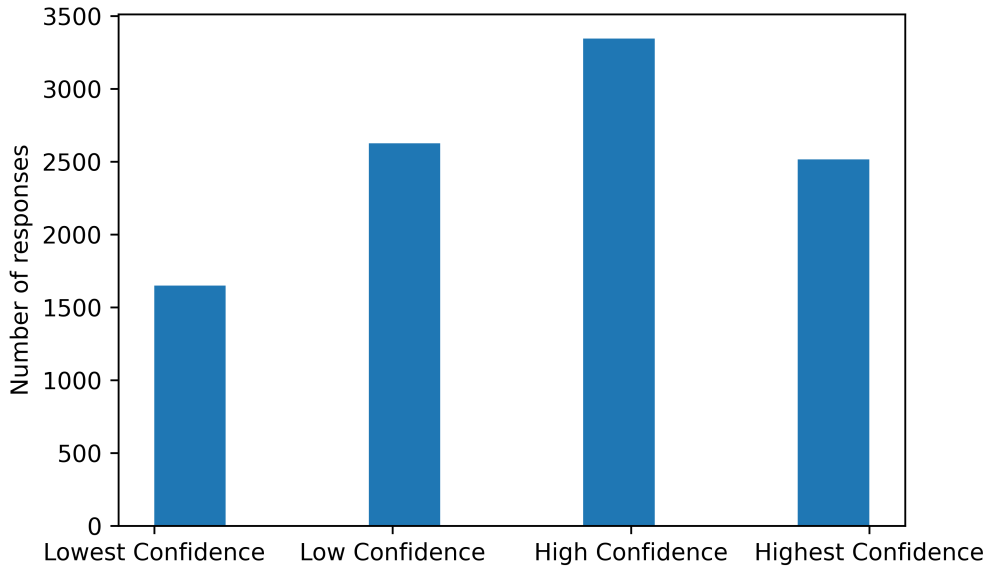


Figure B.1: Histogram of confidence ratings over all trials over the entire set of participants.

Figure B.2 illustrates the distribution of participants' AUROC2 values, highlighting the variability in metacognitive sensitivity across the group. AUROC2 scores range from 0 to 1, with higher values indicating stronger alignment between participants' confidence levels and their actual performance, expressing better metacognitive sensitivity.

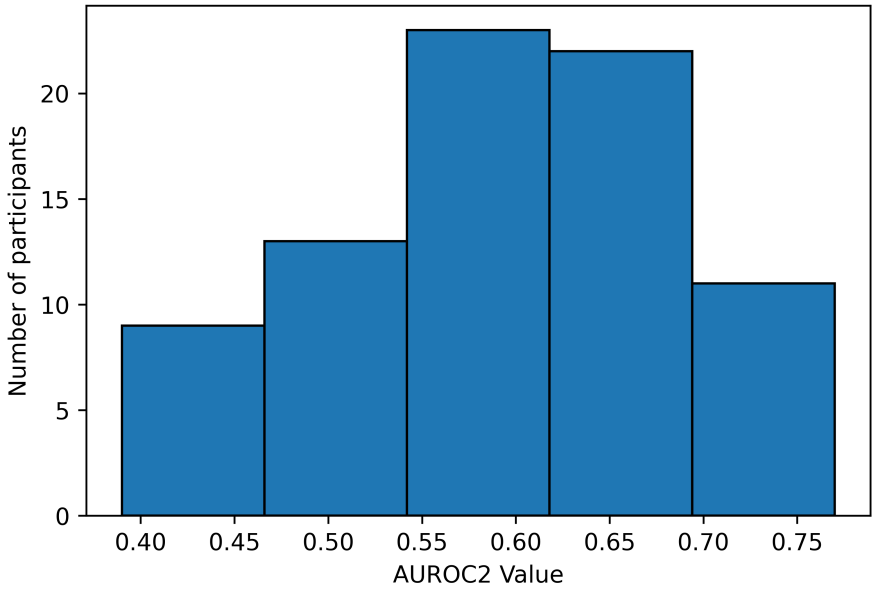


Figure B.2: Histogram of AUROC2 values over the entire set of participants.

Lastly, Figure B.3 illustrates the Type 2 ROC curve for a representative participant. The curve exemplifies how metacognitive sensitivity is derived from True Positive Rate (H2) and False Positive Rate (FA2) according to participants' performance and self-assessment.

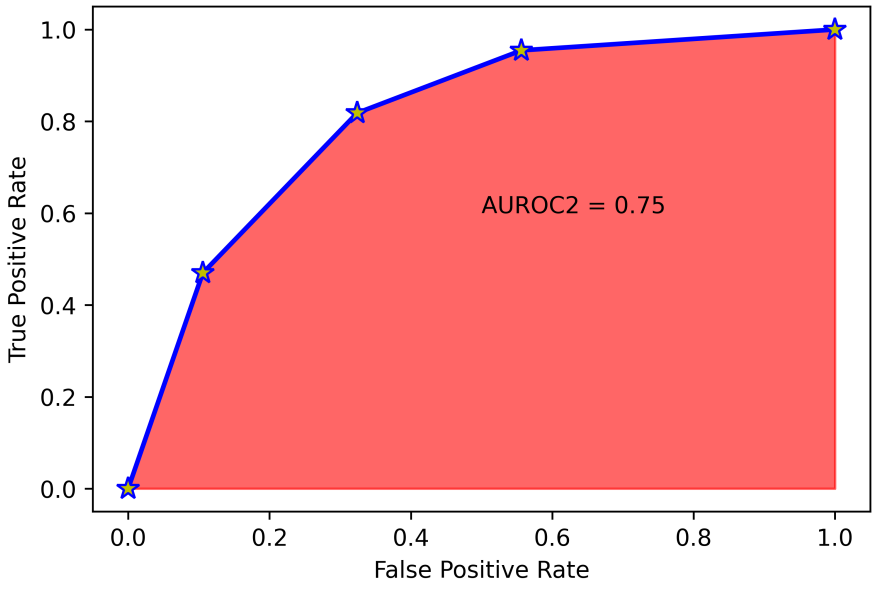


Figure B.3: Corresponding AUROC Curve for an Example Participant.

