UNIVERSITY OF CALIFORNIA,
IRVINE

Human Mental Models of Self, Others, and AI Agents

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Cognitive Science

by

Aakriti Kumar

Dissertation Committee:
Professor Mark Steyvers, Chair
Professor Megan Peters
Professor Padhraic Smyth

2024

DEDICATION

To Abhishek Sharma & Aanjaneya Kumar,
For being my constants in life and academia.

And to Mumma, Paa, & Chaa,
For being the wind behind my sails. I owe it all to you.

Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

Doing a PhD is like navigating a maze that you create with the help of a few others. It has limited rewards and many dead-ends, and demands an inner motivation and resilience that I had to continually muster. I am incredibly grateful to this process for making me a scientist, for showing me the value of collaboration, and most importantly, for giving me a framework to approach life and its puzzles.

Just like it takes a village to raise a child, it also takes a village to support an adult through the rigors of a PhD. I have many villagers to thank for their love, guidance, and support.

My brilliant mentors and collaborators:

Mark Steyvers, for being my guide, for teaching me the art and science of research, for nurturing interdisciplinary ideas, and for making this journey fun. Your passion for research is infectious.

Padhraic Smyth and Megan Peters, for your kindness, continued support, and feedback on my work. Aaron Bornstein, Ramesh Srinivasan, and Jeff Krichmar for your guidance and feedback on previous versions of this work.

Jeff Rouder and Rebecca Martinez, for welcoming me to Irvine and their home. Joachim Vandekerckhove and Holly Westfall, for supporting me throughout the PhD journey and for all the Halloween spirit. Michael Lee, for your wit, wisdom, and all the cricket conversations, and Helen Braithwaite, for all the book recommendations. Ramesh Srinivasan, for your guidance and support, when I most needed it.

Prathap Haridoss and Rahul Marathe, my mentors at IIT Madras, for your wisdom and kindness during my formative years, and for guiding me towards higher education.

Krysta Chauncey, Sara Garver, Kumar Akash, Shashank Mehrotra, and Pradeep Shenoy, for being the most kind and brilliant managers one can have.

My brilliant collaborators— Helio Tejeda, Aaron Benjamin, Trisha Patel, Julia Haaf, Markelle Kelly, Catarina Belem, Padhraic Smyth, Miguel Eckstein, Scott Brown, Sheer Karny, Xinyue Hu, Lukas Mayer — I have learned a lot through our work together and your insights have been invaluable in shaping the research presented in this dissertation.

Other amazing members of the department, Xiaojue Zhou, Paulina Silva, Kexin Chen, Jinwei Xing, Jeff Coon, Bobby Thomas, Manuel Villareal Ulloa, Brandon Hackney, Nora Bradford, Ari Khoudary, Fayette Klassen, Pele Schramm, Alexander Bower, and Arseny Moskvichev for your friendship and for always inspiring me. And to Clara Schultheiss and MyLee Mary Ryan-Glass, for keeping the department going and for all your help.

My many wonderful friends:

Isha, my wine and whine partner from Day 1 at UCI, I hope life continues to guide us to the same cities as it has all these years.

Priyan Das and Helio Tejeda, for being my closest confidants in the lab. Nora Harhen, for being the nicest roommate and friend. Nidhi Banavar, my bro, for all your wisdom and laughter. Adriana Felisa Cha vez De la Pena and Jaime Islas Farias, for welcoming me into your home whenever I needed one. Lauren Montgomery, for your brilliant puns and boba recommendations.

Emily Sumner, for being a selfless mentor and helping me navigate the industry maze. I promise to pay it forward. Alex Etz, for all the student's tea.

Cheeku, Reddy, Harsha, Swar, Priyan, Sanket, for making Boston feel like home.

Shruti, Chetana, Divyasree, Raveena, Sripriya, and Neelima, my dearest misfits, for all the love, laughter, and friendship.

Sehr, for being my constant in an ever-changing world.

Piyush, Faizan, and Snehal, for all the star-gazing, NFTs, and poetry.

Tamhane, for your craziness. And Himanshu, for all the music.

And my beautiful family:

Vasundhara, Kritika, and Tanisha, for being my lifetime source of love and inspiration, and for reminding me what crazy work hours actually look like. Chayan and Prateek, for your humor and friendship. Ilu, Kirti, Chandni Di, Chandan Jiju, Chayanika Di, Nishant Zeezoo, Kaustubh Bhaiya, Shantala Bhabhi, Manik, Madhav, Aadi, Parth, and Mahek for their unconditional love and frienship.

Nana for being my first source of awe and inspiration. I think you would be proud. And Dada, we now have another psychologist in the family. My extended family back home - Mummy, Papa, all my Mousis, Mamus, Mamis, Chacha, Fua - who cheered me on from India.

Rekhu for impressing upon me the value of education from a very young age. Paa for teaching me the value of discipline and hard work. Chaa for introducing me to the wonders and rigors of science. And to all three of them for always believing in my abilities.

Aanjaneya, for being my friend since the day he was born, and for being a constant source of support and WhatsApp stickers.

Abhishek, for being my partner in life and this PhD. 19-year-old us would not have imagined a future of sitting together and talking research over chai. To many more years of fruitful collaboration!

The PhD journey is unique—it is both a personal quest and a collaborative adventure. It has taught me how to be my own friend and collaborator, and for that, I am deeply grateful. It has also shown me the value of having your community, and to each person who has been a part of my journey and woven into the fabric of my story, thank you for being my village.

VITA

Aakriti Kumar

EDUCATION

Doctor of Philosophy in Cognitive Science — **2024 (expected)**
University of California, Irvine — *Irvine, California, USA*

Masters in Statistics — **2022**
University of California, Irvine — *Irvine, California, USA*

Bachelor of Technology in Engineering — **2015**
Indian Institute of Technology, Madras — *Chennai, Tamil Nadu, India*

RESEARCH EXPERIENCE

Graduate Research, MADLAB, UC Irvine
Jan'20–Mar'24 | Advisor: Prof. Mark Steyvers

Graduate Research, Structure in Perception and Cognition Lab, UC Irvine
Sep'18–Dec'20 | Advisor: Prof. Jeffrey Rouder

INDUSTRY EXPERIENCE

Human-Computer Interaction Research Intern, Motional
Apr–Dec, 2023

Human Behavior Modeling Intern, Honda Research Institute
Jun–Sep, 2022

Statistical Consultant, Google Research, India
Jan–Mar, 2021

REFEREED JOURNAL PUBLICATIONS

[1] **A. Kumar**, P. Smyth, M. Steyvers (2023). Differentiating mental models of self and others: A hierarchical framework for knowledge assessment. *Psychological Review*

[2] M. Steyvers & **A. Kumar**, (2023). Three Challenges for AI-Assisted Decision-Making. *Perspectives on Psychological Science*

[3] H. Lemus, **A. Kumar**, P. Smyth, M. Steyvers (2022). AI-assisted Decision-Making: A Cognitive Modeling Approach to Infer Latent Reliance Strategies. *Computational Brain & Behavior*

[4] **A. Kumar**, A. Benjamin, A. Heathcote, M. Steyvers (2021). Comparing models of learning and relearning in large-scale cognitive training data sets. *NPJ Science of Learning.*

[5] J. Rouder, **A. Kumar**, J. Haaf (2023). Why Most Studies of Individual Differences With Inhibition Tasks Are Bound To Fail. *Psych Bulletin & Review*

REFEREED CONFERENCE PUBLICATIONS

[C1] **A. Kumar**, T. Patel, A. Benjamin, M. Steyvers (2021). Explaining Algorithm Aversion with Metacognitive Bandits. *Cognitive Science (Vol. 43)*

[C2] **A. Kumar***, S. Chatterjee*[1], P. Shenoy (2022). Meta-Learning of Dynamic Policy Adjustments in Inhibitory Control Tasks. *Cognitive Science (Vol. 44).*

[C3] H. Lemus, **A. Kumar**, M. Steyvers (2022). An Empirical Investigation of Reliance on AI-Assistance in a Noisy-Image Classification Task. *Proceedings of the First International Conference on Hybrid Human-Machine Intelligence.*

[C4] H. Lemus, **A. Kumar**, M. Steyvers (2022). How Displaying AI Confidence Affects Reliance and Hybrid Human-AI Performance. *Proceedings of the Second International Conference on Hybrid Human-Machine Intelligence.*

[C5] **A. Kumar**, K. Akash, S. Mehrotra, T. Misu, M. Steyvers (2023). When Do Drivers Intervene In Autonomous Driving?. *Human Robot Interaction (Late-Breaking Report) 2023*

[C6] M. Kelly, **A. Kumar**, P. Smyth, M. Steyvers (2023). Capturing Humans' Mental Models of AI: An Item Response Theory Approach. *FAccT 2023*

PAPERS IN PROGRESS

[P1]**A. Kumar***, R. Tham*, M. Steyvers. Assessing the Impact of Differing Perspectives in Advice-Taking Behavior. *Under Review in Decision*

[P2] **A. Kumar**, P. Petrides, K. Chauncey. Impact of System Delays on User Stress & Behavioral Intention: A Mixed Methods Approach.

[P3] M. Steyvers, H. Tejeda, **A. Kumar**, S. Karny, X. Hu, L. Mayer, P. Smyth. The Calibration Gap between Model and Human Confidence in Large Language Models. *Under Review in FAccT 2024*

TEACHING EXPERIENCE

Teaching Assistant, Probability and Statistics in Psychology I *Fall 2019*

[1]* indicates equal contribution

Teaching Assistant, Probability and Statistics in Psychology II,	*Winter 2020*
Teaching Assistant, Probability and Statistics in Psychology III	*Spring 2020*
Lead Teaching Assistant, Introduction to Human Memory	*Winter 2019*
Teaching Assistant, Introduction to Psychology	*Fall 2018*

ABSTRACT OF THE DISSERTATION

Human Mental Models
of Self, Others, and AI Agents

By

Aakriti Kumar

Doctor of Philosophy in Cognitive Science

University of California, Irvine, 2024

Professor Mark Steyvers, Chair

Collaboration with other agents requires people to reflect on several aspects of their collaborator's capabilities: How good are they at a task? Do they have access to the same information as I do? How useful is their advice? This dissertation takes a closer look at how people answer these questions while interacting with other humans and more importantly, with AI agents. First, we discuss the role of mental models in facilitating people's interactions with other agents. In Chapter 2, we present a computational framework to understand how people assess the ability of other humans and AI agents. Our research reveals a discrepancy in people's assessments: individuals accurately gauge another human's ability on a task, but consistently overestimate an AI agent's performance on the same task. In Chapter 3, we examine how people navigate advice when a human or an AI advisor has access to different information than they do. We find that while people take into account this difference in information, they consistently underestimate the value of advice. Finally, in Chapter 4, using two case studies we demonstrate the utility of cognitive modeling in inferring people's latent reliance on AI. The first case study models people's decision to accept advice, and the second case study models people's decision to intervene in the AI's course of action. We conclude by discussing avenues for future work. Altogether, this dissertation uses a cognitive science lens to improve our understanding of human-AI interaction, highlighting the role of mental

models and drawing valuable lessons from human-human interaction.

Chapter 1

Introduction

In recent years, the landscape of technology has undergone a radical transformation. The term "technology" no longer evokes images of bulky desktop computers or industrial machinery. Instead, it is a spectrum from predictive models aiding experts in medical and judicial decisions, to chatbots like ChatGPT that respond to user queries in natural language. The broad umbrella term for these predictive and generative innovations is Artificial Intelligence (AI). Today's AI has human-like capabilities, is widely accessible, and is evolving rapidly. Simple tasks can now be automated by digital assistants like Siri, Alexa, and ChatGPT. Advanced driver assistance systems (ADAS) aim to improve our driving experience and minimize the occurrence of avoidable accidents. Recommender systems on media platforms supply personalized playlists that include both our favorites and new content we might enjoy. Such widespread integration of AI into our lives holds the promise of saving human effort and avoiding blind spots in human decisions.

This seismic shift in human-technology interaction, especially with AI assistants [1], warrants a closer investigation of how the human mind perceives and understands AI's capabilities.

[1]We use the terms AI assistant / algorithms / machines as catch-all terms to describe a black-box model that takes in inputs and can output a prediction.

This dissertation adopts a Cognitive Science lens to investigate such human-AI interaction. It explores this theme through online behavioral experiments and the development of cognitive models, offering insights into the nuances of human engagement with AI assistance.

A critical determinant of the effective use of AI assistance is the human's *mental model* of the AI. This mental model encapsulates the human's beliefs about AI as well as expectations concerning the effects of interacting with it. In general, mental models are simplified representations of the world that are constructed by humans to allow them to integrate new information and make predictions while expending little mental effort (Craik, 1952; Smyth et al., 1994). People's mental models of themselves and other humans play a critical role in supporting human-human interactions and have been studied extensively. However, our knowledge about people's mental models of AI and how it differs from their mental models of humans is rather limited. This thesis is a step towards bridging this gap in the context of AI-assisted decision-making. A deeper understanding of people's mental models of AI can facilitate the design of workflows that can aid humans in developing appropriate reliance strategies and consequently lead to improved team performance.

1.1 What are Mental Models?

The notion of mental models was introduced by (Craik, 1952) who described them as "small-scale models of reality that are used to reason, explain, and anticipate events in the world". Mental models did not attract much attention until they were reintroduced by (Johnson-Laird, 1983), who defined them as the basic structure of cognition. Later, (Gentner & Stevens, 1983) use mental models to study the interaction between humans and technological systems. Their work suggests that people develop mental models of themselves, as well as of systems, human or otherwise, with which they interact. (Battaglia et al., 2013) proposes that humans use mental models like an 'intuitive physics engine' to understand the physical structure of

real-world scenes, and to make predictions about the dynamics of objects in these scenes. These models provide people with predictive and explanatory power to understand their interactions with other systems (Norman, 2014). In summary, mental models are simplified representations of the world constructed by humans which allow them to integrate new information and make predictions with little mental effort (Smyth et al., 1994).

1.2 Reflecting on the Self: Metacognition

A human building a mental model of themselves engages in *metacognition*. Metacognition refers to 'cognition about cognition' or more simply 'thinking about thinking'. It supports social interaction by enabling people to reflect on their own cognitive processes (Shimamura, 2000), and explain their actions to others (Frith & Frith, 1999). It is quantified as a combination of two measures: 1) a person's performance on a task, also called a first-order judgment; and 2) a person's assessment of their own performance, usually expressed as the confidence in their first-order judgment. This assessment is called a second-order judgment (Fleming & Lau, 2014).

Although metacognition is not the primary focus of our work, it significantly influences how humans perceive other agents – humans' mental model of themselves influences their mental model of other agents. Chapter 2 demonstrates a strong correlation between humans' perceptions of others' abilities and their self-perceived abilities. This correlation does not hold for AI agents; humans' views of AI capabilities are distinctly separate from their own abilities. Chapter 4 further explores this, revealing that the likelihood of accepting advice from an AI agent hinges on an individual's confidence in their own decision-making.

1.3 Understanding Others: Theory of Mind & Machine

Throughout their lives, humans engage in extensive interactions with others. To effectively collaborate with others, humans use *theory of mind* (ToM). ToM, or mentalizing, can be described as a form of metacognition. In particular, it is a form of *explicit metacognition* that requires inferring the reasons underlying behavior of another agent by observing their actions. Traditional accounts of ToM highlight the importance of understanding the other agent's intentions as a key component to understanding their 'mind' (Frith & Frith, 1999).

Humans are adept at planning based on beliefs, goals, and resource constraints (Baker et al., 2009; Gopnik & Meltzoff, 1997; Lieder & Griffiths, 2020). They can do inverse-planning to infer beliefs and goals from the observed behavior of other agents (Shum et al., 2019; Tauber & Steyvers, 2011). However, to collaborate with other agents on knowledge-based tasks, it suffices to build mental models of the agents' abilities as opposed to modeling their beliefs and goals. This is the setting we focus on. In Chapter 2, we present a model of a human's assessment of another agent's performance on a classification task. In Chapter 3, we empirically investigate a human's assessment of another agent's performance on a task, but now the human and the other agent have different information available while making decisions.

Interacting with an AI agent requires a similar cognitive process called *theory of machine* analogous to theory of mind (Logg, 2017). Like theory of mind, theory of machine requires people to think about the internal processes of an AI agent. Recent work in the human-computer interaction community has been focused on empirical investigations of theory of machine. However, the correspondence with theory of mind makes it clear that cognitive science can offer important insights on theory of machine and human-AI collaboration in general. These parallels motivate our investigation of how humans build mental models of other agents.

1.4 Understanding Artificial Agents: People's Mental Models of AI

Several studies have investigated people's prior beliefs where participants are asked how they *would* use AI advice in various hypothetical scenarios. The results are strongly dependent on the framing of the scenario including the task domain, the amount of information that is provided about AI performance, as well as individual differences (Abraham et al., 2017; Bigman & Gray, 2018; Castelo et al., 2019; Lubars & Tan, 2019). People tend to prefer to rely on humans over AI in highly consequential scenarios (Castelo et al., 2019), especially when considering hypothetical moral scenarios where life and death hang in the balance (Bigman & Gray, 2018). For tasks with a high perceived degree of objectivity (e.g. involving quantifiable facts as opposed to personal opinion and intuition), human preference shifts to AI (Castelo et al., 2019). In several low-stakes quantitative tasks such as estimating the weight of a person from a photograph or predicting the popularity of songs, people prefer to take advice from algorithms over other humans (Logg et al., 2019). In addition, people's preference to use AI increases when performance data about the AI is provided (Castelo et al., 2019). The willingness of people to consider using automation also depends on demographic factors. For example, younger users are more willing to use automation in vehicles (Abraham et al., 2017). Understanding these preferences for task delegability and expectations about AI performance is important as they might impact the willingness to accept AI advice when people do interact with AI decision support systems.

Another set of studies has investigated people's mental models of AI after some initial exposure to the AI. At first glance, these experiments seem to present a mixed picture of people's understanding of the AI and the effectiveness of their reliance decisions. For example, (Dietvorst et al., 2015) showed that participants prefer to rely on human decision-making over algorithms once they see the performance of an algorithm, which includes instances where

the algorithm makes mistakes, even though the algorithm actually outperforms the human decision-makers on average. This result has been interpreted to suggest that experience with the AI, especially exposure to errors made by the AI leads to *algorithm aversion*, presumably because people expect algorithms to perform better than they actually do (e.g., see (Burton et al., 2020) for an overview). However, there is an important limitation of these experimental studies. While individuals did become familiar with the algorithm's performance, they were asked only *once* about a delegation decision and they did not become familiar with the consequences of that delegation decision. Therefore, these results cannot be used to answer questions about whether humans use AI advice *selectively*.

In contrast, in recent studies (Liang et al., 2022; Tejeda et al., 2022), participants are provided numerous opportunities to make reliance decisions and allow the decision-maker to selectively use algorithmic advice. These experiments have not confirmed general algorithm aversion. Instead, the results by (Tejeda et al., 2022) show that participants adopt a flexible reliance strategy where reliance depends on the decision-makers own confidence state, the confidence expressed by the AI, and the overall AI performance. In addition, the results showed that the reliance strategy is effective and does not differ substantially from optimal reliance strategies. Other studies have found that people can take into account the accuracy of algorithmic advice (Liang et al., 2022; Yin et al., 2019). Interestingly, even without any feedback about accuracy, it is still possible for people to adjust their reliance on AI (Lu & Yin, 2021) as people can use the trials where they are very confident about their own performance to assess AI performance.