C

Software Design and Metadata of Human-AI Experiment

C.1 Experiment Data Variables

The following table provides an overview of the variables data collected during the study.

Group	Variable Label	Explanation
Environment Setup	`participant_id`	Human's unique anonymized identifier.
	`delay_1`	Injected delay time of 1st Robot.
	`delay_2`	Injected delay time of 2nd Robot.
	`timestamp`	exact date and time when experiment data was recorded.
	`location`	Initial location of Robots.
	`orientation`	Initial orientation of Robots.
	`door_config`	Configuration of door in each trials.
Participant Data	`user_inference`	Initial inference made by human at the end of each trial.
	`correct_inference`	The correct inference of the robot with less control delay.
	`user_confidence`	Human's initial confidence.
	`if_correct`	Indicator (1=correct, 0=incorrect) of Human's inference.
	`final_user_infer`	Human's final inference after AI-DSS inference.
	`final_user_conf`	Human's final confidence.
AI-DSS System	`ids_type`	Confidence-calibration of AI-DSS.
	`ids_robot_inf`	Inference made by the AI-DSS.
	`ids_confidence`	Confidence of AI-DSS.
	`if_ids_correct`	Indicator (1=correct, 0=incorrect) of AI-DSS's inference.

Table C.2: Description of variables used in our human-AI joint inference study.

C.2 Human Participant Demographics

We collected responses from 100 participants in total via Prolific Platform [1]. 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 [25].

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.

[1] www.prolific.com

References

[1] Huaping Liu, Di Guo, and Angelo Cangelosi. "Embodied intelligence: A synergy of morphology, action, perception and learning". In: *ACM Computing Surveys* 57.7 (2025), pp. 1–36.

[2] Manolis Chiou et al. "Experimental analysis of a variable autonomy framework for controlling a remotely operating mobile robot". In: *IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE. 2016.

[3] Manolis Chiou et al. "Robot-assisted nuclear disaster response: Report and insights from a field exercise". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2022.

[4] Harvey Stedman et al. "Evaluating immersive teleoperation interfaces: Coordinating robot radiation monitoring tasks in nuclear facilities". In: *Proceedings-IEEE International Conference on Robotics and Automation*. 2023.

[5] Katharina M Gabrecht. "Human factors of semi-autonomous robots for urban search and rescue". PhD thesis. University of Nottingham, 2016.

[6] Mitchell Pryor et al. "Radiation Surveys in Active Nuclear Facilities with Heterogeneous Collaborative Mobile Robots". In: *arXiv preprint arXiv:2402.15008* (2024).

[7] Geoffrey A Landis. "Robots and humans: Synergy in planetary exploration". In: *AIP Conference Proceedings*. Vol. 654. 1. American Institute of Physics. 2003, pp. 853–860.

[8] Yongchang Zhang et al. "Progress, challenges, and prospects of soft robotics for space applications". In: *Advanced Intelligent Systems* 5.3 (2023), p. 2200071.

[9] Induni N Weerarathna, David Raymond, and Anurag Luharia. "Human-robot collaboration for healthcare: a narrative review". In: *Cureus* (2023).

[10] Kaiqiang Zhang et al. "Radiation tolerance testing methodology of robotic manipulator prior to nuclear waste handling". In: *Frontiers in Robotics and AI* (2020).

[11] Duo Dong et al. "Research on key technology and application progress of rescue robot in nuclear accident emergency situation". In: *Nuclear Engineering and Technology* (2025), p. 103457.

[12] Benjamin Lambert et al. "Trustworthy clinical AI solutions: A unified review of uncertainty quantification in Deep Learning models for medical image analysis." In: *Artif. Intell. Medicine* 150 (2024), p. 102830.

[13] Chengchao Shen. "Multimodal AI-driven object detection with uncertainty quantification for cardiovascular risk assessment in autistic patients". In: *Frontiers in Cardiovascular Medicine* 12 (2025), p. 1606159.

[14] Sauliha Rabia Alli et al. “The potential of artificial intelligence tools for reducing uncertainty in medicine and directions for medical education”. In: *JMIR Medical Education* 10.1 (2024), e51446.

[15] Ransalu Senanayake. “The role of predictive uncertainty and diversity in embodied ai and robot learning”. In: *arXiv preprint arXiv:2405.03164* (2024).

[16] Thomas Timm Andersen et al. “Measuring and modelling delays in robot manipulators for temporally precise control using machine learning”. In: *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. 2015.

[17] Ugur Tumerdem. “Multilateral teleoperation under asymmetric time delays: L 2 stability and robustness”. In: *International Journal of Advanced Robotic Systems* (2017).