Overall, empirical results suggest that people's mental models of AI depend on the degree of familiarity with the AI as well as the degree of familiarity with the outcomes of their reliance decisions. It is possible that people who are somewhat familiar with the AI's performance but not with the consequences of decisions to delegate or rely on AI advice might have an incomplete mental model and might not accurately represent the differential capabilities of the

AI relative to oneself. Perhaps their mental assessment of the AI is (correctly) downgraded after exposure to inevitable AI errors, but is not correctly reflecting the fact that they might not fare any better on the same problem, and in fact might perform even worse. However, the results on studies where people gain familiarity with the consequences of their reliance decisions suggest that people develop richer mental models of the AI that allow for flexibility in depending on their own decision or the AI.

Other factors such as the complexity of the AI and decision-making task likely impact the mental model fidelity as well. In some laboratory tasks, relatively simple behavioral tasks are used that might not require much learning to develop effective reliance strategies. However, in complex industrial systems or military applications with higher levels of automation, the decision-maker might not fully understand how the system works, and the decision-maker might default to simplistic strategies such as indiscriminately relying on AI (Cummings, 2017).

1.5 The Role of Mental Models in Shaping AI-Assisted Decision-Making

In line with the old adage, "two minds are better than one", collaborative human-AI decision-making is expected to increase the efficacy of decisions. The more accurate a human's mental model of an AI, the more likely it is that the human will use AI assistance appropriately (Bansal et al., 2019). Similarly, the ineffective use of AI might be driven by incomplete and/or incorrect mental models of the AI. Such incorrect mental models may lead to inappropriate levels of reliance on or miscalibrated trust in the AI. Recent work shows mixed results: while some studies report that decisions made jointly by the human and AI are more effective than either the human or the AI working independently (Patel et al., 2019; Phillips et al.,

2018; Steyvers et al., 2022; Wright et al., 2017), other studies highlight humans' sub-optimal use of AI-advice and explanations (Bansal et al., 2021; Tan et al., 2018; Y. Zhang et al., 2020). Many empirical investigations of joint human-AI decision-making have indicated that humans are susceptible to biases and errors when working with AI assistance which reflects poor mental models of AI. People may over- or under-rely on the AI's advice leading to sup-optimal performance. While there is a lot to be done to improve AI accuracy, reliability, explainability, and so on, there is also a concurrent need to understand and quantify how humans work with AI teammates. Humans assisted by AI must have a good understanding of the AI agent's ability and knowledge (National Academies of Sciences & Medicine, 2021). An in-depth understanding of human-human interaction can serve as a starting point for studying human-computer interaction. It would enable us to understand the effect of AI-assistance in a principled way and make predictions about human behavior.

1.6 Overview

There is considerable evidence suggesting that to effectively collaborate with another agent, it is essential to develop an accurate mental model of their knowledge and abilities. When collaborating with others, people must simultaneously build mental models—both of their own expertise at the task (metacognition) and of the collaborator's performance on the task (ToM). Metacognition enables them to look inward and understand their own abilities and shortcomings in performing a task. While theory of mind enables them to use their own knowledge of the task—and the world in general—to build mental models of other agents' knowledge, beliefs, goals, and emotions. These mental models are then continuously updated as information about performance—theirs and of the collaborator—is revealed.

This dissertation combines empirical research and computational modeling to better understand people's mental models of other humans and AI agents. It is divided into an

introduction, 3 main chapters, and a conclusion.

Chapter 2 presents a framework that allows us to understand people's assessment of other agents - human or AI. Additionally, through an image classification task and a trivia task, we empirically investigate people's ability to predict their own performance and the performance of another human or AI. Using this experimental data, we demonstrate the use of our framework for examining research questions about people's perceptions of AI agents and other people. The results presented in Chapter 2 have been peer-reviewed and published. The knowledge assessment framework was first published as *'Differentiating mental models of self and others: A hierarchical framework for knowledge assessment'* in the journal *Psychological Review*. Extension of the knowledge assessment paradigm to humans assessing AI agents was published as *'Capturing Humans' Mental Models of AI: An Item Response Theory Approach'* in the *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. This work was led by M. Kelly.

In Chapter 3, the focus shifts to advice-taking scenarios characterized by asymmetric information. It examines whether individuals can appropriately vet the value of advice from another person or an AI, especially when their advisor has different information compared to them. This is assessed via a counting task where participants are asked to estimate the number of objects in a jar. If given the option to estimate the number of objects themselves or to delegate the task to another person or AI with a different view of the jar, do people optimally delegate based on who has the better 'view'? Alternatively, if given the option to change their estimate by integrating information from another person or AI's estimate, do people appropriately integrate the target's estimate into their own estimate? Chapter 3 is currently under review.

Chapter 4 introduces two instances of computational cognitive models that capture how people decide to rely on AI advice across two assisted decision-making contexts. First, we look at people's performance on an assisted image classification task. We demonstrate that our

model's predicted reliance strategies closely track the strategies employed by humans in the experiment. Second, we investigate people's decision to intervene in an autonomous driving simulator. We propose a cognitive model that allows us to quantify and compare the relative perceived risk of different scenarios for the two mobility types. Results presented in Chapter 4 are based on two publications - 1) *'AI-assisted decision-making: A cognitive modeling approach to infer latent reliance strategies'* published in the journal *Computational Brain and Behavior*, and 2) *'When Do Drivers Intervene In Autonomous Driving?'* published in the *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*.

The final chapter outlines the conclusions and discusses avenues for future work. Each chapter is written to stand on its own.

Chapter 2

Capturing Humans' Mental Models of Self, Others' and AI Ability: A Hierarchical Framework for Knowledge Assessment

Abstract

Developing an accurate model of another agent's knowledge is central to communication and cooperation between agents. We propose a hierarchical framework of knowledge assessment that explains how people construct mental models of their own knowledge and the knowledge of others. Our framework posits that people integrate information about their own and others' knowledge via Bayesian inference. To evaluate this claim, we conduct an experiment in which participants repeatedly assess their own performance (a metacognitive task) and the performance of another person (a type of theory of mind task) on the same image classification

tasks. We contrast the hierarchical framework with simpler alternatives that assume different degrees of differentiation between mental models of self and others. Our model accurately captures participants' assessment of their own performance and the performance of others in the task: initially, people rely on their own self-assessment process to reason about the other person's performance, leading to similar self- and other-performance predictions. As more information about the other person's ability becomes available, the mental model for the other person becomes increasingly distinct from the mental model of self. Simulation studies also confirm that our framework explains a wide range of findings about human knowledge assessment of themselves and others.

We also extend this framework to model humans' perception of AI agents. We apply this framework to real-world experiments, in which each participant works alongside another person or an AI agent in a question-answering setting, repeatedly assessing their teammate's performance. Using this experimental data, we demonstrate the use of a multidimensional extension of our framework for testing research questions about people's perceptions of both AI agents and other people. We contrast mental models of AI teammates with those of human teammates. Our results indicate that people expect AI agents' performance to be significantly better on average than the performance of other humans, with less variation across different types of problems.

2.1 Introduction

Understanding and comparing the knowledge states of others to our own knowledge is a fundamental skill that supports social interaction in daily life. Does Akira know what I know? Would Georgina perform better than me on this task? Will this problem be as difficult for Keith as it is for me? Humans constantly make predictions about their abilities at different tasks and how well other people might fare at the same task relative to themselves.

For an individual making predictions about the difficulty of a task for others, a potential starting point is to base it on their own experience with the task (Nickerson, 1999) such as remembering information (Jameson et al., 1993; Koriat & Ackerman, 2010) or solving problems (Kelley & Jacoby, 1996). One's mental model about oneself may often lead to accurate predictions about others. However, previous research has not explored how the mental model of another person can be differentiated to account for specific information learned about them. When we observe another person over time, what is the process by which an initial undifferentiated mental model of that person becomes tailored towards them?

Our research combines ideas from (i) metacognition which includes processes used to draw inferences about one's own knowledge states and (ii) theory of mind (also known as mindreading), which includes processes used to draw inferences about other people's knowledge states. Recent computational perspectives have suggested that reasoning processes about self and others are closely intertwined (Fleming, 2021). For example, a recent model for metacognition has been motivated by considering self-evaluation as a "second-order" computation distinct from simpler first-order accounts in which the same internal state guides decisions and self-evaluation (Fleming & Daw, 2017). Such second-order computation is also required when assessing the knowledge states of other people. Similarly, computational models for mindreading have been motivated by inverse planning – the process by which other people's goals and beliefs are inferred by applying one's own mental model to the observed actions (Aboody et al., 2021; Baker et al., 2017; Baker et al., 2009; Berke & Jara-Ettinger, 2021; Tauber & Steyvers, 2011). Empirical studies have provided increasing support for commonalities between metacognition and theory of mind based on shared cognitive resources (Nicholson et al., 2021), overlapping brain structures (Vaccaro & Fleming, 2018), and overlapping developmental trajectories ((Gopnik & Astington, 1988), (Paulus et al., 2014), but see (Baer et al., 2021)). Taken together, there is substantial evidence for a close correspondence between reasoning about self and others.

We present a hierarchical framework for knowledge assessment that explains how people assess their own knowledge and the knowledge of others. The framework is inspired by the connection between metacognition and theory of mind, and has significant implications for understanding knowledge assessment in general. We focus on the relationship between *self-assessment* (i.e., predicting one's performance on a task) and *other-assessment* (i.e., predicting how well another person performs on the same task). There are two types of empirical results that the hierarchical framework is designed to address. First, the model can be used to explain the relationship between self- and other-assessment in situations where there is a lack of information about the other person being judged. For example, people are asked to assess the percentage of randomly selected students who know the answer to a given question (Nickerson et al., 1987; Tullis, 2018) or their relative placement in a population (Dunning, 2011; Moore & Healy, 2008). These studies have shown that people tend to predict that they are better than others on easy tasks but worse than others on challenging tasks (Moore & Cain, 2007). In these tasks, people consider comparisons to randomly sampled other individuals from a population. In later sections, we show how our framework may be applied to these experimental settings and demonstrate its ability to explain the empirical results observed in the literature. Second, the hierarchical framework also accounts for situations where people learn to make predictions about a *specific* person as information about that person becomes available. Our framework can also explain how people assess a specific other person by observing their performance on a task over time. To test our framework's predictions, we conduct a behavioral experiment where participants classify images and assess their own performance and the performance of a specific other person on this task. This experimental setup allows us to investigate two distinct aspects of assessing others: how individuals assess another individual without any explicit information about the other's ability, and how this assessment changes as information about the other's performance becomes available. We also apply our framework to explain other assessment in paradigms where no information is provided about the other person (Moore & Healy,

2008; Tullis, 2018). Throughout this chapter, we assume that performance is indicative of a person's knowledge or ability. However, our proposed framework could also be applied to other domains that are not related to knowledge. For example, inferring a person's strength when observing them perform specific exercises in a gym, or assessing the skill of drivers by observing them in challenging parking situations.

In the following sections, we provide a detailed overview of our modeling framework. We then present data from a knowledge assessment task in which people assess their own performance and the performance of one other person on an image classification task. We apply our proposed framework and simpler alternative models to this empirical data and demonstrate that the predictions of our hierarchical model closely match the trends observed in the data. We also show how our framework supports other findings in the empirical literature on knowledge assessment. Finally, we discuss the significance and implications of this framework for future research.

2.2 A Hierarchical Framework for Knowledge Assessment

We propose a hierarchical framework for knowledge assessment that describes the computational problem which people solve when assessing themselves or another person. We posit that both self-assessment and other-assessment are inference problems that people solve through Bayesian inference. Figure 2.1 illustrates the different levels of the framework and the graphical model corresponding to it. The central idea underlying our framework is that reasoning about the performance of oneself or another person occurs at three different levels:

1. Population level: The top level corresponds to the population level (ω) which encodes information about the population of individuals to which the self and the other belong.

2. Individual-specific level: The middle level pertains to information about specific people

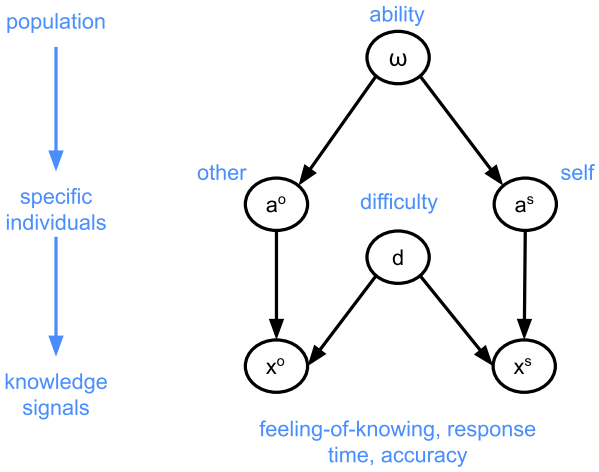


Figure 2.1: Three levels of the hierarchical model used to reason about one's own as well as other people's performance. People may have access to different kinds of knowledge signals such as feeling-of-knowing, response time, and accuracy when assessing their own knowledge or another person's knowledge.

(including self and others) such as the ability of self and other (a^s, a^o), the difficulty of the task perceived by self and other (d).

3. Knowledge-signals level: The bottom-most level concerns knowledge signals (x) which include observed performance outcomes for self and/or others and internal metacognitive signals that people may have access to when doing a task.

We assume that people can reason across the three levels and make inferences about self- or other performance a^s, a^o, as well as task difficulty d using the observed knowledge signals x. To enable reasoning across abilities of people and difficulties of items in tasks, the hierarchical framework adopts concepts from item-response theory (IRT, (Fox, 2010; van der Linden & Hambleton, 2013)) to describe the relationship between x and a_s, a_o, d. Item-response theory has recently been used to model self-assessment (Jansen et al., 2021; Jansen et al., 2020). Similar to the model by (Jansen et al., 2021), we hypothesize that people make errors in their self-assessment such that their predicted performance deviates from the actual performance that would be predicted by an item-response model. Specifically, we assume that people

combine a *subjective* estimate of ability with a *subjective* estimate of task difficulty to estimate the performance on a task.

To support inferences about ability and task difficulty, our work builds on previous research (Koriat, 1997; Moore & Healy, 2008; Nickerson, 1999; Thomas & Jacoby, 2013) which identifies a variety of signals that people use for assessment. In our framework, we assume that people may have access to two kinds of knowledge signals (x) while performing a task. The first kind is based on *external signals*, such as feedback on people's assessment of self or other, information about the correct or optimal solution to a problem, or information about the other's performance. For example, in some tasks, people may receive feedback about their accuracy which could be used as an external signal to infer their ability and predict future performance. The second kind of signals are *internal signals* that arise from reflecting on one's internal metacognitive processing. These include how long it takes people to arrive at a solution (Thomas & Jacoby, 2013; Tullis, 2018), their confidence in their response (Hart, 1965; Leibert & Nelson, 1998; Nelson & Narens, 1980), or their feeling-of-knowing about the problem at hand (Koriat, 2000). We use feeling-of-knowing to refer to the intuition that one may have about being able to solve a problem or answer a question without actually attempting to solve the problem or answer the question (e.g., when reading a general knowledge question, one may feel the question is answerable based on the familiarity with the words in the question).

Knowledge signals allow people to make estimates of individual-specific parameters such as the ability of self and other, and perceived difficulty of the task. Depending on the available signals, our framework suggests two ways in which people may infer the ability of others:

1. In the absence of specific information about others (e.g., the inference is about a randomly sampled person from the population), people may use the knowledge signals regarding their own performance and metacognition (x^s) to reason about the ability of others. This corresponds to inferring $p(a^o|x^s)$.

2. If some information about the other person is available, people may also consider a combination of their own and others' knowledge signals to infer $p(a^o|x^s, x^o)$.

The first inference problem maps directly onto previous research where no information is provided about others (Moore & Healy, 2008; Nickerson, 1999; Tullis, 2018). The second inference problem has not been studied previously. In the next section, we present results from an experimental paradigm where participants track the performance of a specific other person and are provided with an increasing amount of information about the other person's performance. The framework also extends to assessing multiple other people. Note that, in many real-world contexts, people already have an estimate of their own ability on a variety of tasks: they gather information about their ability over time through varied interactions with other agents and environments. Hence, a^s may be partially or fully observed in these cases. In comparison, people typically have less information about other people's abilities. Therefore, in most cases, a^o is unobserved and must be inferred. As a result, people's assessment of their own abilities and knowledge will be less noisy than their assessment of others (Moore & Healy, 2008).

People must also reason about the task at hand when doing self- or other-assessment. External signals such as accuracy may enable people to better assess the difficulty (d) of the task at hand. Internal signals such as the time it takes people to solve a problem may provide additional information about the difficulty of the task and help predict how others would fare at the same task. For example, people may infer that questions that take them longer to answer are more difficult, and may take others longer to answer as well. Together, these internal and external signals provide information that people may use to infer task difficulty (Kelley & Jacoby, 1996).

The top-level of the hierarchy formalizes the assumption that any person's ability, including one's own, is a sample from a population-ability distribution which is denoted by ω. Note

that ω may vary across tasks and population composition. Consider a Chemistry teacher who is about to begin teaching a lesson on stoichiometry to a group of students who have never studied it. She has however observed other students of the same grade in the past, and can easily make inferences about how well the new batch of students might fare on a test before and after her lesson. This is because the teacher assumes that any new student may be considered a random sample from the population of all students. She would also have a reasonable understanding of what questions the students might find difficult. On the other hand, if asked to compare her own knowledge of stoichiometry to another Chemistry teacher, she would think about the population of Chemistry teachers (which also includes herself) and her placement in this population. Therefore, people's assessment of the ability of others starts with assumptions about the population they are evaluating. In this work, we focus on people's assessment of others from the same population as themselves. However, it is straightforward to extend our framework to model how people assess individuals from different populations or even artificial agents. One way to do this is to add another level to the current hierarchy: two populations may be considered samples from a super-population of agents.

2.2.1 Three Instantiations of the Hierarchical Framework

Within this hierarchical approach to knowledge assessment, we explore three classes of models for connecting the subjective estimates of self and other as illustrated in Figure 2.2. These models correspond to different substantive assumptions about the psychological process of other assessment in terms of the assumed connections between the different layers of the hierarchy. The first instantiation, *differentiated by ability* is equivalent to the full hierarchical model. The second instantiation, the *fully differentiated* model, assumes that self- and other-assessment are distinct processes. The *undifferentiated* model assumes no distinction between self- and other-assessment. We will also refer to these models with the short-hand

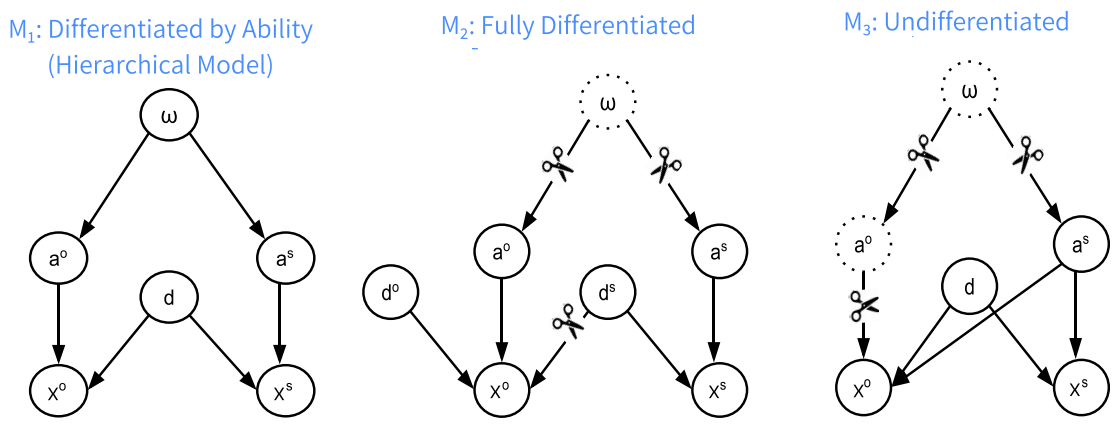


Figure 2.2: Schematic graphical models connecting the subjective estimates of self and other, corresponding to different substantive assumptions about the psychological process of other assessment: 1) Differentiated by Ability model (M_1) which is equivalent to the full hierarchical model, 2) Fully Differentiated model (M_2) which ignores population level information, and 3) Undifferentiated model (M_3) which ignores the individual-specific level of the full framework.

notation M_1, M_2, and M_3 respectively.

Differentiated by Ability Model (M_1)

This model maps directly to the proposed hierarchical model of knowledge assessment. One way to formalize the reasoning process in this model is that people separately assess their own ability (a^s) and the ability of another person (a^o). However, because the hierarchical structure imposes connections between the self and other ability (e.g., with no knowledge of the other person, the best estimate of another person equals that one of one's own ability, $a^o = a^s$), it is conceptually convenient to assume that people evaluate the ability of others relative to their own abilities. Specifically, $\delta = a^o - a^s$ captures the *differential ability*, the amount by which the ability of others is different from one's own ability. Hence, we refer to this model as the *differentiated by ability* model[1]. As shown in Figure 2.2, this model considers inference at all three levels: population, specific individuals, and knowledge signals.

[1]note that assessing differential ability δ and a^s is equivalent to separately assessing a^s and a^o

As more information becomes available via external knowledge signals such as performance feedback, it is possible to learn whether the other person is better ($\delta > 0$) or worse ($\delta < 0$) relative to themselves.

Additionally, this model assumes that estimates of perceived difficulty of the problem (d) are the same for both self and the other person. Hence, the participant uses their perceived item difficulty when estimating the other person's score on the same task. This is a key feature of the model. In contrast to the next model (M_2), it allows a person to draw meaningful insights from their experience with the task. When predicting the other's score for a target problem, the prediction can be informed by information gained about differential ability from previous problems and the participant's own perceived problem difficulty for the target problem. Therefore, this model predicts correlated scores between self- and other-estimated scores.

An equivalent formulation of the differentiated by ability model is a 'differentiated by difficulty' model. Intuitively, the differentiated by difficulty model suggests that people assume equal ability for self and other but would experience the same task as having different difficulties. Due to the interconnected relationship between the ability and difficulty parameters, the two models would make similar predictions.

Fully Differentiated Model (M_2)

This model assumes that other-assessment is not informed by any self-assessed estimates, consistent with a *fully differentiated* model of the other. As shown in Figure 2.2, this model assumes that inference about self and others is disjointed. As a consequence, there is no information sharing at the individual level. The fully differentiated model suggests that people draw no information from their own experience with the task when reasoning about another person. According to this model, in the absence of feedback, the participant possesses

no meaningful information that can be used to inform predictions of the other person's performance. The participant starts with arbitrary priors about the other person's ability and perceived item difficulty and proceeds to learn about the other by solely observing their scores (in the feedback condition) and ignoring any insights from their own experience. As more observations become available over time, the estimated other ability can be updated and can inform the prediction for the next set of problems. Note that, because people do not rely on their experience with the task to assess the other person, this model does not allow the person to learn any meaningful estimates of difficulty as experienced by the other person. Both ability and difficulty estimates of the other are evaluated independent of the ability and difficulty estimates of the self.

Undifferentiated Model (M_3)

The last model assumes that the predicted other scores are highly constrained as the process of other-assessment uses the exact same information as the process used for self-assessment. As shown in Figure 2.2, this formulation ignores inference at the specific individual or the population levels of the proposed hierarchical framework. Therefore, this model suggests that people rely only on their assessment of themselves to make predictions about the other person. Overall, this model predicts no differentiation in ability as more information about the other person becomes available.

2.2.2 Overview of Experiments and Modeling

Up to this point, we have explained the hierarchical framework and model variants primarily at a conceptual level. In the next sections, we will apply the framework to specific empirical paradigms. First, we will describe an empirical paradigm based on an image classification task where participants sequentially make predictions about the performance of themselves as

well as the performance of another person. We evaluate how the self- and other predictions differentiate over time as more information about the other person becomes available and test which of the three instantiations of the hierarchical model best accounts for the observed data. Second, we will use the hierarchical model to account for previous empirical findings about other assessment in tasks where no specific information about the other person is available and participants reason about the other person and relative placement in the population using a combination of internal and external knowledge signals.

2.3 A Sequential Knowledge Assessment Task

We develop an empirical paradigm similar to observer paradigms (Jameson et al., 1993) where there are multiple rounds of assessing one's own performance as well as the performance of another target person, allowing people to update their mental models of the target person. In this empirical paradigm, participants go through a series of problem-sets, where each problem-set consists of a series of classification problems involving images of different species of animals (see Figure 2.3 for examples). After each problem-set, participants self-assess their own performance ("how many items do you think you answered correctly?") as well as the performance of a target person who previously performed the task ("how many items do you think Akira answered correctly?"). The target person is referenced with a made-up name but the associated data is based on an actual person who performed the experiment. In the no-feedback condition of the experiment, no information is provided about the actual performance of the target person and assessment is based on a priori predictions. In the feedback condition, the performance of the target person can be used by the participant to update their mental model of the other person's ability. In the example in Figure 2.3, when the participant is predicting how many items Akira answered correctly in the first problem-set (involving birds), no feedback has been presented yet. However, after learning

that Akira answered 9 out of 12 items correctly while the participants themselves answered only 7 items correctly, this provides an opportunity for the participants to adjust their mental model of the other person. This differentiated mental model can then be applied in the assessment phase for the second classification problem-set (dogs) and further refined after receiving feedback. We apply an instantiation of the proposed framework to behavioral data collected via the sequential knowledge assessment task, extending the work by (Jansen et al., 2021) on other-assessment. We assume that other-assessment proceeds in a similar fashion as self-assessment by combining a subjective estimate for the perceived ability of the other person with estimates of the perceived difficulty for the other person. We use this framework to assess the degree of differentiation between the mental model of self (containing ability and problem difficulty estimates for self) and the mental model of others (containing ability and problem difficulty estimates for the other person). Consistent with previous research that has shown that one's own perceived difficulty in retrieving information or solving problems can be used to predict the difficulty experienced by others (Jameson et al., 1993; Kelley & Jacoby, 1996; Nickerson, 1999; Nickerson et al., 1987), we show that the subjective estimates of problem difficulty are shared between the self- and other mental models. In addition, we show that the other-person model differentiates from the self model based on differences in perceived ability. As information becomes available about the other person's performance, the differential ability can be updated, leading a person to upgrade or downgrade the predictions relative to their own ability.

2.3.1 Notation

Before describing the computational model, we introduce some notation and define the scope of the model. In our empirical paradigm, each person i is paired with a single other person. That is, each person reasons about their own performance and one other person's performance throughout the experiment. Therefore, we will omit from the notation which specific other

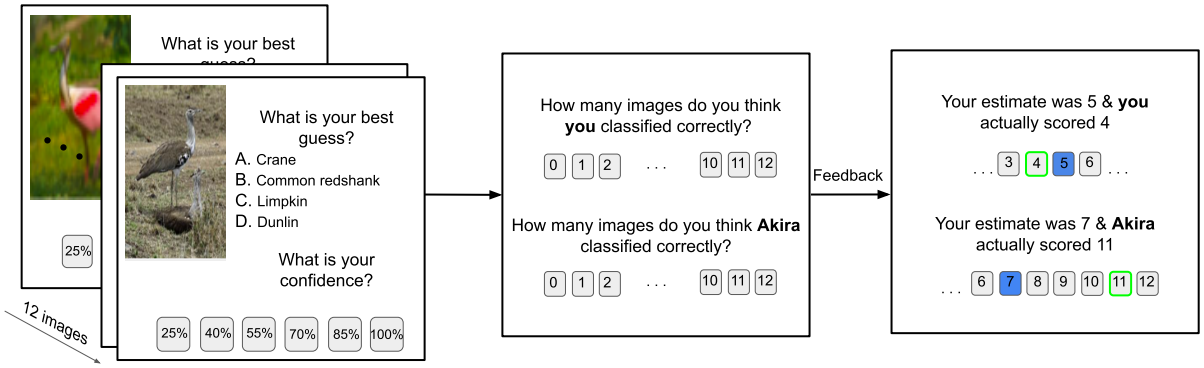


Figure 2.3: Illustration of the empirical paradigm for self- and other assessment. Participants go through a series of classification problem-sets requiring participants to discriminate between different types of animals in a four-alternative forced-choice task. After classifying twelve images that constitute a problem-set, participants proceed to the assessment phase, where they estimate the number of items they and another person answered correctly. The assessment phase is followed by feedback (if provided) on the actual number of items answered correctly. Numbers in blue and green show estimates and true scores respectively. The scores of the other (target) person are based on selected participants who previously went through the experiment. A number of different names, including Akira, are used to reference the other person.

person i the self is reasoning about. We instead use the superscripts s (self) and o (other) to denote both the true scores of a person or of the assigned other person, and subjective estimates of a person about their own or the other person's performance respectively. We will use subscript j to index the problem-set, where $j \in \{1, \ldots, L\}$.

For example, $x^s_{i,j}$ represents the number of items person i answered correctly in problem-set j, and $x^o_{i,j}$ represents the number of items answered correctly in problem-set j by the other person paired with i. $\hat{x}^s_{i,j}$ represents the number of items person i *estimates* they answered correctly on problem-set j. Similarly, $\hat{x}^o_{i,j}$ represents the estimated performance of the other person from the viewpoint of person i, i.e., how many items person i *believes* the other person answered correctly for problem-set j. Both true and estimated scores are limited to the number of classification items (M) within each set, $x_{i,j} \in \{0, \ldots, M\}, \hat{x}_{i,j} \in \{0, \ldots, M\}$, where $M = 12$ throughout our experiments. In the empirical paradigm, the order in which the

problem-sets are presented varies across participants. We will use subscript $t = 1, 2, \ldots, T$ to refer to the order in which problem-sets are presented, and j to refer to the specific type of problem-set. For example, the bird problem-set in Figure 2.3 could correspond to $t = 1$ and (say) $j = 4$. For person i in this particular example and for $t = 1$, the number of estimated and true self- and other answered correctly are $\hat{x}^s_{i,t} = 5, \hat{x}^o_{i,t} = 7, x^s_{i,t} = 4, x^o_{i,t} = 11$, with $M = 12$.

2.3.2 Modeling actual performance

To formalize actual performance, we start with a model from Item Response Theory (IRT, (Fox, 2010; van der Linden & Hambleton, 2013)) which accounts for the observed performance differences across people and problem-sets. The IRT model will also form the basis for the two other parts of the model (self- and other-assessment). To simplify the application of the IRT model across the three parts, we will use a basic Rasch model (Rasch, 1993) extended for ordered polytomous categories (i.e., the responses $x \in \{0, \ldots, M\}$). The key assumption of the Rasch modeling approach is that the number of items answered correctly, $x_{i,j}$ for person i and problem j, is modeled by combining two latent factors, the ability a_i of each person i and the difficulty d_j for problem-set j:

$$\begin{aligned} \theta_{i,j} &= a_i - d_j \\ p_{i,j} &= \frac{1}{1 + \exp(-\theta_{i,j})} \\ x_{i,j} &\sim \text{OrderedProbit}(p_{i,j}, v, \sigma) \end{aligned} \tag{2.1}$$

Note that a_i and d_j represent the objective ability of person i and the objective difficulty of problem j measured using the IRT model. $\theta_{i,j}$ represents the latent score of person i on problem-set j on a logit scale ($-\infty < \theta < \infty$) which is modeled as a sum of a_i, the ability of person i, and d_j, the difficulty for problem-set j. Therefore, a higher score is expected for people with high ability or problems with low difficulty. The variable $p_{i,j}$ represents the latent

score for person i and problem-set j converted to a value between 0 and 1. The ordered probit model[2] is a simple probabilistic process that maps the latent score $p_{i,j}$ to a discrete score, $x_{i,j} \in \{0, \ldots, M\}$. In this process, normally distributed noise with zero mean and standard deviation σ is added to the latent score $p_{i,j}$ and the placement of the resulting value in a set of intervals (defined by the cutoff points v) determines the observed score. The variable σ represents the uncertainty in mapping from latent to observed scores (see Appendix A.1 for details).

In this particular model, we have assumed that ability is one-dimensional – all variations in ability can be characterized by changes along a single overall ability scale. In Section 2.7, we consider multidimensional extensions of this model, analogous to multidimensional item response theory (Reckase, 2009) that allow for differences in ability along a number of dimensions.

2.3.3 Modeling self-assessment

For the self-assessment model, we assume that each person i's estimate of their own ability a_i^s and estimate of the problem difficulty for problem-set j, $d_{i,j}^s$, are noisy and distorted versions of the true values. Both a_i^s and $d_{i,j}^s$ may be interpreted as subjective estimates made by each person i on problem j. These subjective estimates are related to the objective measures of ability (a_i) and difficulty (d_j) from Eq. 4.8 according to:

$$\begin{aligned} a_i^s &\sim \mathrm{N}\left(a_i, \sigma_{a,i}\right) \\ d_{i,j}^s &\sim \mathrm{N}\left(\gamma d_j + \lambda, \sigma_{d,i}\right) \end{aligned} \tag{2.2}$$

where γ and λ parameter are scaling parameters that can capture systematic deviations of

[2]There are alternative generative models for ordered responses including the graded response model (Greene & Hensher, 2010). We have found that the use of this alternative construction does not change the qualitative results.

people's estimates from the true values of difficulty (d_j). Specifically, when $\lambda > 0$, problem difficulty will be overestimated leading to underestimates of scores. Similarly, when $\lambda < 0$, problem difficulty will be underestimated leading to overestimates of scores. The linear transformation of the problem difficulty is similar to the linear-in-log-odds models that have been used to model distortions in probability estimation in a variety of cognitive tasks (Turner et al., 2014; H. Zhang & Maloney, 2012).

An estimated score $\hat{x}^s_{i,j}$ by person i for problem-set j is produced by combining the self-estimated ability and problem difficulty by following the same general process as in Eq. 4.8:

$$\begin{aligned} \theta^s_{i,j} &= a^s_i - d^s_{i,j} \\ p^s_{i,j} &= \frac{1}{1 + \exp(-\theta^s_{i,j})} \\ \hat{x}^s_{i,j} &\sim \text{OrderedProbit}(p^s_{i,j}, v, \sigma^s) \end{aligned} \tag{2.3}$$

Overall, there are two sources of noise that can produce distortions in self-estimation. The subjective ability might not reflect the true ability and the subjective problem difficulty might systematically deviate from the actual problem difficulty.

Note that the self-assessment model in Eqs. 2.2-2.3 is similar to the IRT model in Eq. 4.8 but that it plays a very different role in our approach conceptually. The IRT model in Eq. 4.8 serves the purpose of a data analysis model to estimate the true abilities and true item difficulties whereas the self-assessment model in Eqs. 2.2-2.3 formulate a cognitive model to explain the process of self-assessment. We use the ordered probit model as a link function to map a person's subjective latent probability of being correct, $p^s_{i,j}$, to a score between 0 and 12. However, as we will show in a later section of this chapter, we may easily modify this to accommodate cases where different knowledge signals are available (e.g., feeling-of-knowing or response time).

2.3.4 Modeling other-assessment

For this model we make the assumption that the way people reason about the other person's performance is through the lens of their own self-assessment process. That is, once a person i has an estimate of the ability of the other person (a_i^o) and an estimate of the problem difficulty for problem-set j as experienced by the other person ($d_{i,j}^o$), we assume that scores for the other person can be predicted by applying the same cognitive model as Eq. 2.3:

$$\begin{aligned} \theta_{i,j}^o &= a_i^o - d_{i,j}^o \\ p_{i,j}^o &= \frac{1}{1+\exp(-\theta_{i,j}^o)} \\ \hat{x}_{i,j}^o &\sim \text{OrderedProbit}(p_{i,j}^o, v, \sigma^s) \end{aligned} \tag{2.4}$$

Note that in this model, a_i^o and $d_{i,j}^o$ are not the true ability and problem difficulty of the other. Instead, they represent i's estimate of the true ability of other and the estimate of the difficulty for the other.

2.3.5 Hypotheses about the Relationship between the Self- and Other Model

Now that the basic models for self- and other assessment have been formalized, we specify how the three hypotheses, the differentiated by ability (M_1), fully differentiated (M_2), and undifferentiated model (M_3) translate to different computational assumptions about how the estimates of the other ability and problem difficulty are formed. The underlying computational assumptions of the three hypotheses are summarized in Table 2.1 in terms of the notation above. Note that these relationships describe different *beliefs* held by the person making inferences about the other person. In other words, these are psychological assumptions about how people use available information to draw inferences in their cognitive model of the other

person.

Model	Hypothesized Dependencies	
	a_i^o and a_i^s	$d_{i,j}^o$ and $d_{i,j}^s$
M_1: Differentiated by Ability	$a_i^o = a_i^s + \delta_i$	$d_{i,j}^o = d_{i,j}^s$
M_2: Fully differentiated	unrelated	unrelated
M_3: Undifferentiated	$a_i^o = a_i^s$	$d_{i,j}^o = d_{i,j}^s$

Table 2.1: Model-based hypotheses about the relationship between self- and other-mental model parameters. Each hypothesis is associated with a different cognitive model for other-assessment.

M_1: Differentiated by ability model

The differentiated by ability model (M_1) assumes that for each type of problem-set j, the difficulty for another person is the same as the difficulty for one's self (i.e. $d_{i,j}^o = d_{i,j}^s$). However, it allows for the possibility that there is a difference, δ_i in ability between self and other from the viewpoint of person i. This differential ability is inferred as information about the performance of the other person becomes available over time.

The inference process can be stated as a sequential updating problem. After t problem-sets, person i has received information about the other person's performance $x_{i,1}^o, \ldots, x_{i,t}^o$. (e.g., if after $t = 3$ rounds of problem-sets, the other person scored 11, 7, and 8 correct out of 12, we have $x_{i,1}^o = 11$, $x_{i,2}^o = 7$, and $x_{i,3}^o = 8$). On the basis of this information, a prediction for the performance on the next problem-set, $\hat{x}_{i,t+1}^o$, can be made by first making an inference about the differential ability δ_i from the viewpoint of person i:

$$\begin{aligned} p(\delta_i | x_{i,1}^o, \ldots, x_{i,t}^o) &\propto p(x_{i,1}^o, \ldots, x_{i,t}^o | \delta_i, d_{i,1}^s, ..., d_{i,t}^s) p(\delta_i) \\ &= \left(\prod_{\tau=1}^{t} p(x_{i,\tau}^o | \delta_i, d_{i,\tau}^s) \right) p(\delta_i) \end{aligned} \tag{2.5}$$

Note that the second line follows from the first because of conditional independence. The term in the product can be evaluated by Eq. 2.4 by using the model assumption $a_i^o = a_i^s + \delta_i$.

In the next step, on the basis of the posterior estimates of a_i^o the score of the other person for the next problem-set presented at time $t+1$, $p(x_{i,t+1}^o|a_i^o, d_{i,t+1}^o)$, can be predicted by applying Eq. 2.4. Here, $d_{i,t+1}^o$ is the same difficulty as inferred by the self using the self-assessment model ($d_{i,t+1}^s$). The term $p(\delta_i)$ reflect person i's prior about the differential ability. We assume that this prior is centered around zero, such that at the start of learning, the mental model of self and other are undifferentiated.