[18] Derek H Arnold et al. “On why we lack confidence in some signal-detection-based analyses of confidence”. In: *Consciousness and cognition* 113 (2023), p. 103532.

[19] Duc-An Nguyen et al. “Group Decision-Making in Robot Teleoperation: Two Heads are Better Than One”. In: *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 2025.

[20] Duc-An Nguyen et al. “Towards Joint Human-AI Inference for Robot Missions: A Confidence-Based Approach”. In: arXiv preprint arXiv. 2025.

[21] Duc-An Nguyen et al. “Beyond Confidence: A Bandit-Based Framework for Joint Human-AI Inference in Robot Missions”. In: arXiv preprint arXiv. 2026.

[22] Stephen M Fleming, Josefien Huijgen, and Raymond J Dolan. “Prefrontal contributions to metacognition in perceptual decision making”. In: *Journal of Neuroscience* 32.18 (2012), pp. 6117–6125.

[23] Xiaoyu Jia, Weijian Li, and Liren Cao. “The role of metacognitive components in creative thinking”. In: *Frontiers in psychology* 10 (2019), p. 2404.

[24] Jan Treur and Gerrit Glas. “A multi-level cognitive architecture for self-referencing, self-awareness and self-interpretation”. In: *Cognitive Systems Research* 68 (2021), pp. 125–142.

[25] Marion Rouault et al. “Human metacognition across domains: insights from individual differences and neuroimaging”. In: *Personality Neuroscience* 1 (2018).

[26] Leona Polyanskaya. “I know that I know. But do I know that I do not know?” In: *Frontiers in Psychology* 14 (2023), p. 1128200.

[27] Stephen M Fleming and Hakwan C Lau. “How to measure metacognition”. In: *Frontiers in Human Neuroscience* 8 (2014), p. 443.

[28] Manuel Rausch, Hermann J Müller, and Michael Zehetleitner. “Metacognitive sensitivity of subjective reports of decisional confidence and visual experience”. In: *Consciousness and cognition* 35 (2015), pp. 192–205.

[29] Stephen M Fleming and Raymond J Dolan. “The neural basis of metacognitive ability”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1594 (2012), pp. 1338–1349.

[30] Linda Bol and Douglas J Hacker. “Calibration research: Where do we go from here?” In: *Frontiers in psychology* 3 (2012), p. 229.

[31] Songyu Liu and Mark Steyvers. "Combining Human and AI Strengths in Object Counting under Information Asymmetry". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. 2024.

[32] Sascha Meyen et al. "Group decisions based on confidence weighted majority voting". In: *Cognitive research: principles and implications* 6.1 (2021), p. 18.

[33] Tsutomu Harada. "Three heads are better than two: Comparing learning properties and performances across individuals, dyads, and triads through a computational approach". In: *PLOS ONE* 16 (2021). Ed. by Darrell A. Worthy.

[34] Asher Koriat. "When two heads are better than one and when they can be worse: The amplification hypothesis." In: *Journal of Experimental Psychology: General* 144.5 (2015), pp. 934–950.

[35] Ali Mahmoodi et al. "Learning to Make Collective Decisions: The Impact of Confidence Escalation". In: *PLoS ONE* 8.12 (2013). Ed. by Stephen C. Pratt, e81195.

[36] Niccolo Pescetelli, Geraint Rees, and Bahador Bahrami. "The perceptual and social components of metacognition." In: *Journal of Experimental Psychology: General* 145.8 (2016), pp. 949–965.

[37] Marion Rouault et al. "Human Metacognition Across Domains: Insights from Individual Differences and Neuroimaging". In: *Personality Neuroscience* 1 (2018), e17.

[38] Mihaela Aldea et al. "Molecular Tumor Boards: A consensus statement from the International Association for the Study of Lung Cancer". In: *Journal of Thoracic Oncology* (2025).

[39] Martín Angel et al. "The value of virtual molecular tumor boards for informed clinical decision-making". In: *The Oncologist* 29.7 (2024), pp. 554–559.

[40] Tetiana Shmelova et al. "Decision-Making Models by the Aircraft Crew in Emergency" Depressurization"." In: *CITRisk*. 2022, pp. 132–145.

[41] Judith Orasanu. "Shared mental models and crew decision making". In: *12th Annual Conference. CSS Pod*. Psychology Press. 2022, pp. 1066–1066.

[42] Dan Bang et al. "Confidence matching in group decision-making". In: *Nature Human Behaviour* 1.6 (2017), p. 0117.

[43] Dan Bang and Chris D Frith. "Making better decisions in groups". In: *Royal Society open science* 4.8 (2017), p. 170193.

[44] Cecilia Heyes et al. "Knowing ourselves together: The cultural origins of metacognition". In: *Trends in cognitive sciences* 24.5 (2020), pp. 349–362.

[45] Chris D Frith. "The role of metacognition in human social interactions". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1599 (2012), pp. 2213–2223.

[46] Bahador Bahrami et al. "What failure in collective decision-making tells us about metacognition". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1594 (2012), pp. 1350–1365.

[47] Bahador Bahrami et al. "Optimally interacting minds". In: *Science* 329.5995 (2010), pp. 1081–1085.

[48] Bahador Bahrami et al. “Together, slowly but surely: the role of social interaction and feedback on the build-up of benefit in collective decision-making.” In: *Journal of Experimental Psychology: Human Perception and Performance* 38.1 (2012), p. 3.

[49] Bahador Bahrami et al. “What failure in collective decision-making tells us about metacognition”. en. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1594 (May 2012), pp. 1350–1365.