M_2: Fully differentiated model

The most unconstrained of the three hypotheses is the fully differentiated model (M_2). In this model, the estimates in the mental self model are unrelated to the estimates in mental other model (i.e. a_i^o is unrelated to a_i^s and $d_{i,j}^o$ is unrelated to $d_{i,j}^s$). This model posits that people use no insights from their experience with the task when assessing the other person.

A prediction for the performance on the next problem-set $t+1$, $\hat{x}_{t+1}^o$, can be made by making an inference about the ability of the other person (a_i^o) and difficulty for the other person $(d_{i,1}^o, \ldots, d_{i,t}^o)$:

$$\begin{aligned} p(a_i^o, d_{i,1}^o, \ldots, d_{i,t}^o|x_{i,1}^o, \ldots, x_{i,t}^o) &\propto p(x_{i,1}^o, \ldots, x_{i,t}^o|a_i^o, d_{i,1}^o, \ldots, d_{i,t}^o)p(d_{i,1}^o, \ldots, d_{i,t}^o)p(a_i^o) \\ &= \left(\prod_{\tau=1}^{t} p(x_{i,\tau}^o|a_i^o, d_{i,\tau}^o)p(d_{i,\tau}^o)\right)p(a_i^o) \end{aligned} \tag{2.6}$$

The terms $p(a_i^o)$ and $p(d_i^o)$ reflect a person's priors about the other person and we have assumed independence between these priors. Note that the second line follows from the first because of conditional independence. The score of the other person for the next problem-set, $p(x_{i,t+1}^o|a_i^o, d_{i,t+1}^o)$, can be predicted by applying Eq. 2.4 to the posterior estimates of a_i^o and drawing a sample from the posterior of d_i^o.

Note that the flexibility of this other-assessment model allows for the possibility that a

problem-set has differing levels of difficulty across people. When the same type of problem-set occurs over time, this model will allow a person to potentially make accurate predictions for the other person's performance. However, in an environment where problem-sets do not repeat (as in our empirical paradigm), this model does not generalize well as the information acquired for each type of problem-set is not utilized in the future.

M_3: Undifferentiated model

The most constrained of the three models is the undifferentiated model (M_3). In this model, the mental models of self and other are the same and remain undifferentiated as new information becomes available about the performance of the other individual. Therefore, the process for producing predictions for the problem-set presented at time t for self ($\hat{x}^s_{i,t}$) and other ($\hat{x}^o_{i,t}$) in Eqs. 2.3-2.4 are based on the same parameters. Note that in this model, the predicted self- and other scores can still deviate from each other because of the noise process of producing discrete scores in Eqs. 2.3-2.4.

2.4 Experiments

We conduct two image classification experiments to investigate self- and other-assessment and develop and test the computational models. In Experiment 1, we collect behavioral data from 68 participants on the basic experimental paradigm that only includes self-assessment. Experiment 2 follows the same experimental paradigm but also includes other-assessment of participants from Experiment 1. There were 128 individuals in total serving as "self" in Experiment 2. Specifically, the best and worst performing 16 participants from Experiment 1 served as the "other" individuals that participants in Experiment 2 are learning about.

2.4.1 Methods

Participants

Participants were recruited through Amazon Mechanical Turk. 68 and 128 participants were recruited for Experiment 1 and Experiment 2 respectively. To be eligible for the studies, participants were required to meet the following criteria: 1) have greater than or equal to 80% Human Intelligence Task (HIT) approval rate for all requesters' HITs; 2) be located in the United States and; 3) be 18-years-old or older. All participants provided informed consent before taking part in our study and were compensated $6 for their participation. The median time to complete the experiment was 33 minutes.

Images

There were 192 unique images in total used in the experiments, divided equally into 4 categories (birds, dogs, primates, and reptiles). Each category was associated with T = 4 x 4 = 16 problem sets in total, with each problem-set containing M = 12 individual classification problems. In each classification instance, the goal is to classify images according to four different labels corresponding to a specific category. For example, for one of the bird problem-sets the labels are *crane*, *common redshank*, *limpkin*, *dunlin*, and for one of the dog problem-sets the labels are *Afghan hound*, *Ibiza hound*, *Norwegian elkhound*, *redbone coonhound* (See Appendix A.3 for a list of the 16 classification problem-sets). The images and labels for the classification problems are based on the ImageNet Large Scale Visual Recognition Challenge (ILSRVR) 2012 database (Russakovsky, Deng, Su, Krause, Satheesh, Ma, Huang, Karpathy, Khosla, Bernstein, et al., 2015b). ImageNet is an image dataset where the labels for each image are hierarchically organized according to the WordNet hierarchy (Miller, 1995). We selected 16 classification problem-sets equally divided among the

4 categories. For each classification problem-set, we randomly selected 12 images (3 images per label) from the validation set of ImageNet. Each image was center-cropped and scaled to 256 x 256 pixels.

Procedure

In both Experiments 1 and 2, participants went through 16 problem-sets where each problem-set included 12 classification problems of a particular category as well as a prediction task where participants assessed their own performance (Experiment 1 and 2) and also assessed another person's performance (Experiment 2 only). For each problem-set, a participant first classified 12 individual images (Figure 2.3). For each image, the participant selected a label from four response alternatives (e.g. *little blue heron*, *oystercatcher*, *dowitcher*, and *great egret*). The response alternatives remained the same during each problem-set. The participant also selected a discrete confidence level from six alternatives (25%, 40%, 55%, 70%, 85%, and 100% confidence). The 25% and 100% confidence levels had additional text labels "Guessing" and "Absolutely Certain" respectively. No feedback was provided during this classification phase. The confidence ratings and individual classifications were not used for the purpose of this research.

At the end of each problem-set, the 12 images from the preceding classification task were presented simultaneously on the screen. In both Experiments 1 and 2, participants were instructed to predict the number of images they classified correctly by selecting a response option between 0 and 12 (self-assessment). In Experiment 2, they were also asked to predict the performance of another person by selecting a number between 0 and 12 (other-assessment). This person was referred to by a name, sampled randomly from a set of 7 male and 7 female names (e.g. "*Vince*", "*Glenda*"). The participant was told that this was not the real name of the other person but that the other person was an actual person who participated previously in the experiment (the same name was used throughout the experiment).

In Experiment 1, after the predictions were made for each problem-set t, participants were provided feedback and were told the actual number of correct responses (e.g., "You classified 8 out of 12 images correctly"). Participants were given an option to see which individual images they classified incorrectly. The correct label was not shown. After this feedback, participants proceeded to the next problem-set $t+1$. In Experiment 2, in the feedback condition, feedback was provided about the number of correct self as well as other-responses (e.g. "Vince scored 6 out of 12 images correctly"). In the no-feedback condition, this feedback about self- or other-performance was omitted.

Overall, each participant provided 192 image classifications with corresponding confidence levels and provided 16 predictions about their performance across 16 different types of classification problem-sets.

Design

The 16 best and 16 worst performing participants from Experiment 1 served as the other person to learn about in Experiment 2. We will refer to these two groups of other people as top and bottom respectively. In the feedback condition, a participant in Experiment 2 received feedback about the particular other person assigned to the participant. In the no-feedback condition, no such information was provided. The assignment of the 16 top and 16 bottom participants from Experiment 1 to the 128 participants in Experiment 2 was counterbalanced across the two feedback conditions – each target participant from Experiment 1 was assigned to exactly four participants in Experiment 2, two in the feedback and two in the no-feedback conditions. This study was not preregistered. All data sets analyzed in this work can be accessed from: https://osf.io/68347/.

Metrics for Assessment Performance

For both self- and other-assessment, we report results based on three different metrics to provide a more comprehensive picture of assessment performance (Dunning & Helzer, 2014). Note that because our assessment task of estimating the number of items scored correctly does not relate to a binary detection task, various standard metacognition measures such as metacognitive sensitivity and efficiency (Fleming & Lau, 2014) cannot be applied.

The first metric is the coefficient of predictive ability (CPA) (Gneiting & Walz, 2021), a rank-based measure that generalizes the Area under the Curve (AUC) to ordinal and continuous variables (for details, see Appendix A.4). In our context, the CPA evaluates how well people can discriminate in their assessment between different true scores. More specifically, the CPA is a weighted probability that under random sampling of problem-sets, a problem-set with a higher true score is self-assessed with a higher score than a problem-set with a lower true score[3]. The weights in CPA are based on the distance between the ranks of the true scores. Therefore, a person who is able to assign different scores to closely ranked true scores will achieve a higher CPA. The CPA measure is theoretically appropriate for a number of reasons: the CPA is equivalent to AUC when applied to binary outcomes, and equivalent to Kendall's tau rank-order correlation when there are no ties in the true scores. It is also closely related to the Goodman Kruskal's Gamma coefficient that has been used to assess metacognitive sensitivity (Nelson, 1984). Because of the rank-based nature, CPA is insensitive to bias. Any changes to the estimated scores that preserve ranking will result in the same CPA. The CPA attains values between 0 and 1. A value of 1 is attained when there is a perfect correspondence between estimated and true scores. A value of $1/2$ is attained when the estimated scores are independent of the true scores.

Second, we report a bias measure to measure the systematic deviations between the true

[3]ties between the self-assessed scores are resolved at random.

and estimated score, defined as Bias = $(1/N)\sum_{i=1}^{N}(\hat{x}_i - \bar{x})$ where $\hat{x}$ is the estimated score through self- or other assessment and $\bar{x}$ is the mean of true scores across problem-sets. If the assessment scores are consistently overestimating or underestimating the true performance, the bias score will be positive and negative respectively.

Third, to facilitate comparison to previous reported results on assessment (e.g. (Zell & Krizan, 2014)), we also report the Pearson correlation coefficient (ρ) between the true and estimated scores.

2.4.2 Model Inference

We used Markov Chain Monte Carlo (MCMC) sampling to infer model parameters for the cognitive models presented in Figure A.2 and obtain samples from the posterior distribution. We chose the Stan computing environment for posterior inference (Stan Development Team, 2020). Model inference proceeds in a sequential fashion. We begin with actual performance assessment, followed by self-assessment and finally other-assessment. We start by estimating the parameters (a, d, σ) that account for actual performance of the participants using the true scores x^s. These parameters were estimated using a standard 1-parameter IRT model described in section 4.8 on modeling actual performance. In the next stage of our inference, we treat the posterior means of a, d, σ as observed data to infer the parameters of our self-assessment model $(a^s, d^s, \sigma^{a,i}, \sigma^{d,i}, \sigma^s, \lambda, \gamma)$ using participant's estimates of their true scores $(\hat{x}^s)$. Inference on the self-assessment model gives us the estimated perceived ability of self (a^s) and perceived difficulty of items (d^s) for every individual. We ignore learning over time when estimating these self-assessment parameters as we did not observe any such learning in our empirical data. Finally, the posterior means of the parameters from the self-assessment model serve as the starting point for the other-assessment models.

We use the three variants of the other-assessment model to simulate participants' estimates

of the other person's scores. To do inference, we condition on a^s, d^s, σ^s, and x^o. Figure A.2 shows the graphical models corresponding to each model variant. At the first time step, depending on the variant of the other-assessment model, we either use priors for a^o and d^o (M_3) or values of a^s and d^s (M_1, M_3) to predict the participant's first estimate of the other person's performance (here, the participant has not received any information about the other person). At each subsequent time step, participants may learn about the other person in the feedback condition. Simulating from the undifferentiated model (M_3) requires no learning: we simply use self estimates (a^s and d^s) to predict the participant's estimated scores of the other person on each time step. To simulate the participant's estimates using the fully differentiated model (M_2), we use the mean posterior estimates of a^o and d^o from the previous time step to predict estimated scores of the other person. For the differentiated by ability model (M_1), we use the the mean posterior estimates of a^o from the previous time step and d^s for the current item to predict the participant's estimated score of the other person $\hat{x}^o$.

Our experimental and modeling setup allows us to simulate a participant's estimate of any other person's score, i.e, we can use a participant's inferred self-ability and item difficulties from the self-assessment model to predict their estimates of any randomly picked other person's scores. For Figures 2.7 and 2.8, we increased the number of simulated other-assessments fourfold in order to more clearly visualize the differences in model predictions from the three different linkage hypotheses. In these simulations, for every participant, we simulate their other assessment separately for four randomly assigned participants as their 'other persons'. We then use the other-assessment procedure described above to make predictions about the participant's estimates of the new others' scores.

Implementing the IRT model requires careful attention to the selection of priors on both ability and difficulty to avoid potential identifiability issues. For the actual performance model, we used normal priors of ability and difficulty IRT parameters: $a_i \sim \mathcal{N}(0, 1)$, $d_j \sim \mathcal{N}(\mu_d, \sigma_d)$,

where $\mu_d \sim \mathcal{N}(0,1)$, $\sigma_d \sim \text{Cauchy}(0,5)$. Additionally, for the self-assessment model we used Normal priors for $\lambda \sim \mathcal{N}(0,1)$, $\gamma \sim \mathcal{N}(0,1)$ and Cauchy priors for standard deviation parameters $\sigma_{a,i}$, $\sigma_{d,i} \sim \text{Cauchy}(0,5)$. Finally, for the differentiated-by-ability model, we use a normal prior on $\delta_i \sim \mathcal{N}(\mu_\delta, \sigma_\delta)$ where $\mu_\delta \sim \mathcal{N}(0,1)$ and $\sigma_\delta \sim \text{Cauchy}(0,5)$. Throughout the inference process, we ran the sampler with 2 chains with a burnin of 1000 iterations before taking 1000 samples per chain. The chains mixed appropriately based on Rhat values (close to 1). Stan code for self- and other-assessment models can be accessed from: https://osf.io/68347/.

2.4.3 Empirical Results

Classification performance

Participants substantially differed in overall performance. From the worst to the best performing participant, the mean proportion correct varied between 33% to 81% across Experiments 1 and 2. Classification performance improved slightly within each problem-set. Across the first, middle, and last 4 classification items in a problem-set, average performance was 53%, 55%, and 57% respectively. This improvement is likely due to participant strategies of adjusting their classifications after seeing a larger range of images. Across problem-sets, no apparent learning took place (keep in mind that each problem-set involved new classification problems with a unique set of labels). The average accuracy, grouped by 4 consecutive problem-sets was 56%, 56%, 53% and 55%.

2.4.4 Assessment performance

While many metrics have been introduced to evaluate metacognition, they are typically applied to binary decision tasks (Fleming & Lau, 2014). Given that the self- and other

	Across participants			Per participant			
Type / Condition	CPA	Bias	ρ	Mean CPA	Mean Bias	Mean ρ	N
Self-assessment							
Exp. 1, Feedback (All)	0.75	-1.41	0.52	0.79 (0.011)	-1.41 (0.19)	0.62 (0.019)	68
Exp. 1, Feedback (TB)	0.75	-1.24	0.53	0.80 (0.015)	-1.24 (0.30)	0.62 (0.029)	32
Exp. 2, Feedback	0.82	-1.41	0.65	0.82 (0.011)	-1.41 (0.14)	0.64 (0.022)	64
Exp. 2, No Feedback	0.78	-1.54	0.57	0.80 (0.009)	-1.54 (0.21)	0.64 (0.018)	64
Other-assessment							
Exp. 2, Feedback	0.70	-0.08	0.40	0.63 (0.013)	-0.08 (0.14)	0.28 (0.027)	64
Exp. 2, No Feedback	0.63	-0.60	0.27	0.69 (0.016)	-0.60 (0.26)	0.41 (0.032)	64

Table 2.2: Self- and other-assessment performance across experiments and conditions. For the analysis per participant, the statistics are calculated at the individual participant level and then averaged; numbers between parentheses are standard errors. N is the number of participants. For the analysis across participants, we ignore individual differences and report a single outcome across participants and problem-sets. TB refers to the subset of participants who were part of the top and bottom performers

estimated and true scores are based on discrete counts with more than two outcomes, we adopt a relatively new measure, the coefficient of predictive ability (CPA, (Gneiting & Walz, 2021)) to assess metacognitive sensitivity, the ability to discriminate between different true scores.

Table 2.2 shows the self- and other assessment performance based on CPA as well as Bias (See Methods for details), and Pearson correlation coefficient (ρ) between true and estimated scores[4]. According to the CPA as well as the Pearson correlation, participants' self- and other assessment is well above chance level (note that chance level for CPA is 0.5). For self-assessment, the Pearson correlation coefficients are in the 0.5-0.7 range which is well above the 0.2-0.3 range reported for many other self-assessment tasks (Zell & Krizan, 2014).

Figure 2.4 shows the self-estimated score as a function of the true score for a particular problem-set. The data for this analysis is combined across Experiments 1 and 2 (see Supplementary for the results separated by Experiment). The results show a small range

[4]Although not used in this research, we also collected confidence scores for individual questions in each problem set. There is a close correspondence between the mean of the estimated probabilities across items and the estimated score (.78 (sd = .14) for Experiment 1, and .82 (sd =.13) for Experiment 2), suggesting that participants' estimates are based on aggregates of individual confidence scores.

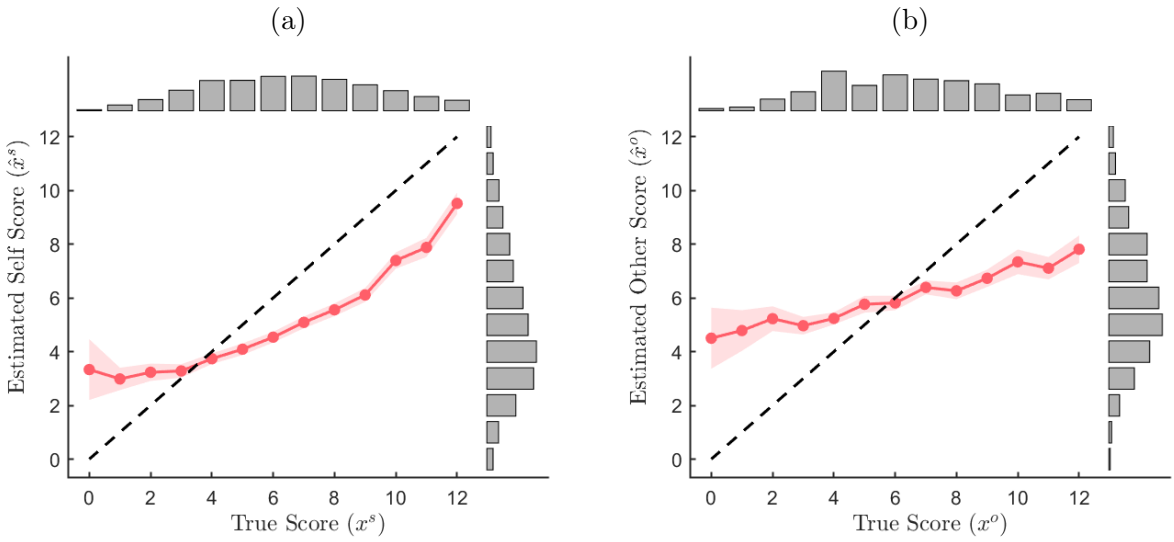


Figure 2.4: Mean estimated self score (a) and other score (b), each as a function of actual performance for a particular problem-set. For the self-scores, the data is combined across Experiments 1 and 2. Histograms show the marginal distribution of scores. The colored areas shows 95% confidence intervals.

of true scores associated with a pattern of overestimation. For a larger range of true scores, there was a pattern of underestimation. Generally, this pattern of systematic deviations is consistent with previous findings in self-assessment (Jansen et al., 2021; Kruger & Dunning, 1999) and is consistent with the general pattern of over- and underestimation in subjective assessment tasks (H. Zhang & Maloney, 2012). However, it is important to note that there were few problem-sets where participants produced the low true scores that are associated with the overestimation pattern (see the marginal distribution at the top of the figure). Overall, there was a tendency to underestimate performance, as revealed by the negative bias values in Table 2.2. Across Experiments 1 and 2, there were 169 participants with more under- than overestimates in the self-assessment and only 19 participants with more over- then underestimates.

Other-assessment is a more challenging task than self-assessment leading to somewhat lower performance. However, participants' accuracy in assessing other participants (i.e., the

participants in Experiment 1) is not far off from the ability of those participants to predict their own performance (i.e., see self-assessment results from Experiment 1, top/bottom performers). Across participants, feedback improves other-assessment on all performance metrics including bias[5].

Figure 2.5 demonstrates that individual participants are tracking the performance of other people in the feedback condition. In the feedback condition, when participants make predictions about the other person for the first problem set, no feedback has been provided yet and the results show that predictions are the same across top- and bottom other performers. However, the estimated mean scores diverge within a few problem-sets depending on the type of other person they are learning about. In the no feedback condition, participants' estimated scores cannot (by definition) reflect differences between other people. Instead, without feedback, estimates have to be based on prior knowledge only. Generally, these prior predictions underestimate true performance (i.e., negative bias).

Finally, the other assessment shows patterns of over- and under-estimation that are similar to self-assessment. Figure 2.4(b) shows that for particular problem-sets that lead to low (high) true scores, participants tend to over (under) estimate performance. This pattern is similar across feedback conditions.

2.4.5 Relationship between self- and other-assessment

Figure 2.6 shows that there is a close correspondence between self- and other-assessment. In the no feedback condition, there is a strong tendency to link the estimate of the other score to the estimate of the self score, suggesting that when people believe a problem is

[5]At the individual participant level, discrimination (CPA) and correlation (C) is higher in the absence of feedback which suggests that feedback lowers the ability to discriminate between different levels of performance. However, it should be noted that each participant in the feedback condition tracks the performance of either a top or bottom performing other person. Therefore, for those participants, there is a restricted range of scores to discriminate which reduces CPA and C.

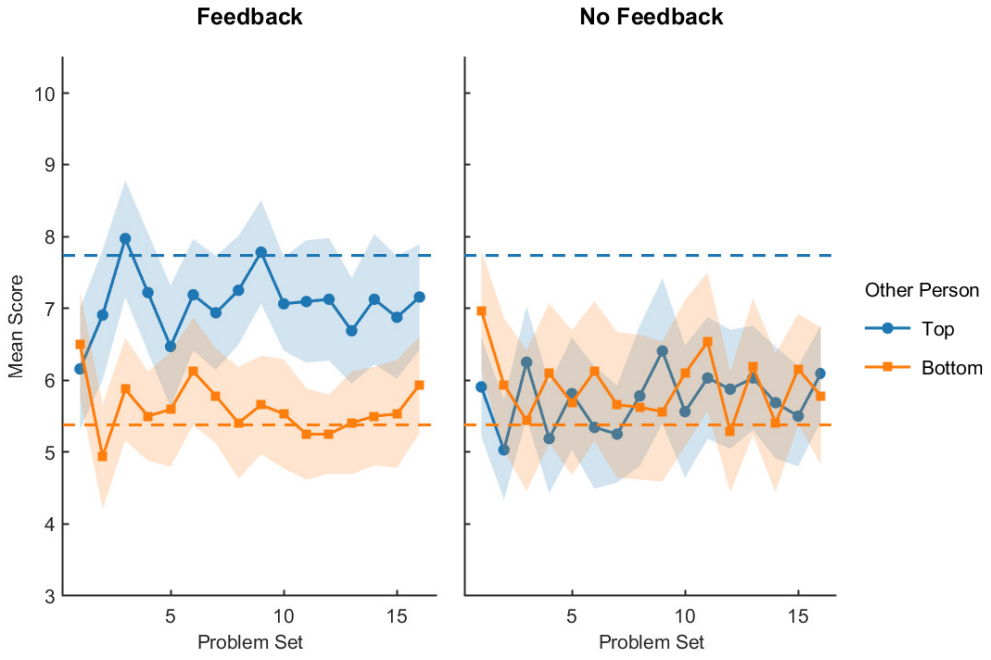


Figure 2.5: Mean estimated score of the other person across feedback conditions and performance levels of the other person. Dashed lines show the mean true score across the top and bottom performing other people. Note that the no feedback condition (right panel) shows the a priori predictions of participants. The colored areas show 95% confidence intervals.

challenging for themselves, they believe it is likely to be challenging for other people as well. In the feedback condition, the results show the same pattern but the predictions are differentiated by the type of other person they are learning about with higher predicted scores for a top-performer. Therefore, in the feedback condition, the results suggest that two factors affect the other-assessment, the estimated overall performance of the other person and the perceived problem difficulty.

2.4.6 Discussion of Empirical Results

Our empirical results are consistent with the hypothesis that participants are developing and updating a mental model that allows them to make inferences about the overall level of performance of the other person. Figure 2.5 shows that participants' estimates of top and

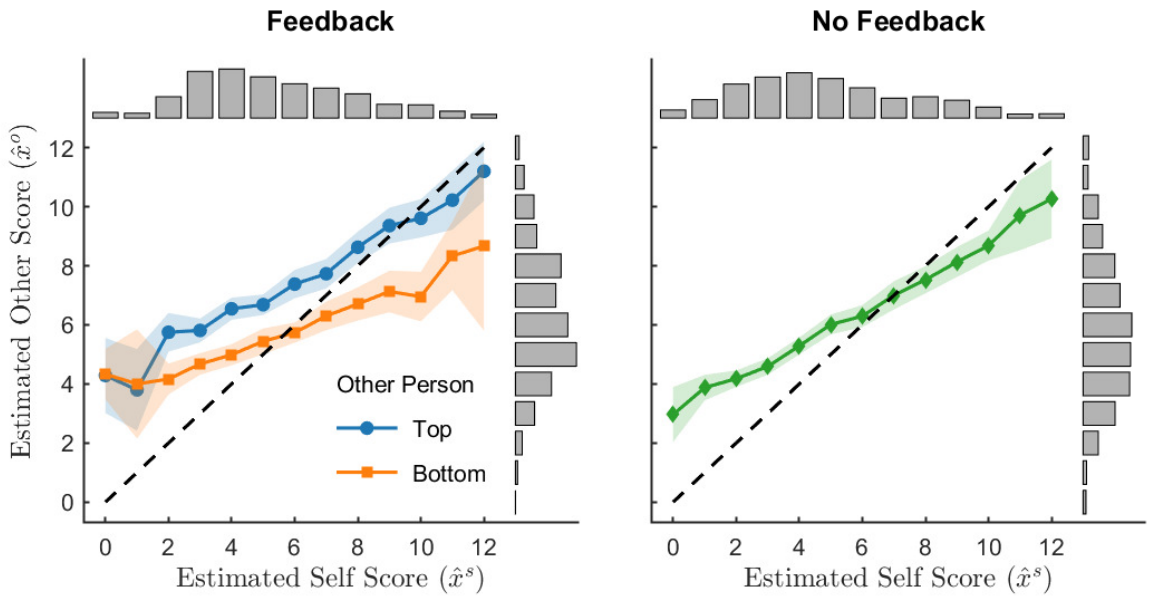


Figure 2.6: Estimated score for the other person ($\hat{x}^o$) conditional on the estimated self score ($\hat{x}^s$). The results for the feedback condition are separated by the overall performance of the other person. Histograms show the marginal distribution of scores. The colored areas show 95% confidence intervals.

bottom other performers diverges within a couple of feedback rounds. This suggests that people employ an efficient mental representation of the other that enables them to quickly distinguish their own performance from the other person's.

Our results are consistent with previous studies of predicting general knowledge in self and others (Jameson et al., 1993). Target participants in Experiment 1 were more accurate in assessing themselves than the observers in Experiment 2 who assessed the targets and received feedback. In turn, the observers who received feedback were more accurate than the observers who did not receive feedback. However, without feedback performance is still well above chance. Figure 2.6 hints that observers without feedback use their own perceived ability and their self-assessed problem difficulty as predictors, assuming that what is difficult for them is also difficult for another person. This guessing strategy is effective in situations where the perceived problem difficulty for self correlates with the actual problem difficulty faced by other people (Fussell & Krauss, 1991; Jameson et al., 1993; Nickerson et al., 1987).

2.5 Model-based Results

Our primary modeling objective is to understand the mechanisms at play when humans make inferences about the ability and performance of other individuals. To do so, we simulate the three qualitatively different models described above and relate them to the key empirical findings in our experiments. We use two methods to evaluate model adequacy. First, we perform a qualitative model evaluation by assessing the models' ability to replicate the qualitative patterns we observed in the empirical data. We do this through posterior predictive simulation. For all three hypotheses, we use the existing behavioral data from the set of participants and problem-sets to estimate posterior distributions of the parameters. We then simulate the behavior of new participants and new problem-sets by sampling from the posterior predictive distribution (i.e., these are predictions for a replication of the experiment with a new set of participants and new problem sets). We use this simulated data to compare the qualitative predictions of our models to our empirical findings on 1) the relationship between self- and other-assessment, and 2) people's ability to differentiate between good and bad performances of other participants when given feedback. Our second method for model evaluation is through out-of-sample predictive checks using cross-validation. In this approach, we use the posterior distributions for the actual set of participants and problem-sets in the experiments and compare the model predictions for held-out problem-sets against the observed data.

2.5.1 Relationship between self- and other-assessment

Previous investigations of neural-activity during self- and other-assessment (Frith & Frith, 1999; Jenkins et al., 2008; Mitchell et al., 2005) have revealed a close correspondence between people's metacognition and their theory of mind. Our empirical results also indicate that self-assessment is closely tied to other-assessment. Figure 2.7 shows the relationship between

self- and other-assessment as predicted by the three models. These results are based on a combination of experimental data and simulated data. We simulate participants' assessment of others' performance for four randomly assigned participants as their 'other persons'.

Compared to the observed empirical data in Figure 2.6, we see that the Differentiated-by-ability model (M_1) most closely captures the trend observed in the empirical data in both the feedback and no feedback conditions. When feedback is provided, it predicts a strong association between the self and other estimates while allowing for learning of differential ability of the other. This is consistent with what we see in our empirical data where people's estimates of their own performance are closely tied to their performance of the other. People draw on their experience with the task to make inferences about the other person's experience and assume that their subjective difficulty on any item must be commensurate to the difficulty experienced by the other person. Throughout the experiment, their estimates of the other person's performance are anchored by their own scores.

In contrast, without any informative priors about ability or difficulty, the fully differentiated model (M_2) fails to predict any association between self- and other-assessment. Alternatively, the undifferentiated model (M_3) relies too heavily on priors and predicts that people's estimates of others' performances are closely tied with their assessment of their own performance. Note that in the case of no feedback, M_1 is similar to M_3. With no information to learn from, people are forced to rely heavily on their own metacognitive assessments of their ability and difficulty of each item as a prior for the other person. Hence both models predict similar trends between self and other scores in the no feedback condition.

2.5.2 Differentiating between good and bad performers

In Figure 2.5 we observed that participants are able to distinguish between good and bad performances of other participants in the feedback condition. On the first trial, people

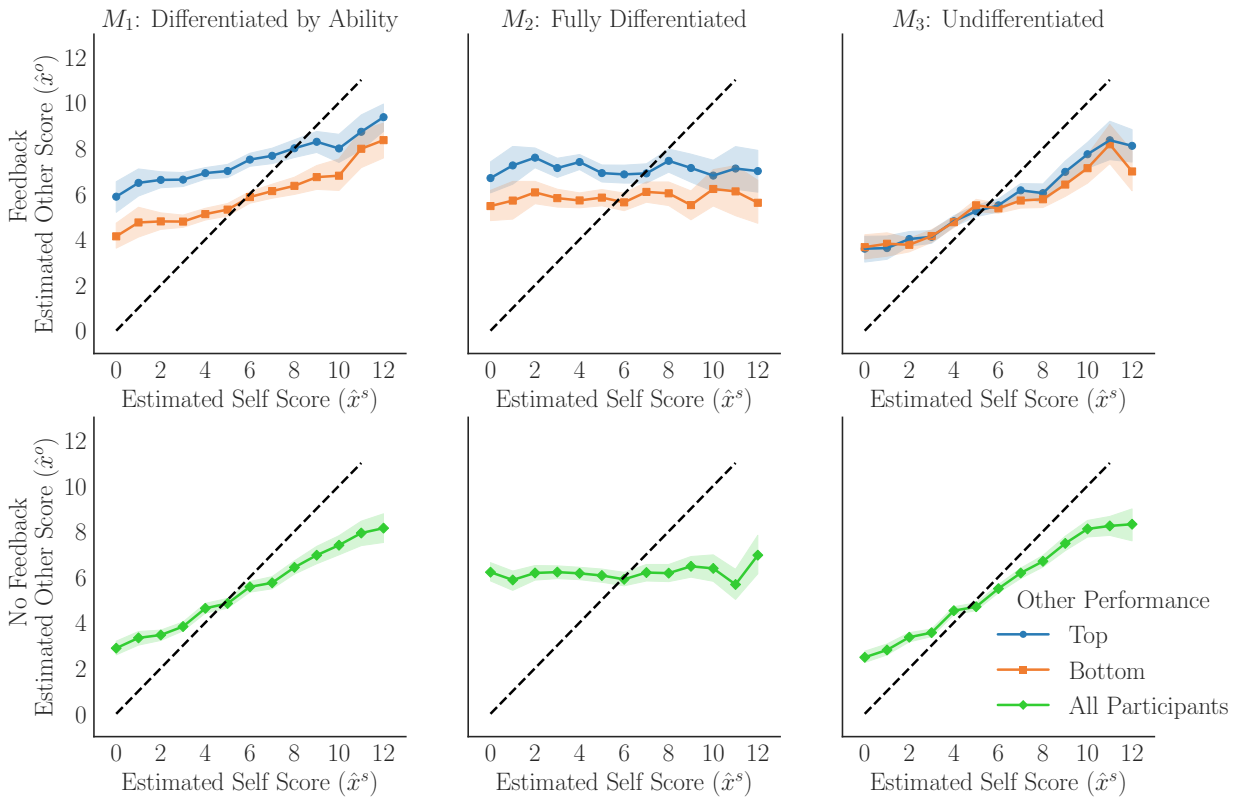


Figure 2.7: Model predictions for the relationship between estimated other score and estimated self performance. The results are separated by the feedback condition and performance levels of the other person. Note that in the no feedback condition, participants can't differentiate between top and bottom performers. Dashed line indicates exact equivalence between estimated self and other scores. The colored areas show 95% confidence intervals.

use their prior beliefs about the other person's ability and difficulty to estimate others' scores. Subsequently, in the presence of feedback, people adjust their beliefs about the other participant's ability to make their estimates. The corresponding model predictions are shown in Figure 2.8. The results show that the differentiated-by-ability model (M_1) accurately emulates this behavioral pattern. The simulated participants' estimates of the good and bad performances diverge after they receive a single data point as feedback. On the other hand, while M_2 does better than M_3 at capturing the dependence of other-assessment on self-assessment (Figure 2.7), it does not capture people's ability to learn and differentiate between good and bad performances by the other. This is an important feature of the feedback condition in our experiment - people quickly learn the differential ability of the

other person. Both M_2 and M_3 fail to capture this critical empirical feature.

Figure 2.8: Model predictions for the mean estimated score of the other person over problem-sets. The results are separated by the feedback condition and performance levels of the other person. Dashed lines show the mean true score across the top and bottom performing other people. The colored areas show 95% confidence intervals.

2.5.3 Quantitative Assessment of Model Performance

Table 2.3 shows how well each of the three models are able to capture the other-assessments in the empirical data. The sequential nature of our models allow us to make out-of-sample predictions for other-assessment at each time-step. For example, when making a prediction at time $t+1$, the model only receives information about the other person's true performance up to time t.

The table shows the mean squared error (MSE) and Pearson Correlation (ρ) between the predicted estimates of other-performance as evaluated by the models and the actual estimates of other-performance made by participants in the experiment. These values indicate how

	Across participants		Per participant		
Model	MSE	ρ	Mean MSE	Mean ρ	N
M_1: Differentiated by Ability	**8.92**	**0.39**	**8.92 (5.515, 12.324)**	**0.359 (0.241, 0.478)**	64
M_2: Fully Differentiated	15.95	0.15	15.95 (12.726, 19.172)	0.076 (-0.067, 0.219)	64
M_3: Undifferentiated	10.60	0.26	10.60 (7.254, 13.945)	0.276 (0.137, 0.414)	64

Table 2.3: Other-assessment across models M_1, M_2, and M_3. For analysis per participant, the statistics are calculated at the individual participant level and then averaged; numbers between parentheses are 95% confidence intervals. N is the number of participants. For the analysis across participants, we ignore individual differences and report a single outcome across participants and problem-sets.

closely model estimates resemble the true data. We only compare the models on their performance on the feedback condition. Overall, we see that the differentiated-by-ability model (M_1) outperforms the two other models (M_2 and M_3). This model provides the best quantitative fit to the data when the correspondence is assessed for each individual participant as well as across participants. Other statistics such as CPA follow the same trends as shown in Table 2.3. We focused on MSE because it is a standard way to evaluate the predictive performance of models.

2.5.4 Discussion of Model-based Results

We contrasted three models and assessed the ability of the models to capture the qualitative patterns as well as match the human predictions in a quantitative way. The best performing model was the differentiated-by-ability (M_1) model. It is a model with relatively few parameters that makes an assumption that there is a simple link between the mental model of self and other. Model M_1 learns only one differential ability parameter linking self- to other-assessment. Note that this is one of many ways to formulate how self- and other-assessment are tied together. Our claim is that for simpler tasks and with small amounts of data this link between self- and other-assessment remains low dimensional. How quickly these models grow in complexity needs to be explored in future work.

Predictions from the differentiated-by-ability model (M_1) replicate the qualitative pattern we see in our empirical results while also being quantitatively closest to the observed data as shown in Table 2.3. The other two models (M_2 and M_3) fail to simultaneously capture the relationship between estimated self- and other-scores (Figure 2.7) and the divergence of estimated scores for top and bottom performers (Figure 2.8). In contrast, in the absence of feedback, people only have their own encounter with the task to rely on. This reliance is best captured by models M_1 and M_3. In M_3, the estimated ability and problem difficulty are assumed to be the same for the other person, leading a person to predict similar performance in self- and other assessment.

2.6 Explaining Previous Empirical Findings on Knowledge Assessment

Up to this point, we have shown how the hierarchical knowledge assessment model can explain a variety of findings from an empirical paradigm that we specifically designed to test how people differentiate between their own and others' performance. However, the hierarchical model can also be applied to other empirical paradigms. In this section, we demonstrate the model's ability to explain how people's assessment of other's performance changes as different knowledge signals are made available to them (Tullis, 2018) and how people place themselves relative to others (Moore & Healy, 2008). For each of the experiments, we qualitatively compare model predictions from the hierarchical model to the observed data. The details of the simulations are presented in Appendices A.5 and A.6.