[50] Sébastien Massoni and Nicolas Roux. “Optimal group decision: A matter of confidence calibration”. In: *Journal of Mathematical Psychology* 79 (2017), pp. 121–130.

[51] Benedetto De Martino et al. “Confidence in value-based choice”. In: *Nature neuroscience* 16.1 (2013), pp. 105–110.

[52] Ali Mahmoodi et al. “Equality bias impairs collective decision-making across cultures”. In: *Proceedings of the National Academy of Sciences* 112.12 (2015), pp. 3835–3840.

[53] Benedetto De Martino et al. “Confidence in value-based choice”. In: *Nature Neuroscience* 16.1 (2013), pp. 105–110.

[54] Bahador Bahrami et al. “Together, slowly but surely: The role of social interaction and feedback on the build-up of benefit in collective decision-making.” en. In: *Journal of Experimental Psychology: Human Perception and Performance* 38.1 (Feb. 2012), pp. 3–8.

[55] Dan Bang et al. “Does interaction matter? Testing whether a confidence heuristic can replace interaction in collective decision-making”. In: *Consciousness and Cognition* 26 (2014).

[56] Stephen M. Fleming and Nathaniel D. Daw. “Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation.” In: *Psychological Review* 124.1 (2017), pp. 91–114.

[57] Asher Koriat. “When are two heads better than one and why?” In: *Science* 336.6079 (2012), pp. 360–362.

[58] Dan Bang et al. “Does interaction matter? Testing whether a confidence heuristic can replace interaction in collective decision-making”. In: *Consciousness and Cognition* 26 (2014), pp. 13–23.

[59] Ali Mahmoodi et al. “Learning to make collective decisions: the impact of confidence escalation”. In: *PLoS One* 8.12 (2013), e81195.

[60] Ralf HJM Kurvers et al. “Boosting medical diagnostics by pooling independent judgments”. In: *Proceedings of the National Academy of Sciences* 113.31 (2016), pp. 8777–8782.

[61] Saugat Bhattacharyya et al. “Target detection in video feeds with selected dyads and groups assisted by collaborative brain-computer interfaces”. In: *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE. 2019, pp. 159–162.

[62] Douglas Guilbeault, Samuel Woolley, and Joshua Becker. “Probabilistic social learning improves the public’s judgments of news veracity”. In: *Plos One* 16.3 (2021), e0247487.

[63] Eeshan Hasan et al. “Improving Medical Image Decision-Making by Leveraging Metacognitive Processes and Representational Similarity”. In: *Topics in Cognitive Science* 14.2 (2022), pp. 400–413.

[64] Vito Trianni et al. “Hybrid collective intelligence for decision support in complex open-ended domains”. In: *HHAI 2023: Augmenting Human Intellect*. IOS Press, 2023, pp. 124–137.

[65] Ralf HJM Kurvers et al. “Automating hybrid collective intelligence in open-ended medical diagnostics”. In: *Proceedings of the National Academy of Sciences* 120.34 (2023), e2221473120.

[66] Rajeeva Nagarakanti. “Human-AI Collaboration in Financial Services: Augmenting Decision-Making with Cloud-Native Intelligence”. In: *European Journal of Computer Science and Information Technology* 13 (Apr. 2025), pp. 23–41.

[67] Musaab Mohammad Alhaddad. “Artificial intelligence in banking industry: a review on fraud detection, credit management, and document processing”. In: *ResearchBerg Review of Science and Technology* 2.3 (2018), pp. 25–46.

[68] Fatema Tuz Johora et al. “AI Advances: Enhancing Banking Security with Fraud Detection”. In: *2024 First International Conference on Technological Innovations and Advance Computing (TIACOMP)*. 2024, pp. 289–294.

[69] Kathleen M Vogel et al. “The impact of AI on intelligence analysis: tackling issues of collaboration, algorithmic transparency, accountability, and management”. In: *Intelligence and National Security* 36.6 (2021), pp. 827–848.

[70] Liguo Zhao et al. “Artificial intelligence analysis in cyber domain: A review”. In: *International Journal of Distributed Sensor Networks* 18.4 (2022), p. 15501329221084882.

[71] Mark Steyvers et al. “What large language models know and what people think they know”. In: *Nature Machine Intelligence* (2025), pp. 1–11.

[72] Charvi Rastogi et al. “A taxonomy of human and ML strengths in decision-making to investigate human-ML complementarity”. In: *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*. 2023.

[73] Zahra Zahedi and Subbarao Kambhampati. “Human-AI symbiosis: A survey of current approaches”. In: *arXiv preprint arXiv:2103.09990* (2021).

[74] Mark Steyvers and Megan AK Peters. “Metacognition and Uncertainty Communication in Humans and Large Language Models”. In: *arXiv preprint arXiv:2504.14045* (2025).

[75] Heliodoro Tejeda Lemus, Aakriti Kumar, and Mark Steyvers. “An Empirical Investigation of Reliance on AI-Assistance in a Noisy-Image Classification Task”. In: *HHAI2022: Augmenting Human Intellect*. IOS Press, 2022, pp. 225–237.

[76] Mark Steyvers et al. “Bayesian modeling of human-AI complementarity”. In: *Proceedings of the National Academy of Sciences* 119.11 (2022), e2111547119.

[77] Heliodoro Tejeda et al. “AI-assisted decision-making: A cognitive modeling approach to infer latent reliance strategies”. In: *Computational Brain & Behavior* 5.4 (2022), pp. 491–508.