Metacognitive Cue Utilization for Knowledge Assessment

The availability of certain performance related signals influences people's assessment of their performance on a task (Jost et al., 1998; Nelson et al., 1998; Tullis, 2018). In addition to assessing one's own knowledge, (Nickerson, 1999) proposes that the same signals may also guide one's assessment of others. For example, when asked to assess another person's performance on a task without doing the task themselves, a person may rely on a vague feeling-of-knowing about the task. However, if the person does the task themselves before assessing another person, they have access to additional information about their performance through signals such as the time it takes for them to perform the task. This information may enable the person to make a more informed assessment of another person's performance on the same task. (Tullis, 2018) proposes a theory of knowledge estimation as cue utilization that builds upon these previous accounts on self and other knowledge assessment (Koriat, 1997; Nickerson, 1999; Thomas & Jacoby, 2013). In this theory, the degree of overlap between self-assessment and other-assessment depends on the cues available to oneself. These cues may depend on an individual's interactions with the task, information about the specific other person being assessed, or general information about the population.

Through a series of experiments, (Tullis, 2018) demonstrates that the bases and accuracy of assessment of others depends on the conditions under which the assessment is elicited. In Experiment 1 in (Tullis, 2018) , participants judged the percentage of other participants who would know the answer to a series of trivia questions. There were two experimental conditions. In the *answer before* condition, on each trial, participants first answered the trivia question and then subsequently estimated the proportion of other participants who would know the answer. In the *answer after* condition, participants first estimated for each trivia question the proportion of other participants who would know the answer and then answered the trivia questions. Experiment 2 included four experimental conditions. As in Experiment 1, participants were either required to answer trivia questions before estimating

other participants' performance or they were asked to estimate the other participants' performance without needing to answer the question themselves. In addition, feedback was manipulated: participants either did or did not received corrective feedback about the correct answer after answering each question. Table 2.4 describes the four conditions in Experiment 2 and the corresponding metacognitive signals available to the participants.

The left panels of Figures 2.9 and 2.10 summarize the key empirical findings. Results are reported as gamma correlations between 1) predictions of other's knowledge and the time needed for the person to answer the question themselves and 2) predictions of other's knowledge and the accuracy of the participant themselves. Figure 2.9A shows that participants' predictions of others' knowledge were more strongly tied to their own performance when they were required to answer trivia questions themselves before estimating others' knowledge on the same questions. This is consistent with our hypothesis that people draw information through the process of answering questions when assessing others. The results also show that participants' assessment of others' improved when they were provided feedback about the accuracy of their answer (left panel of Figure 2.9B). This additional cue helped participants better assess the difficulty of each question and hence make better assessments of others' performance. Moreover, negative gamma correlations between participant's predictions for others and the time they took to answer the questions suggests that participants expected others to perform worse on questions that took them longer to answer. This supports our assumption that participants use response time as a signal to assess the difficulty of problems and therefore to inform their assessment of others. However, there was no significant difference in this effect between the feedback and no-feedback conditions.

To apply the hierarchical knowledge assessment framework to the other-assessment task presented in (Tullis, 2018), we will assume that the experimental conditions determine which metacognitive cues or knowledge signals are available to a person when assessing themselves and others. We will use $x_{i,j}^{FK}$, $x_{i,j}^{RT}$, and $x_{i,j}^{ACC}$ to denote the three types of knowledge signals

potentially available to participant i for problem j: *feeling of knowing* (FK), *response time* (RT), and *performance feedback* (ACC) respectively. We assume that these knowledge signals are produced according to:

$$x_{i,j}^{FK} \sim f(p_{i,j}^{s}, \eta), \quad x_{i,j}^{RT} \sim g(p_{i,j}^{s}, \nu), \quad x_{i,j}^{ACC} \sim h(p_{i,j}^{s}) \tag{2.7}$$

where functions f, g, and h link the knowledge signals to a person i's estimate about their probability of being correct on problem j ($p_{i,j}^{s}$) and η, ν encode the noise in the mapping to the observed knowledge. The mappings encode simple monotonic relationships between the probability correct and the knowledge signals. For example, feeling-of-knowing ($x_{i,j}^{FK}$) is modeled as linearly related to $p_{i,j}^{s}$ - the more likely a person is correct, the stronger their feeling-of-knowing. In contrast, we expect people's response times $x_{i,j}^{RT}$ to be inversely related to $p_{i,j}^{s}$ - the longer it takes people to solve a problem the harder they think it is. Note that in this experimental setup, participants only have access to their estimates of their response time. They do not observe the response time of other participants.

In Experiment 1 in (Tullis, 2018), in the answer after condition, participants judge other participants' performance before answering the question themselves, and hence participants only have a feeling of knowing signal available to make knowledge assessments, i.e. $x_{i,j}^{s} = \{x_{i,j}^{FK}\}$. In contrast, in the answer before condition, participants are required to answer the questions before evaluating others. Therefore, they have access to their response time in addition to the FK signal, i.e. $x_{i,j}^{s} = \{x_{i,j}^{FK}, x_{i,j}^{RT}\}$. Table 2.4 details the assumptions about the types of knowledge signals available to people across different conditions and experiments.

In the experimental task, participants have to estimate the percentage of other participants who know the answer to a series of trivia questions. This can be thought of as assessing the performance of an average person instead of a specific individual. Since participants do not have access to any knowledge signals (x^{o}) pertaining to the other person, they can only make

estimates about an average other person. In the absence of x^o, our modeling setup assumes that a^o is a random draw from the population and hence represents the ability of an average person. Therefore, we frame the inference problem for the participant to estimate a^o and problem difficulty d on the basis of the observed knowledge signals x^s. Since we do not have access to the raw experimental data from the paper, we first simulate experimental data for Experiments 1 and 2 using simple assumptions about individual differences in ability, variability of question difficulty as well as basic assumptions about the functional forms used in Eq. 2.7. Next, we apply the differentiated by ability model to simulate the inference process based on the simulated experimental data (see Appendix A.5 for details). The qualitative results shown here do not depend critically on the choice of simulation parameters.

Our model's predictions closely track the qualitative trends observed in the experimental data for Experiments 1 and 2, as demonstrated in Figure 2.9. In Figure 2.9A, the model predictions are consistent with the empirical observation that participants in the answer before condition showed a significantly stronger negative correlation between the time they took to answer a question and their accuracy of other assessment than participants in the answer after condition (i.e., participants estimated lower scores for others on questions that took them longer to answer). Additionally, the model predicts a positive correlation between participants' accuracy and their predictions of others' knowledge (i.e., participants tend to estimate higher scores for others on questions they themselves answered correctly). Similarly, for Experiment 2 (2.9B), the model predicts that participants estimate lower scores for others on questions that took them longer to answer. This effect is stronger in the feedback condition than in the no feedback condition. Additionally, the model captures the finding that participants tend to estimate higher scores for others on questions they themselves answered correctly. Figure 2.10 shows that the model predicts, consistent with the empirical observations, that participants' estimates of others improved when they were required to answer the question themselves and then were provided feedback. Overall, these results show that our model is able to accurately capture knowledge assessment across different

experimental conditions.

	Condition	Types of Knowledge Signals
Exp 1	Answer After	FK
	Answer Before	FK, RT
Exp 2	Answer Not Required, Feedback Not Given	FK
	Answer Not Required, Feedback Given	FK, ACC
	Answer Required, Feedback Not Given	FK, RT
	Answer Required, Feedback Given	FK, RT, ACC

Table 2.4: Assumptions about the types of knowledge signals available to people for the different conditions in Experiment 1 and 2 in Tullis (2018). FK=Feeling of Knowing; RT=Response Time; ACC=Accuracy

Overestimation and Overplacement

People's assessment of their own performance and the performance of others is known to be biased in several ways (Dunning, 2011; Larrick et al., 2007; Moore, 2007; Moore & Healy, 2008; Tullis, 2018). In particular, people tend to believe that they are less likely than average to exhibit extraordinary abilities and more likely than average to exhibit ordinary abilities (Moore, 2007). These beliefs about ability also depend on task difficulty.

(Moore & Healy, 2008) showed that on difficult tasks, people tend to overestimate their performance but incorrectly believe that they are worse than others. Whereas, on easy tasks, people tend to underestimate their performance but incorrectly believe they are better than others (Dunning, 2011; Moore & Healy, 2008). These findings can be attributed to two forms of overconfidence that people often display: *overestimation* and *overplacement*. For example, in the experimental paradigm from (Moore & Healy, 2008), participants answered trivia questions and predicted their own score and the score of a randomly selected participant at three different stages of the experiment. First, participants made predictions about themselves and the other participant before they had any specific information about the quiz they were about to take. Second, they answered quiz questions and then estimated their own scores and the other participant's score again. This is termed their *interim* estimate. Finally,

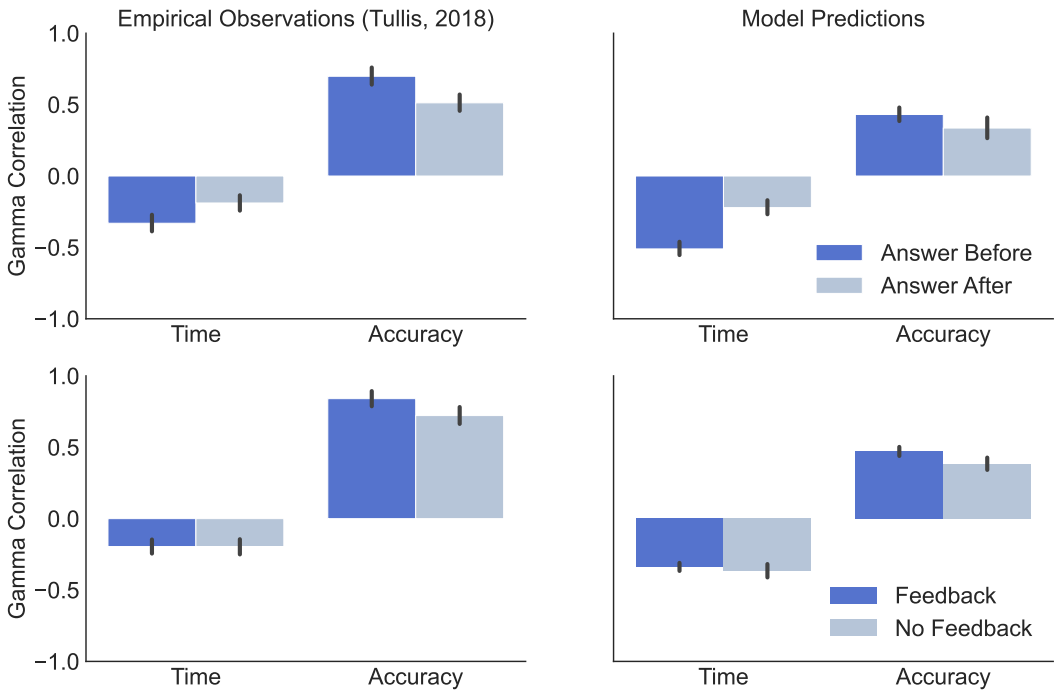


Figure 2.9: Observed and model-predicted correlations between a person's prediction of others' knowledge and the time needed for the person to answer the question themselves and their accuracy. The observed data is from (Tullis, 2018). The top row shows the results from the answer before and answer after conditions in Experiment 1. The bottom row shows results for the feedback and no feedback conditions in Experiment 2.

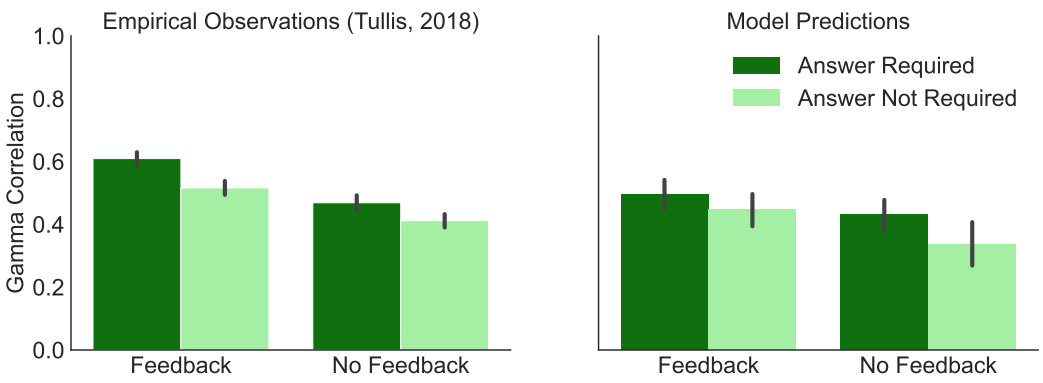


Figure 2.10: Observed and model-predicted correlations between a person's prediction about others' knowledge and the (sign reversed) difficulty of the questions. The observed data is from Experiment 2 from (Tullis, 2018) across the feedback and no feedback conditions. Note that the difficulty of a question for the empirical observations was based on the empirical proportion of participants that answered the question correctly. For the model predictions, the difficulty of the questions is the inferred latent difficulty.

participants were shown the correct answers to the quiz and asked to make final estimates about their performance and the other participant's performance.

The hierarchical knowledge assessment model is consistent with the theory presented by (Moore & Healy, 2008). The authors present a theory of overconfidence which assumes that people have imperfect information about their own performances and even worse information about the performances of others. As a result, people's estimates of themselves are regressive, but their estimates of others are even more regressive. The left panel of Figure 2.11 exemplifies the theory's prediction of participants' regressive estimates about performance of self and others. The right panel of Figure 2.11 demonstrates that our model predictions are consistent with the predictions of their theory of overconfidence and the empirical data presented in (Moore & Healy, 2008)- people's estimates of others' performance are more regressive than their estimates of their own performance. This qualitative trend is observed for a broad range of parameter values in our simulations. The main difference between the two theories is that the hierarchical model was designed to apply to a broader variety of empirical manipulations and tasks. The hierarchical framework provides explicit ways to model manipulations of

problem difficulty, feedback, ordering of answering relative to other assessment, as well as situations that lead to knowledge signals specific to other people.

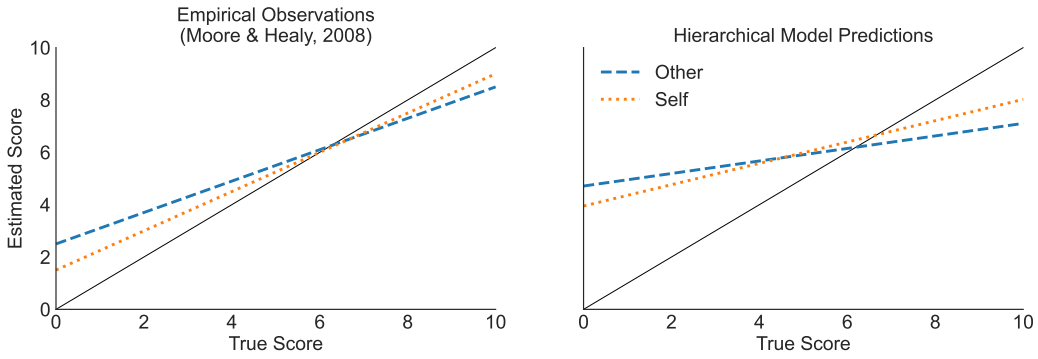


Figure 2.11: Relationship between the estimated performance of self and other and true performance of self and other as predicted by the theory of overconfidence (Moore & Healy, 2008) and as predicted by the hierarchical model.

The empirical observation columns in Table 2.5 show the degree of participants' overplacement and overestimation in the interim phase of the experiment. Higher positive values correspond to higher levels of overestimation and overplacement, and negative values correspond to underestimation and underplacement. The degree of overestimation was evaluated by the difference between the estimate of their performance and the person's true performance (i.e., $\hat{x}^{s,ACC} - x^{s,ACC}$). The degree of overplacement was evaluated by a difference of two differences: first, the difference between the estimated performance of self and other and second, the difference between the actual performance of self and other (i.e., $(\hat{x}^{s,ACC} - \hat{x}^{o,ACC}) - (x^{s,ACC} - x^{o,ACC})$). This can be understood as the difference between a person's estimate of how much better they are when compared to another person $(\hat{x}^{s,ACC} - \hat{x}^{o,ACC})$ and the true difference between the two people $(x^{s,ACC} - x^{o,ACC})$. The empirical results show that participants tend to overestimate their performance on hard problems and underestimate their performance on easier problems. Furthermore, participants overplace their performance on easy problems and underplace their performance on difficult problems.

We simulated the hierarchical knowledge assessment model for the interim stage of the experiment using the same setup and simulation parameters as used for the simulations

	Overestimation		Overplacement	
Difficulty	Empirical Observations	Model Predictions	Empirical Observations	Model Predictions
Easy	-.22 (.93)	-1.09 (2.27)	.48 (2.59)	.06 (2.53)
Medium	.01 (1.27)	2.2 (3.04)	.04 (3.91)	-.01 (3.36)
Hard	.79 (1.50)	4.33 (2.66)	-1.36 (2.39)	-.41 (2.78)

Table 2.5: Empirical observations from (Moore & Healy, 2008)and model predictions for *overestimation* and *overplacement* when making self and other knowledge assessment at the interim phase for three different question difficulties (standard deviations in parentheses).

of the (Tullis, 2018) experiments (See Appendix A.6 for details). At the interim stage of the experiment, we assume that participants have access to feeling-of-knowing and response time signals, similar to the answer-before condition in Experiment 1 of (Tullis, 2018), i.e. $x^s_{i,j} = \{x^{FK}_{i,j}, x^{RT}_{i,j}\}$. We use the model to simulate the knowledge signals available to participants in the experiment. We also simulate a distribution of problem difficulty and refer to the highest 33% difficulty values as hard, the lowest 33% as easy, and the rest as medium. Next, we simulate the task faced by the participant: the problem of inferring $x^{s,ACC}$ and $x^{o,ACC}$ (i.e., producing estimates $\hat{x}^{s,ACC}$, $\hat{x}^{o,ACC}$) given the available knowledge signals x^s. Finally, to analyze the model predictions, we assess the degree of overestimation and overplacement using the same evaluation approach used to analyze the empirical data. The model prediction in Table 2.5 demonstrate our model's ability to capture the relationship between task difficulty and people's tendency to overplace or overestimate their performance. In line with the empirical observations, our model predicts that people underplace but overestimate their performance on difficult problems, and people overplace and underestimate their performance on easy problems.

At first glance, it may seem that the quantitative predictions of our model in Table 2.5 significantly diverge from the empirical observations. However, it is important to recognize that the empirical effects of overplacement and underplacement reported in (Moore & Healy, 2008) are relatively small. Our primary objective was not to achieve exact quantitative matches but rather to demonstrate that the hierarchical model makes qualitatively accurate predictions for self- and other-assessment phenomena reported in the literature. Additionally,

it is worth noting that we use the same set of assumptions and parameter values to simulate data for all the experiments from (Tullis, 2018) and (Moore & Healy, 2008).

2.7 Extending to Humans' Assessment of AI Ability

Based on the paper 'Capturing Humans' Mental Models of AI: An Item Response Theory Approach' (Kelly et al., 2023). This work was led by M. Kelly. Author Contributions: M.K., A.K., M.S., and P.S. designed research; M.K. performed research; M.K. analyzed data; and M.K., M.S., and P.S. wrote the paper.

A person being advised by an AI agent should be able to accurately perceive the AI's strengths and weaknesses to make good use of its assistance, i.e, they should build an accurate model of the AI's capabilities. Understanding the mental models humans build of AI agents can help predict the development of appropriate trust in AI, when they will defer to an AI agent, and how a human AI team will function overall. We apply our proposed model of knowledge assessment to understand how people assess AI agents. To do this, we conduct another experiment where participants assess themselves, and either a human or an AI agent.