[78] Heliodoro Tejeda Lemus, Aakriti Kumar, and Mark Steyvers. "How displaying AI confidence affects reliance and hybrid human-AI performance". In: *HHAI 2023: Augmenting Human Intellect*. IOS Press, 2023, pp. 234–242.

[79] ZhaoBin Li and Mark Steyvers. "Beyond Accuracy: How AI Metacognitive Sensitivity improves AI-assisted Decision Making". In: *arXiv preprint arXiv:2507.22365* (2025).

[80] Markelle Kelly et al. "Capturing Humans' mental models of AI: An item response theory approach". In: *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*. 2023, pp. 1723–1734.

[81] Zhaobin Li and Mark Steyvers. "The Importance of Metacognitive Sensitivity in Human-AI Decision-Making". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 47. 2025.

[82] Gagan Bansal et al. "Beyond accuracy: The role of mental models in human-AI team performance". In: *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*. 2019.

[83] Gagan Bansal et al. "Does the whole exceed its parts? the effect of AI explanations on complementary team performance". In: *Conference on Human Factors in Computing Systems (CHI)*. 2021.

[84] Gagan Bansal et al. "Is the most accurate ai the best teammate? optimizing ai for teamwork". In: *AAAI Conference on Artificial Intelligence*. Vol. 35. 13. 2021, pp. 11405–11414.

[85] Shray Bansal et al. "A Bayesian Framework for Nash Equilibrium Inference in Human-Robot Parallel Play". In: (June 2020). URL: http://arxiv.org/abs/2006.05729.

[86] Hussein Mozannar et al. "When to show a suggestion? Integrating human feedback in AI-assisted programming". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 9. 2024, pp. 10137–10144.

[87] Gagan Bansal et al. "Challenges in human-agent communication". In: *arXiv preprint arXiv:2412.10380* (2024).

[88] Keiji Nagatani et al. "Emergency response to the nuclear accident at the Fukushima Daiichi Nuclear Power Plants using mobile rescue robots". In: *Journal of Field Robotics* 30.1 (2013), pp. 44–63.

[89] Kwang-Hyun Lee, Usman Mehmood, and Jee-Hwan Ryu. "Development of the human interactive autonomy for the shared teleoperation of mobile robots". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2016.

[90] Ryan Hoque et al. "Fleet-dagger: Interactive robot fleet learning with scalable human supervision". In: *Conference on Robot Learning (CoRL)*. 2023.

[91] Duc-An Nguyen et al. "Model structure identification of hybrid dynamical systems based on unsupervised clustering and variable selection". In: *IFAC-PapersOnLine* 53.2 (2020), pp. 1090–1095.

[92] Changjoo Nam et al. "Models of trust in human control of swarms with varied levels of autonomy". In: *IEEE Transactions on Human-Machine Systems* 50.3 (2019), pp. 194–204.

[93] Lauren Milliken and Geoffrey A Hollinger. "Modeling user expertise for choosing levels of shared autonomy". In: *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2017.

[94] Jessie YC Chen and Michael J Barnes. "Supervisory control of multiple robots: Effects of imperfect automation and individual differences". In: *Human Factors* 54.2 (2012), pp. 157–174.

[95] Yuan Liu et al. "What Affects Human Decision Making in Human–Robot Collaboration?: A Scoping Review". In: *Robotics* 13 (2024).

[96] Kyle B Reed and Michael A Peshkin. "Physical collaboration of human-human and human-robot teams". In: *IEEE Transactions on Haptics* 1.2 (2008), pp. 108–120.

[97] O Reinoso et al. "Mechanisms for collaborative teleoperation with a team of cooperative robots". In: *Industrial Robot: An International Journal* 35.1 (2008), pp. 27–36.

[98] Duc-An Nguyen et al. "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)*. IEEE. 2020, pp. 3714–3720.

[99] Min Chen et al. "The role of trust in decision-making for human robot collaboration". In: *Workshop on Human-Centered Robotics, RSS*. Vol. 3. 2017.

[100] Stefanos Nikolaidis, David Hsu, and Siddhartha Srinivasa. "Human-robot mutual adaptation in collaborative tasks: Models and experiments". In: *The International Journal of Robotics Research* 36.5-7 (2017), pp. 618–634.

[101] Min Chen et al. "Trust-aware decision making for human-robot collaboration: Model learning and planning". In: *ACM Transactions on Human-Robot Interaction (THRI)* 9.2 (2020), pp. 1–23.

[102] Min Chen et al. "Planning with trust for human-robot collaboration". In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 2018.

[103] Valerie Chen et al. "Understanding the role of human intuition on reliance in human-AI decision-making with explanations". In: *Proceedings of the ACM on Human-computer Interaction* 7.CSCW2 (2023), pp. 1–32.

[104] Clarissa Costen et al. "Shared autonomy systems with stochastic operator models". In: International Joint Conferences on Artificial Intelligence Organization. 2022.

[105] Clarissa Costen et al. "Planning with hidden parameter polynomial MDPs". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 10. 2023, pp. 11963–11971.

[106] Clarissa Costen et al. "Mixed observability MDPs for shared autonomy with uncertain human behaviour". In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization. 2021.

[107] Matthias Scheutz, Scott A DeLoach, and Julie A Adams. "A framework for developing and using shared mental models in human-agent teams". In: *Journal of Cognitive Engineering and Decision Making* 11.3 (2017), pp. 203–224.

[108] Phillip Richter, Heiko Wersing, and Anna-Lisa Vollmer. "Improving Human-Robot Teaching by Quantifying and Reducing Mental Model Mismatch". In: *arXiv preprint arXiv:2501.04755* (2025).

[109] Aaquib Tabrez, Matthew B Luebbers, and Bradley Hayes. "A survey of mental modeling techniques in human–robot teaming". In: *Current Robotics Reports* 1.4 (2020), pp. 259–267.

[110] Stefanie Tellex et al. "Asking for help using inverse semantics". In: (2014).

[111] Aaquib Tabrez, Shivendra Agrawal, and Bradley Hayes. "Explanation-based reward coaching to improve human performance via reinforcement learning". In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2019, pp. 249–257.

[112] Aaquib Tabrez and Bradley Hayes. "Improving human-robot interaction through explainable reinforcement learning". In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2019, pp. 751–753.

[113] Matthew Arnold et al. "FactSheets: Increasing trust in AI services through supplier's declarations of conformity". In: *IBM Journal of Research and Development* 63.4/5 (2019), pp. 6–1.

[114] Tathagata Chakraborti et al. "Plan explanations as model reconciliation–an empirical study". In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Ieee. 2019, pp. 258–266.

[115] Stefan-Dan Ciocirlan, Roxana Agrigoroaie, and Adriana Tapus. "Human-robot team: effects of communication in analyzing trust". In: *2019 28th IEEE international conference on robot and human interactive communication (RO-MAN)*. IEEE. 2019, pp. 1–7.

[116] Zahra Zahedi et al. "Towards understanding user preferences for explanation types in model reconciliation". In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2019, pp. 648–649.

[117] Aditya Yedetore et al. "How poor is the stimulus? Evaluating hierarchical generalization in neural networks trained on child-directed speech". In: *arXiv preprint arXiv:2301.11462* (2023).

[118] Tatsuki Kuribayashi et al. "Large language models are human-like internally". In: *arXiv preprint arXiv:2502.01615* (2025).

[119] Duc-An Nguyen et al. "Joint Decision-Making in Robot Teleoperation: When are Two Heads Better Than One?" In: *arXiv preprint arXiv:2503.15510* (2025).

[120] Talha Siddique et al. "A survey of uncertainty quantification in machine learning for space weather prediction". In: *Geosciences* 12.1 (2022), p. 27.

[121] Soyed Tuhin Ahmed et al. "Scale-dropout: Estimating uncertainty in deep neural networks using stochastic scale". In: *arXiv preprint arXiv:2311.15816* (2023).

[122] Jakob Gawlikowski et al. "A survey of uncertainty in deep neural networks". In: *Artificial Intelligence Review* 56.Suppl 1 (2023), pp. 1513–1589.

[123] Alex Kendall and Yarin Gal. "What uncertainties do we need in bayesian deep learning for computer vision?" In: *Advances in neural information processing systems* 30 (2017).

[124] MMA Valiuddin et al. "A Review of Bayesian Uncertainty Quantification in Deep Probabilistic Image Segmentation". In: *arXiv preprint arXiv:2411.16370* (2024).

[125] Theodore Papamarkou et al. "Position: Bayesian deep learning is needed in the age of large-scale AI". In: *arXiv preprint arXiv:2402.00809* (2024).

[126] Abdullah A Abdullah, Masoud M Hassan, and Yaseen T Mustafa. "Leveraging Bayesian deep learning and ensemble methods for uncertainty quantification in image classification: A ranking-based approach". In: *Heliyon* 10.2 (2024).

[127] Sunday Ochella, Fateme Dinmohammadi, and Mahmood Shafiee. "Bayesian neural networks for uncertainty quantification in remaining useful life prediction of systems with sensor monitoring". In: *Advances in Mechanical Engineering* 16.7 (2024), p. 16878132241239802.

[128] Wesley J Maddox et al. "A simple baseline for bayesian uncertainty in deep learning". In: *Advances in neural information processing systems* 32 (2019).

[129] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles". In: *Advances in neural information processing systems* 30 (2017).

[130] Muhammad Bilal Shahid et al. "Uncertainty Quantification using Deep Ensembles for Decision Making in Cyber-Physical-Human Systems". In: *2024 AIAA Science and Technology Forum and Exposition (AIAA SciTech Forum)*. 2024.

[131] Sunwoong Yang and Kwanjung Yee. "Towards reliable uncertainty quantification via deep ensemble in multi-output regression task". In: *Engineering Applications of Artificial Intelligence* 132 (2024), p. 107871.

[132] Antoine de Mathelin et al. "Deep Out-of-Distribution Uncertainty Quantification via Weight Entropy Maximization". In: *Journal of Machine Learning Research* 26.4 (2025), pp. 1–68.

[133] Luhuan Wu and Sinead A Williamson. "Posterior uncertainty quantification in neural networks using data augmentation". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2024, pp. 3376–3384.

[134] Yarin Gal and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning". In: *international conference on machine learning*. PMLR. 2016, pp. 1050–1059.

[135] Vladimir Puzyrev and Paul Duuring. "Uncertainty quantification of geochemical data imputation using Monte Carlo dropout". In: *Journal of Geochemical Exploration* 272 (2025), p. 107695.

[136] Mehedi Hasan et al. "Controlled dropout for uncertainty estimation". In: *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE. 2023, pp. 973–980.