2.7.1 Sequential Knowledge Assessment: Trivia Question-Answering Task

Participants complete a multi-category trivia question-answering task and estimate the performance of either another human or an AI agent. Experiments were run via Amazon Mechanical Turk. Each participant answered 16 sets of 12 trivia questions. We chose a trivia setting because it does not require specialized knowledge (and thus is doable by Mechanical Turk workers), is discriminative (it is very unlikely humans or AI will answer all trivia

questions correctly) (Boyd-Graber & Börschinger, 2019), and can be broken up into distinct categories, allowing us to directly investigate multiple dimensions of ability.

Using a dataset of trivia questions from The Question Company, we selected four question topics: *History of Art*, *Video Games*, *Cities*, and *Math*. We chose these topics to achieve, based on preliminary pilot data, (1) variation between participants for each topic, (2) variation among topics for each participant, and (3) varying correlations between pairs of topics across participants. We also chose topics that we expected people would *perceive* as different, e.g. requiring different types of knowledge.

For each category, we created four problem sets of 12 questions each, for a total of 16 problem sets. Participants were presented with questions in four rounds. Each round included one problem set from each trivia category. For each participant, we randomized the order of questions within each problem set, the order of categories within rounds, and the order of problem sets across rounds. Similar to the previous task, after each 12-question problem set, participants were asked to estimate their own performance, to capture self-assessment, and the performance of another agent to capture other-assessment. Participants were randomly assigned to assess either an "AI system" or another person.

Two-thirds of the participants were randomly selected to receive feedback regarding their performance estimation: after they provided their estimates for a given problem set, they were shown the actual performance of themselves and the other agent. The remaining one-third did not receive this feedback. The no-feedback group was included to capture prior expectations and the strategies people use to estimate performance in the absence of feedback.

To obtain the performance data for the other humans, we first performed a pilot study ($n = 34$) using the same problem sets. We selected the top five and bottom five participants (based on overall accuracy). Similar to the previous experiment, participants in the main experiment who were assigned to the other human condition were shown performance data

from one of these 10 participants. The name of this other human was randomly chosen from a set of ten names drawn from a random name generator (e.g. “Anna” or “Felix”).

To reduce the possibility of performance as a confounding variable, we then matched AI performance with the other human performance on a topic-wise basis. This was done by running several variants of UnifiedQA (Khashabi et al., 2020) and Zero-shot-CoT (Kojima et al., 2022), which are large language models designed to generalize to a range of tasks. We then chose two models for each topic, one with similar performance to that of the top five other humans, and another to match the performance of the bottom five other humans. We include agents with both high and low accuracy to improve the generalization of results; mental models could differ depending on whether the other agent has higher or lower performance than the participant. Participants were assigned to an agent type (other human or AI agent) and agent category (high accuracy, low accuracy, or no feedback). Participants were evenly divided between these six experimental condition combinations.

2.7.2 Multidimensional Extension of the Hierarchical Framework

An important simplification in the self- and other-assessment models is that they encode ability as a one-dimensional parameter. We focused on a simple mental model where differentiation was based on a single-dimensional ability. However, we don’t rule out the possibility that people are developing increasingly complex multidimensional mental models of others, as more information is observed.

A straightforward extension of the self- and other-assessment models is to account for differences in ability across different categories presented to the participant. Multidimensional Item Response Theory (MIRT) is often used to analyze performance on tasks where multiple abilities are at play (Ackerman et al., 2003; Hartig & Höhler, 2009). MIRT is a generalization of unidimensional IRT models where the probability of success is modeled as a function of

multiple ability dimensions. Such models can also be applied to instances where mixtures of abilities are required for individual test items.

We use an extension of the hierarchical framework where we use a multidimensional structure for the underlying, self-assessed, and other assessed abilities; δ is a k-dimensional vector capturing topic-wise ability differences (See Appendix A.7 for details).

2.7.3 Results

Empirical Results

Figure 2.12 shows evidence of a positive correlation between self-assessment and other-assessment, with the closest correspondence for other humans. Although there is a relationship between self- and other-assessment for both types of other agents, the ability differential is, on average, higher for AI agents. In other words, in the absence of feedback, participants use their own ability as a starting point for estimating the other agent's performance (there is a positive correlation for both types of agent), but expect AI agents to perform much better than other humans.

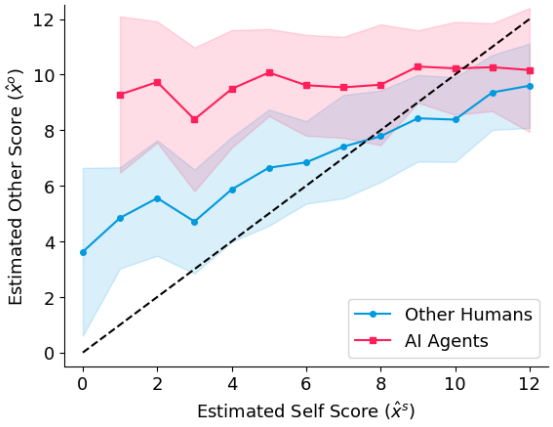


Figure 2.12: Relationship between self-assessed score and other-assessed score in the no-feedback condition. The dotted black diagonal line represents identical self- and other-assessment.

This suggests that people may believe AI has "superpowers" beyond human capabilities; regardless of how participants perceive their performance on a problem set, they expect the AI agent to perform very well on that problem set. This phenomenon is most pronounced when the self-assessed score is low—people tend to overestimate the performance of AI agents most substantially when they perceive their own performance to be poor. In contrast, when participants perceive that they have done poorly on a particular problem set, they accordingly reduce their expectations of other humans.

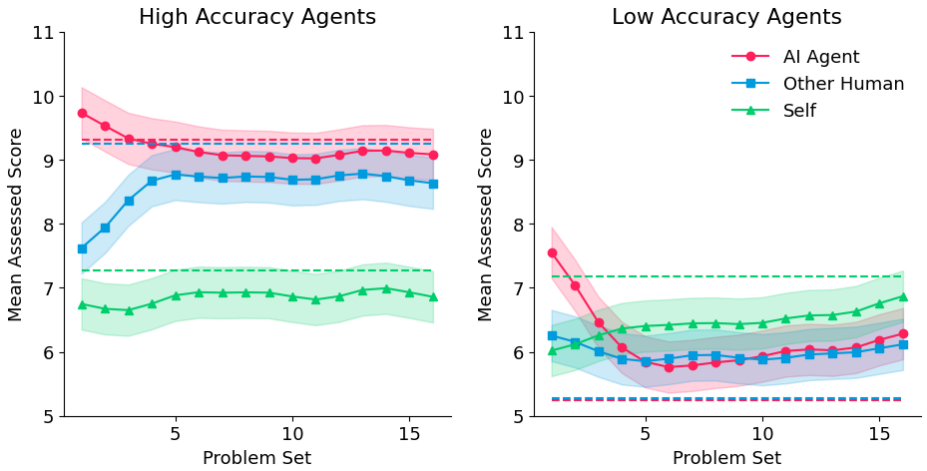


Figure 2.13: Mean assessments of the performance of other agents and self at each problem set. The results are separated by other agents with high accuracy and low accuracy. Dashed lines show corresponding values of actual performance for reference (for self, AI agent, and other human). Results are smoothed across problem-sets to facilitate visual comparison.

Figure 2.13 illustrates how participants update their assessments over rounds of feedback, starting from problem set 1 (when no feedback has been provided yet) up to problem set 16 (when feedback has been provided about all previous 15 problem sets). Participants adapt their assessments of other agents (both AI and human, both high and low accuracy) quickly within the first quarter of problem sets and then change relatively slowly after that. Even

after feedback from 16 problem sets, for the high-accuracy agents, participants systematically assessed AI agents as being roughly 0.6 points more accurate than the human agents, even though the two agents were selected to have approximately the same accuracy (dotted lines).

Model-based Results

For both AI agents and other people, we train three different MIRT models, one for each hypothesized hierarchical structure. To capture mental model development over time, we evaluate models based on next-round predictions, that is, we compute the log-likelihood for round t using a model trained on data from rounds 1 to $t-1$. These next-round log-likelihoods, under the feedback condition, are shown in Table 2.6. There is strong evidence that people differentiate between themselves and other agents. Consistent with the results presented in Section 2.5, when assessing other humans, the differentiated-by-ability model fits the data best, suggesting that participants assess other humans' abilities relative to their own abilities. In contrast, the fully differentiated model best explains the perceived scores of AI agents; this provides evidence that participants' mental models of themselves were less relevant in developing mental models of AI agents (in comparison to those of other humans).

Model	**Humans**	**AI**
M_1: Differentiated by Ability	**-2.00**	-2.18
M_2: Fully Differentiated	-2.13	**-2.13**
M_3: Undifferentiated	-2.69	-3.36

Table 2.6: Held-out next-round log-likelihoods (higher is better) for the different models of knowledge assessment in the feedback condition.

2.7.4 Discussion of Empirical & Model-based Results

We observe that mental models of an AI agent's ability are highly differentiated from self-perceived ability (e.g., Figures 2.12 and 2.13 and Table 2.6). On average, participants expect

AI agents to perform very differently from themselves, especially in the absence of feedback. This bias could lead to under- or over-reliance on an AI agent in a team setting; additional work is needed to better understand and counteract it.

In our experiments, participants' other-assessments did not fully converge to AI agents' true performances, even given feedback (see Figure 2.13). This phenomenon could serve as motivation to give the teammates of an AI agent extra information to aid their mental model development, e.g., a "primer" or onboarding process or prediction explanations.

In summary, our findings indicate that people tend to overestimate the performance of AI agents relative to their own performance and that their mental models fail to develop completely even when given feedback.

2.8 Discussion

Knowing what other agents know is central to communication and cooperation between agents. Much of the current computational work on theory of mind has focused on inferring beliefs and goals of other people by observing intentional behavior in spatial environments (Baker et al., 2017; Baker et al., 2009). However, developing an accurate model of another agent not only requires an understanding of their goals and beliefs which can explain their movements in a physical environment but also their knowledge states which can explain their performance on knowledge tasks. In our theoretical framework, we focus on understanding how people assess the knowledge states of other people in the absence of any physical or verbal cues — they only receive quantitative feedback about their assessment of the other person's performance. The key idea of our work is that people combine their own experience on a task with information received about the other person's performance to make assessments of the other's knowledge states.

Previous research to understand how humans infer knowledge states of other humans was limited to empirical studies (Jameson et al., 1993; Nelson, 1984) and descriptive theories (Nickerson, 1999). However, there is increasing interest in developing models of reasoning about other people's knowledge states (Aboody et al., 2021; Berke & Jara-Ettinger, 2021). (Aboody et al., 2021) present a computational account of how people infer knowledge of another person based on the expectation that the other person maximises epistemic utility when making choices. In this research, we take a complementary view of knowledge assessment of others. Our framework formalizes how humans construct mental models of other humans' knowledge solely based on observed quantitative performance of the other person. We developed and tested three computational models on the basis of a simple empirical paradigm where the participant is asked to make inferences about the other person. As the experiment progresses, limited information about the other person is made available to the participant. For example, after receiving feedback about their first prediction, there is only one data point about the other person that is available to the participant. Still, despite the small amount of information, participants are able to update their mental model of the other person and improve their predictions over subsequent prediction rounds. We suggest that there are two main components that drive people's estimation of the other person's performance. The first is people's tendency to generalise their experience with the task to the other person's behavior. This explains the close association between people's self and other estimates - people use their estimates of task difficulty to adjust their beliefs about the other person's performance. The second component is their capacity to distinguish between their own ability and the other person's ability. This is made apparent by people's quickly diverging estimates of top- and bottom - other performers in our experiment.

2.8.1 Sparse Data Encourages Linking Mental Models of Self and Other

From a computational perspective, people are often faced with situations where not many observations are available about another individual, making it difficult to learn detailed and complex mental models of that individual. Instead, a simpler mental model with few parameters to estimate might be effective (at least in the initial interaction with the individual). In this research, we contrasted three computational models for the inference of knowledge states. The models varied in the degree to which the mental models of self- and other are differentiated. In the simplest mental model of other (M_3; undifferentiated), no parameters need to be updated as the mental model for the other person is the same as the mental model for self. In the most complex mental model of other (M_2, fully differentiated), not only the ability of the other person needs to be estimated but also the experienced difficulty for each type of problem. This model allows for the possibility that what is easy for one's self could be challenging for the other and vice versa. We found evidence for an computational model with an intermediate level of complexity (M_1; differentiated by ability) that involves just a single parameter: the relative ability of the other individual. This simple mental model allows one to quickly extrapolate how likely it is that an individual can successfully perform a task with very few observations.

Our results support our claim that in the presence of feedback, people learn about the other person's ability relative to their own while also drawing information from their own experience from the task. The differentiated-by-ability model that best accounts for the observed data makes the assumption that the way people reason about the other person's performance is through the lens of their own self-assessment process. This assumption is consistent with a second-order model of metacognition which suggests that humans self-reflect and think about others using similar mental processes (Fleming & Daw, 2017). We posit that the same machinery that enables people to estimate their performance also enables them to judge

another person's performance. However, we do not address the issue of the number of systems involved in metacognition and mindreading. Our results simply point out that self-knowledge can be informative and is used by people to make predictions about other people's knowledge.

2.8.2 Proposals for Future Investigations

We now discuss in greater detail how the self- and other-assessment can be extended to handle other interesting situations involving social cues, multiple agents, multidimensional ability, and AI agents assessed by humans and humans assessing AI agents.

Utilizing Social Cues to Assess Others

During social interactions, people have the opportunity to perceive and interpret numerous signals such as facial expressions, natural language, and voice intonations of the person they interact with. These cues can serve as supplementary information when evaluating the other person's knowledge.

An experiment was conducted by (Jameson et al., 1993) where 'observer' participants witnessed 'target' participants answering trivia questions and then made predictions about the targets' performance on those questions. In contrast, 'judge' participants made predictions without observing the targets answering the questions. The findings revealed that observers displayed greater accuracy in predicting the performance of the individuals compared to judges. This divergence in prediction accuracy can be attributed to specific cues exhibited by the observed person which provide additional insights beyond mere performance statistics and task experience. These cues may include the time taken by the target to respond to a question, their facial expressions, the confidence conveyed through their voice, and possibly other factors. Similarly, (Brennan & Williams, 1995) assigned participants the task of listening

to responses given by others to general knowledge questions and evaluated their perception of the "feeling-of-another's-knowing." The results demonstrated that people's assessments of others' knowledge were influenced by changes in intonation, response delays, and the use of filler phrases — confirming that individuals pay attention to the metacognitive information conveyed by speakers regarding their states of knowledge.

Most recent computational approaches to capture knowledge assessment, including our ongoing research, have focused on situations in which individuals possess limited information regarding others. However, there is a need to explore how an expanded set of behavioral and social cues can be quantified and utilized as supplementary information for predicting people's accuracy of other-assessment.

Assessing Multiple Other Agents

More often than not, people work with multiple other agents to accomplish tasks. An important extension of the current work is to see how easily peoples' mental models scale to groups of others, or how well can people make inferences about knowledge states of multiple other teammates when working in a group. For example, when playing a trivia quiz with a group of people, players continuously appraise other players' expertise on a variety of domains. This mechanism of group appraisal and coordination was formalised by (Wegner, 1987) as a transactive memory system (TMS). TMS is a property of a group that consists of knowledge stored in each person's memory and metamemory that encodes different teammates' domains of expertise. (Mei et al., 2017) mathematically formalize TMS as an appraisal network and describe asymptotic properties of the team. However, how people learn such an appraisal network in practice is not well investigated. Here we focused on assessing only one other person and the model that best described the empirical data was a low-dimensional model. It is likely that humans learn a sparse representation of ability to differentiate between multiple teammates. Such parsimony would be essential to manage cognitive overload and resource

constraints.

2.8.3 Conclusions

How a mind understands another mind is a fundamental question in psychology. While there is prior research on how people make theory of mind judgments about the intentions and goals of other agents, there is relatively little investigation of how people assess knowledge of other agents. In this work, we develop a theoretical framework that describes the underlying computation that people employ when assessing the knowledge of other agents. Our empirical results and model predictions demonstrate that people's evaluation of the other person's performance (a theory of mind computation) is linked to their evaluation of their own performance (a metacognitive computation). The models presented in this chapter provide a starting point for a more comprehensive exploration of how humans assess other agents. We also extend our model to analyze people's mental models of AI agents. Our findings indicate that people tend to over-estimate the performance of AI agents relative to their own performance and their mental models fail to develop completely. We anticipate that our modeling framework, and these findings will be useful in both understanding and improving interaction in hybrid human-AI teams.

Chapter 3

Impact of Differing Perspectives in Advice-Taking with Human and AI Advisors

3.1 Abstract

Advice-taking plays a critical role in collaboration. Yet people tend to under-utilize advice, often to their own detriment. In this chapter, we investigate if people's utilization of advice improves when they know the advisor has access to different information compared to them. Across four experiments, we examine how individuals integrate advice in an estimation task, where the advisee and the advisor have access to different perspectives of the same problem. Experiment 1 evaluates participants' independent ability to perform the task of estimating the number of objects in a jar from various viewpoints. The results show that depending on the type of objects, some viewpoints provide more accurate estimates than others. Experiment 2 confirms that participants are able to identify these more informative viewpoints. Building

on these findings, Experiment 3 examines how individuals adjust their estimates when presented with estimates from a human advisor and an AI advisor. Individuals are also given information about the advisor's perspective. Our findings are consistent with egocentric discounting where individuals exhibit a general bias toward their own information. However, this discounting is lower for AI advisors compared to human advisors in our experiment. Our results also show that the advisor's estimate is taken more into account when the advisor has a more favorable viewpoint – for both human and AI advisors. This suggests a potential for optimizing advice-taking behavior by enhancing people's understanding of the advisor's viewpoint. This study furthers our understanding of advice-taking dynamics with human and AI advisors and the role of perspective-taking in decision-making processes.

3.2 Introduction

Research on advice-taking reveals a common tendency: people undervalue suggestions from others compared to their own opinions (Bonaccio & Dalal, 2006) despite evidence that advice typically improves performance (Kämmer et al., 2023). Concurrently, extensive research on overconfidence shows that people perceive their abilities as superior to those of their peers, exhibiting unwarranted confidence in themselves (Moore & Cain, 2007). Working with AI is no different. Research has shown that humans are susceptible to a variety of misjudgements and biases when seeking advice from machines Dietvorst et al., 2015; Goddard et al., 2012; Logg et al., 2019. As humans interface more with AI assistants, it is important to understand how they incorporate AI advice in their decisions. This combination of undervaluing advice and overvaluing personal judgment presents a significant challenge in decision-making and advice utilization.

Several explanations have been proposed to explain why people may underutilize advice. Some suggest that people fail to appreciate the benefit of incorporating diverse perspectives

(Larrick & Soll, 2006), while others attribute it to individual differences in agency (Schultze et al., 2018) or narcissism (Kausel et al., 2015). This tendency to discount others' opinions compared to your own is termed egocentric discounting (Yaniv, 2004; Yaniv & Kleinberger, 2000). It posits that individuals tend to favor their judgment over advice from others and struggle to accurately incorporate alternative viewpoints when evaluating advice. This is further confirmed by (Bailey et al., 2022) through a meta-analysis that shows people generally give more weight to their own beliefs than to advice from others. Interestingly, some studies show that people are more likely to discount advice that significantly differs from their initial beliefs (Himmelstein & Budescu, 2023; Yaniv & Milyavsky, 2007), while others show that advice that mirrors one's own beliefs is also discounted, being perceived as confirmation rather than new information (Allahverdyan & Galstyan, 2014; Himmelstein & Budescu, 2023). This adds complexity to understanding how different degrees of similarity between advisor and advisee opinions influence the advice-taking process.