[137] Hamzeh Asgharnezhad et al. "Enhancing Monte Carlo Dropout Performance for Uncertainty Quantification". In: *arXiv preprint arXiv:2505.15671* (2025).

[138] Daily Milanés-Hermosilla et al. "Monte carlo dropout for uncertainty estimation and motor imagery classification". In: *Sensors* 21.21 (2021), p. 7241.

[139] Doyeon Lee et al. “Metacognitive sensitivity: The key to calibrating trust and optimal decision making with AI”. In: *PNAS nexus* 4.5 (2025), pgaf133.

[140] Jürgen Brandstetter et al. “A peer pressure experiment: Recreation of the Asch conformity experiment with robots”. In: *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2014, pp. 1335–1340.

[141] Payel Kundu and Denise Dellarosa Cummins. “Morality and conformity: The Asch paradigm applied to moral decisions”. In: *Social influence* 8.4 (2013), pp. 268–279.

[142] Axel Franzen and Sebastian Mader. “The power of social influence: A replication and extension of the Asch experiment”. In: *Plos one* 18.11 (2023), e0294325.

[143] Brendan Conway-Smith and Robert L West. “Metacognition in HCI: Designing Systems for Planning and Flexibility”. In: *International Conference on Human-Computer Interaction*. Springer. 2025, pp. 176–187.

[144] Athanasios Drigas, Eleni Mitsea, and Charalabos Skianis. “Metamemory: Metacognitive strategies for improved memory operations and the role of VR and mobiles”. In: *Behavioral Sciences* 12.11 (2022), p. 450.

[145] Kit S Double and Damian P Birney. “Reactivity to measures of metacognition”. In: *Frontiers in Psychology* 10 (2019), p. 2755.

[146] Mohammad Naiseh et al. “Explainable recommendation: when design meets trust calibration”. In: *World Wide Web* 24.5 (2021), pp. 1857–1884.

[147] Benjamin Baird et al. “Regional white matter variation associated with domain-specific metacognitive accuracy”. In: *Journal of cognitive neuroscience* 27.3 (2015), pp. 440–452.

[148] Kai Xue, Medha Shekhar, and Dobromir Rahnev. “A novel behavioral paradigm reveals the nature of confidence computation in multi-alternative perceptual decision making”. In: *Research Square* (2024), rs–3.

[149] Sucharit Katyal and Stephen M Fleming. “The future of metacognition research: Balancing construct breadth with measurement rigor”. In: *Cortex* 171 (2024), pp. 223–234.

[150] Dobromir Rahnev. “A comprehensive assessment of current methods for measuring metacognition”. In: *Nature Communications* 16.1 (2025), p. 701.

[151] Dan Bang et al. “Confidence matching in group decision-making”. en. In: *Nature Human Behaviour* 1.6 (2017), p. 0117.

[152] Arshia Arif et al. “Neurophysiological Approach for Psychological Safety: Enhancing Mental Health in Human–Robot Collaboration in Smart Manufacturing Setups Using Neuroimaging”. In: *Information* 15.10 (2024), p. 640.

[153] Kristin E Schaefer. “Measuring trust in human robot interactions: Development of the “trust perception scale-HRI””. In: *Robust intelligence and trust in autonomous systems*. Springer, 2016, pp. 191–218.

[154] Robert Rousseau et al. “The role of metacognition in the relationship between objective and subjective measures of situation awareness”. In: *Theoretical Issues in Ergonomics Science* 11.1-2 (2010), pp. 119–130.

[155] Hafiza Erum Manzoor and Volha Petukhova. "What Is Going through Your Mind? Metacognitive Events Classification in Human-Agent Interactions". In: *Proceedings of the 18th Joint ACL-ISO Workshop on Interoperable Semantic Annotation within LREC2022*. 2022, pp. 1–9.

[156] Kazuo Okamura and Seiji Yamada. "Adaptive trust calibration for human-AI collaboration". In: *Plos one* 15.2 (2020), e0229132.

[157] Celal Savur and Ferat Sahin. "Survey on Physiological Computing in Human–Robot Collaboration". In: *Machines* 11.5 (2023), p. 536.

[158] Raphaëlle N Roy et al. "How can physiological computing benefit human-robot interaction?" In: *Robotics* 9.4 (2020), p. 100.

[159] Annika Boldt and Nick Yeung. "Shared neural markers of decision confidence and error detection". In: *Journal of Neuroscience* 35.8 (2015), pp. 3478–3484.

[160] Cody Canning and Matthias Scheutz. "Functional near-infrared spectroscopy in human-robot interaction". In: *Journal of Human-Robot Interaction* 2.3 (2013), pp. 62–84.

[161] Leanne M Hirshfield et al. "Toward workload-based adaptive automation: The utility of fNIRS for measuring load in multiple resources in the brain". In: *International Journal of Human–Computer Interaction* 40.22 (2024), pp. 7404–7430.

[162] Megan Strait and Matthias Scheutz. "What we can and cannot (yet) do with functional near infrared spectroscopy". In: *Frontiers in neuroscience* 8 (2014), p. 117.

[163] Melisa Stevanovic et al. "Physiological responses to proposals during dyadic decision-making conversations". In: *PLoS One* 16.1 (2021), e0244929.

[164] Kristina G Baumgart et al. "Neurophysiological correlates of collective perceptual decision-making". In: *European Journal of Neuroscience* 51.7 (2020), pp. 1676–1696.

[165] Nick Hawes et al. "The strands project: Long-term autonomy in everyday environments". In: *IEEE Robotics & Automation Magazine* 24.3 (2017), pp. 146–156.

[166] Matthew Budd et al. "Bayesian reinforcement learning for single-episode missions in partially unknown environments". In: *Conference on Robot Learning (CoRL)*. 2023.

[167] Manolis Chiou, Nick Hawes, and Rustam Stolkin. "Mixed-initiative variable autonomy for remotely operated mobile robots". In: *ACM Transactions on Human-Robot Interaction (THRI)* 10.4 (2021), pp. 1–34.

[168] Javier Corredor, Jorge Sofrony, and Angelika Peer. "Decision-making model for adaptive impedance control of teleoperation systems". In: *IEEE Transactions on Haptics* 10.1 (2016), pp. 5–16.

[169] Jian Shen et al. "A collaborative-shared control system with safe obstacle avoidance capability". In: *IEEE Conference on Robotics, Automation and Mechatronics (ICARM)*. Vol. 1. IEEE. 2004, pp. 119–123.

[170] J Gregory Trafton et al. “Enabling effective human-robot interaction using perspective-taking in robots”. In: *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 35.4 (2005), pp. 460–470.

[171] Andrea Lockerd and Cynthia Breazeal. “Tutelage and socially guided robot learning”. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2004.

[172] Guy Hoffman and Cynthia Breazeal. “Cost-based anticipatory action selection for human–robot fluency”. In: *IEEE Transactions on Robotics* 23.5 (2007), pp. 952–961.

[173] Kenji Sakita et al. “Flexible cooperation between human and robot by interpreting human intention from gaze information”. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2004.

[174] Quanyi Li, Zhenghao Peng, and Bolei Zhou. “Efficient Learning of Safe Driving Policy Via Human-Ai Copilot Optimization”. In: arXiv:2202.10341 (2022). URL: http://arxiv.org/abs/2202.10341.

[175] HCCH Levitt. “Transformed up-down methods in psychoacoustics”. In: *The Journal of the Acoustical society of America* 49.2B (1971), pp. 467–477.

[176] Marion Rouault, Peter Dayan, and Stephen M Fleming. “Forming global estimates of self-performance from local confidence”. In: *Nature communications* 10.1 (2019), p. 1141.

[177] Raunak Bhattacharyya et al. “Towards Intelligent Decision Support Systems in Robotics: Investigating the Role of Self-Confidence Calibration in Joint Decision-Making”. In: *AAAI Spring Symposium Series*. 2024.

[178] Ajith Anil Meera and Pablo Lanillos. “Towards metacognitive robot decision making for tool selection”. In: *International Workshop on Active Inference*. Springer. 2023.

[179] Tjeerd AJ Schoonderwoerd et al. “Design patterns for human-AI co-learning: A wizard-of-Oz evaluation in an urban-search-and-rescue task”. In: *International Journal of Human-Computer Studies* 164 (2022), p. 102831.

[180] Sue E Kase et al. “The future of collaborative human-artificial intelligence decision-making for mission planning”. In: *Frontiers in Psychology* 13 (2022), p. 850628.

[181] Ruben S Verhagen, Mark A Neerincx, and Myrthe L Tielman. “Meaningful human control and variable autonomy in human-robot teams for firefighting”. In: *Frontiers in Robotics and AI* 11 (2024), p. 1323980.

[182] A Hong et al. “Investigating human-robot teams for learning-based semi-autonomous control in urban search and rescue environments”. In: *Journal of Intelligent & Robotic Systems* 94 (2019), pp. 669–686.

[183] Gali Noti and Yiling Chen. “Learning when to advise human decision makers”. In: *International Joint Conference on Artificial Intelligence (IJCAI)*. 2023.

[184] Patrick Hemmer et al. “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.

[185] Leah Chong et al. “Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice”. In: *Computers in Human Behavior* 127 (2022), p. 107018.

[186] Sriraam Natarajan et al. “Human-in-the-loop or AI-in-the-loop? Automate or Collaborate?” In: *AAAI Conference on Artificial Intelligence (AAAI)*. 2025.

[187] Kenny Peng, Nikhil Garg, and Jon Kleinberg. “A no free lunch theorem for human-ai collaboration”. In: *AAAI Conference on Artificial Intelligence (AAAI)*. 2025.

[188] Leila Methnani et al. “Who’s in charge here? A survey on Trustworthy AI in Variable Autonomy Robotic Systems”. In: *ACM computing surveys* 56.7 (2024), pp. 1–32.

[189] Aniketh Ramesh et al. “Experimental evaluation of model predictive mixed-initiative variable autonomy systems applied to human-robot teams”. In: *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 2023.

[190] Maxine T Sherman and Anil K Seth. “Knowing that you know that you know? An extreme-confidence heuristic can lead to above-chance discrimination of metacognitive performance”. In: *Neuroscience of Consciousness* 2024 (2024).

[191] Duc-An Nguyen et al. “Enhancing Joint Human-AI Inference in Robot Missions: A Confidence-Based Approach”. In: *arXiv preprint arXiv:2508.03293* (2025).

[192] Jonathon Love et al. “In human–machine trust, humans rely on a simple averaging strategy”. In: *Cognitive Research: Principles and Implications* (2024).

[193] Qin Ding, Cho-Jui Hsieh, and James Sharpnack. “An efficient algorithm for generalized linear bandit: Online stochastic gradient descent and thompson sampling”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 1585–1593.