Previous research also indicates that people place greater weight on advice when they perceive the advisor to be an expert based on previous interactions with the advisor (Önkal et al., 2017). Beyond egocentric discounting, advisor reputation formation is another critical factor shaping people's advice-taking behavior (Yaniv & Kleinberger, 2000). While the former deals with the inherent human tendency to prioritize personal beliefs, advisor reputation emphasizes the dynamic nature of trust and the importance of past interactions in shaping decisions to rely on advice. However, these findings are based on scenarios where the advisee and the advisor do the same task and have access to identical information. In contrast, the key focus of our work is how individuals integrate advice from others when the information available to the advisor differs from their own. Do people still exhibit egocentric discounting in such scenarios? Does this preference for one's own judgment over others' advice indicate a failure in perspective-taking? Or do individuals possess the capacity to discern between various forms of advice, selectively incorporating them based on the relevance of the advisor's information?

Through a meta-analysis, (Bailey et al., 2022) found that the information about the *quality of advice* was the most significant predictor of advice-taking. In contrast, the key focus of our work is to investigate if people can strategically alter their reliance on the advisor depending on the *quality of information* the advisor has. In particular, we investigate how visual perspective-taking influences advice integration in a visual quantity estimation task. Research on visual perspective taking has demonstrated that people are capable of accurately predicting the visual experience of other agents (Michelon & Zacks, 2006). However, it is unclear if this ability extends to advice integration, especially when the advisor has a superior vantage point to address a particular problem. Our research probes the interplay between egocentric discounting and perspective-taking in the context of advice integration. Through a series of experiments, we examine how individuals integrate the advisor's estimate when individuals know the quality of information available to the advisor. We also contrast people's integration of advice across human and AI advisors.

In our experimental setup, an individual, referred to as the advisee, is tasked with estimating the number of objects in a jar, as shown in Figure 3.2. They are then provided with the estimate made by an advisor, who may be a human or an AI agent, regarding the same jar's contents. This setup mirrors a judge-advisor system (JAS) a widely used framework in decision-making research (Bailey et al., 2022; Himmelstein & Budescu, 2023; Önkal et al., 2017), where a judge initially makes a judgment and then has the chance to revise it after receiving advice. Unlike traditional JAS setups, our experiment presents the advisor and advisee with different views of the jar. These distinct viewpoints make the process of counting the objects more or less challenging.

The implications of our research extend beyond the scope of our experiments. They are relevant in various real-world contexts, including decision-making assisted by AI systems and collaborative problem-solving. The increasing prevalence of AI in decision-making (Grgić-Hlača et al., 2019; Patel et al., 2019; Steyvers & Kumar, 2022) necessitates understanding

the interaction between humans and AI advisors. This includes exploring different workflows, such as making AI advice always available or available only on demand to improve the efficacy of assisted decision-making. Advances in explainability methods also play a critical role in demystifying AI decisions for users (Ignatiev, 2020).

In the following sections, we present details about and insights from three experiments that seek to answer the following questions: Can the advisee, when informed about the advisor's view of the jar, accurately assess the difference in favorability between their view and the advisor's? When given the opportunity to revise their estimate, do individuals appropriately integrate the advisor's assessment into their estimation? How does people's weighting of advice differ across human and AI advisors? Furthermore, do individuals make optimal decisions based on who possesses the more advantageous view? Finally, we discuss the broader significance of this research.

3.3 Overview of Experiments

The goal of our experiments is to assess advice-taking in situations where there is information asymmetry between the advisor and the advisee. We conduct three experiments to explore this theme. In all three experiments, the primary task is to estimate the number of objects in a jar, a classic task used to study the psychology of number estimation (Bevan et al., 1963) and conformity (Jenness, 1932). To facilitate this task, we create stimuli consisting of images of jars containing varying quantities of cylinders, disks, or spheres. Each jar may be viewed from five different angles, as shown in Figure 3.2: 0° (side view), $22^\circ, 45^\circ, 66^\circ$ (intermediate viewing angles), and 90° (top view). Our choice of shapes is deliberate, such that different viewpoints of the jar are advantageous for estimating the number of objects based on the specific characteristics of the object's shape. For example, as shown in Figure 3.2, the top view (90°) allows for counting the cross-sections of cylinders but gives no information about

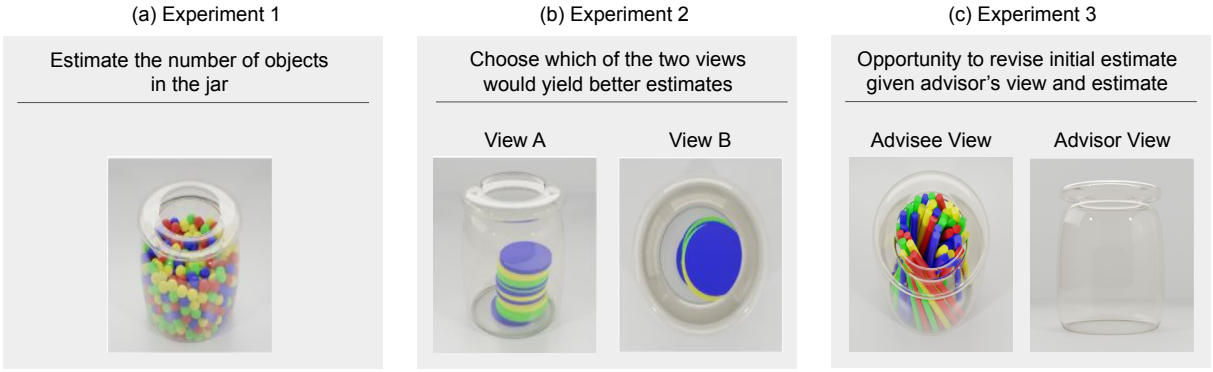


Figure 3.1: Illustration of the behavioral tasks in Experiments 1-3 to investigate advice-taking with information asymmetry between advisee and advisor.

the depth of spheres or disks. Conversely, the side view (0°) is favorable for counting disks but is not suitable for spheres or cylinders.

Figure 3.1 illustrates the behavioral tasks across the three experiments. In Experiment 1, we assess how object shape and jar viewing angle affect estimation accuracy. The estimates are made independently without any advisor. We use the estimates from Experiment 1 as advisor estimates in Experiment 3. In Experiment 2, we assess individuals' ability to identify the most favorable jar view for estimation to establish that individuals are capable of understanding the circumstances where another person would have better quality of information. Experiment 3 examines how advisees integrate advice given the a human or an AI advisor's view of the jar.

3.4 Experiment 1: Eliciting independent estimates across different viewpoints and types of objects

In Experiment 1, we look at how factors such as the shape of the object and viewing angle of a jar, impact people's ability to estimate the number of objects in the jar without any assistance from an advisor. Participants from this experiment serve as advisors to participants

in a later experiment.

3.4.1 Method

Stimuli: Generating Synthetic Images of Object-filled Jars

The images of jars used in the experiments were created using the Python interface of the Blender software (Blender, 2022). Blender is a free and open-source 3D computer graphics software toolset often used for creating animated films, visual effects, and motion graphics. To capture different viewpoints of jars corresponding to the different 'views' accessible to the advisor and the advisee, we designed a standard jar using Blender and placed it in a virtual studio setup available on the platform. Next, we added five cameras to the scene: one positioned horizontally in front of the jar ($0°$), another looking down from the top ($90°$), and three placed in between ($22°, 45°, 66°$). We generated three distinct objects—spheres, cylinders, and disks—that were placed inside the jar. These objects were purposefully selected to create perspectives that are differentially favorable towards counting each type of object. For instance, the side view is optimal for counting disks, whereas the top view works better for counting cylinders. The number of objects in each jar was randomly determined, falling within specified ranges of 50 and 10000 for spheres, 5 and 200 for cylinders, and 5 and 50 for disks. Finally, to virtually place the objects in the jar, we leveraged Blender's physics engine to simulate the dropping of objects into the jar. Five cameras captured snapshots of the jar once the objects were settled in their final positions in the jar. The Python code to create the stimuli is available on https://osf.io/zxftv/.

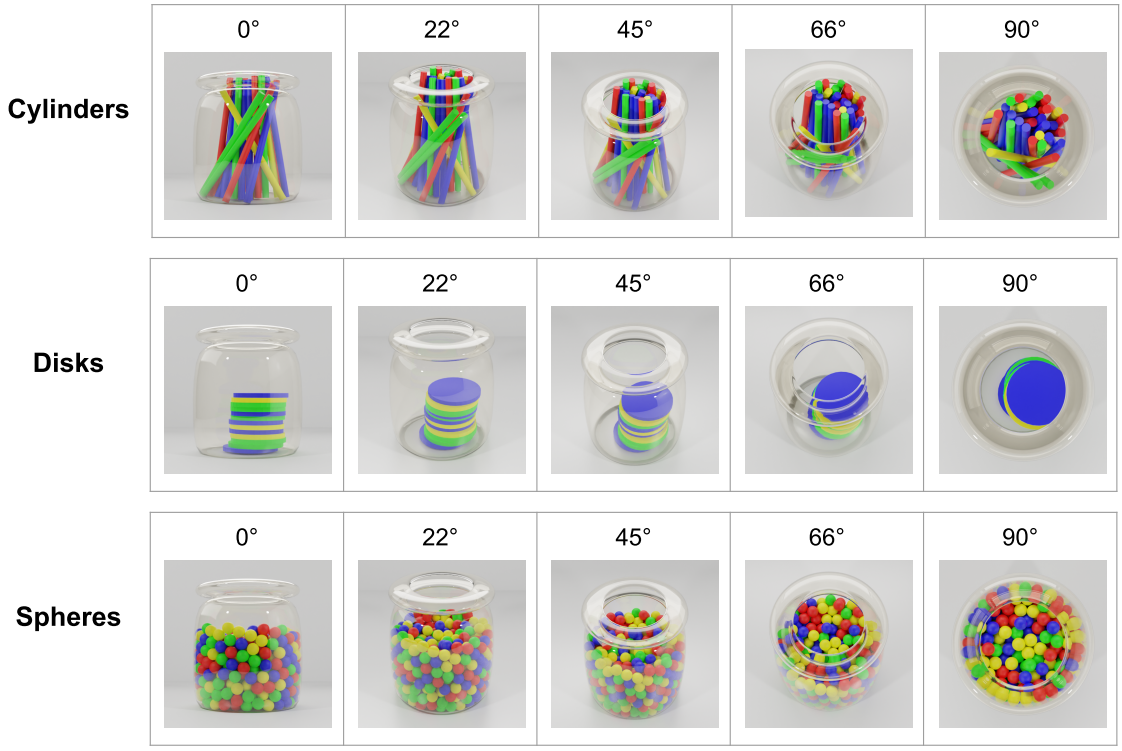


Figure 3.2: Illustration of the stimuli across viewpoints (columns) and types of objects (rows). For each type of object, the images show the jar with the same number of objects in the same configuration.

Participants

One hundred participants were recruited from Prolific. Using Prolific prescreening requirements, we restricted the participant pool to individuals living in the United States who had a minimum approval rate of 90%. Participants from Prolific received $3.00 for their participation and $1.50 for good performance. Participants were expected to take 15 minutes to complete the study. We pre-registered this study at https://osf.io/zjx86. The experimental protocol was approved by the University of California, Irvine Institutional Review Board.

Materials

The stimuli consisted of 150 images: 3 types of objects (spheres, cylinders, and disks) x 5 viewing angles x 10 jars with different numbers of objects. Each jar may be observed from five viewing angles, which include a top view (90°), side view (0°), and intermediate views ($22°, 45°, 66°$).

Procedure

Each subject was assigned the task of estimating the number of objects in 30 jar images. For each trial, participants were instructed to estimate the number of objects contained in the jar in the image displayed. Participants provided their estimates in a text field, which accepted only numeric input. Figure 3.1 (a) shows a schematic of the experiment. The assignment of jar images to participants was counterbalanced, such that there were approximately 20 judgments per unique image.

Analysis

To evaluate participants' estimation error for different images, we computed the Mean Absolute Error (MAE) for each viewing angle and shape combination. If X is the true number of objects in a jar, and X_i is person i's estimate of the number of objects in that jar, the MAE for that trial is $|X - X_i|$. Across the three experiments, we used a logarithmic transformation to standardize participants' estimates.

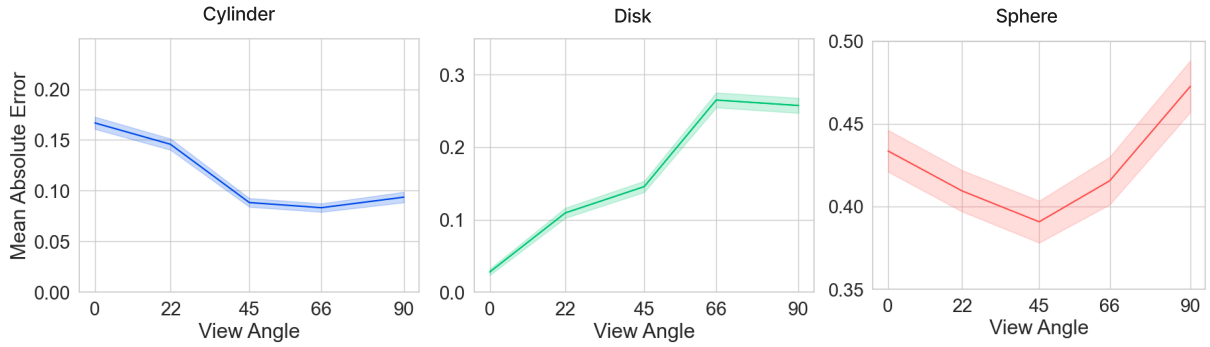


Figure 3.3: Participants' mean absolute errors for cylinders, disks, and spheres across various viewing angles. The colored bands represent the standard error of the mean.

3.4.2 Results

Figure 3.3 shows participants' mean estimation error for each shape and viewing angle. Notably, for cylinders, estimation error decreases as the viewing angle of the jar increases. The opposite pattern is observed for disks, where estimation error increases with the viewing angle. For spheres, the results reveal a non-monotonic and U-shaped trend with the intermediate angles having a lower estimation error than 0 and 90 degrees. In comparison to disks and cylinders, estimation error and variance for spheres are considerably higher. This higher error in estimation can be attributed to the larger number of spheres in the stimuli (Chesney et al., 2015).

Our findings indicate that participants exhibit reasonably good performance in the estimation task, with higher errors for less favorable viewpoints and lower errors for more favorable viewpoints.

3.5 Experiment 2: Identifying the best view for estimating the number of objects

One potential reason for egocentric discounting in advice-taking situations is the challenge of considering another person's perspective (Yaniv & Choshen-Hillel, 2012). Experiment 2 tests whether people can recognize which view of a jar will yield the most accurate estimate and, as a result, understand the differences in information quality when an advisor has a different perspective than the advisee. Experiment 2 results will be used in Experiment 3 to distinguish between favorable and unfavorable advisor views.

3.5.1 Methods

Participants

Two hundred participants were recruited from Prolific. The prescreening requirements were the same as in Experiment 1. Participants from Prolific received $2.00 for their participation and $1.00 for good performance (i.e., if their performance is within the top 30% across all participants). We pre-registered this study at https://osf.io/bzvuf.

Materials

The stimuli consisted of the same jar images as Experiment 1.

Procedure

Each participant completed a total of 30 trials. On each trial, participants were tasked with predicting which of two views of the same jar would lead to the most accurate estimate.

For each participant, the order of the 30 unique jar images was randomized. The first view of the jar, referred to as 'View A', was chosen out of the 5 views. The second view, or 'View B', was chosen out of the 4 remaining views. The combination of views across A and B was counterbalanced such that there were approximately 10 judgments per unique view combination. Figure 3.1 (b) shows an illustration of the experiment.

3.5.2 Results

To quantify the relative preference of View B over View A, we calculate the preference probability for View B. Suppose the number of people who chose View A is X, and the number of people who chose View B is Y.The preference probability for View B is then evaluated as $\frac{Y}{X+Y}$. This metric helps us understand how likely participants are to prefer View B over View A.

3.5.3 Results

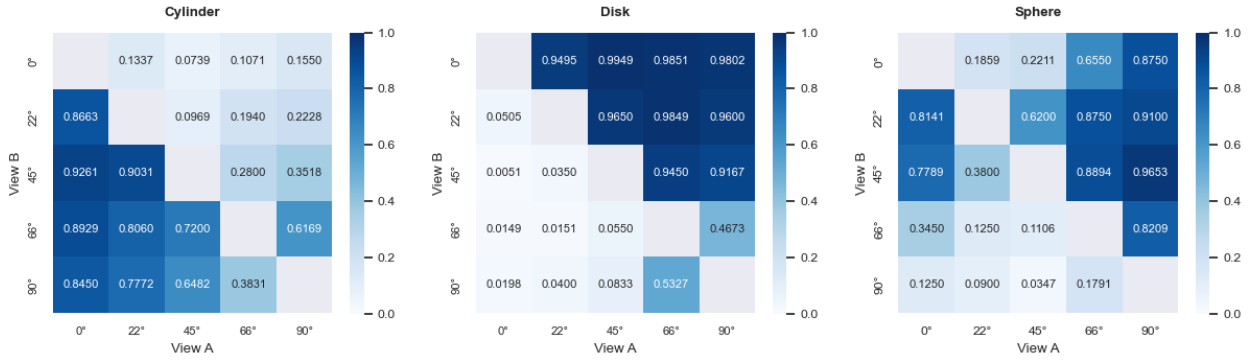


Figure 3.4: Preference probability for View B over View A for each shape and viewing angle combination.

Figure 3.4 illustrates participants' view preferences. Darker shades of blue indicate a stronger preference for View B over View A. Notably, for cylinders, darker cells are primarily clustered

below the diagonal. In this region, as the angle of View A increases, the cells in each row transition from a darker to a lighter shade. This result shows that participants not only favor views with larger angles but also exhibit a stronger preference when the difference between the two viewing angles is larger.

Conversely, for disks, the dark blue cells are concentrated above the diagonal. The shade of this area is darker and more homogeneous compared to the lower triangular area of the cylinder. This shows that participants have a strong preference for lower angles, and the difference between the angles of View B and View A does not strongly impact the magnitude of this preference.

Finally, for the spheres, cells tend to be darker when View B is at a more intermediate angle compared to View A. In addition, the top view (90°) is less preferred over any other viewpoint, presumably because participants intuit that the top view leads to ambiguities in the overall count.

Overall, the results show that participants favor viewing angles that lead to lower estimation errors (as established in Experiment 1) and have an intuitive understanding that some viewpoints lead to lower quality information.

3.6 Experiment 3: Integrating advice from a human or an AI advisor with a different perspective

Experiment 3 investigates how participants evaluate, process, and use the inputs provided by an advisor when they are presented with information regarding the advisor's viewpoint. In one condition, participants are presented with a human advisor's estimates (from Experiment 1). They are tasked with using this advice to formulate their own estimates, aiming for

maximum accuracy. The experiment manipulates the perspective of the jar available to the advisee and advisor. In another condition, participants receive estimates from an AI advisor, along with information about the AI's viewpoint of the jar. We use a Wizard-of-Oz setup wherein the advice, although originating from a human, is portrayed as coming from an AI agent. Participants again have the option to use the AI advisor's estimate to produce an estimate that is as accurate as possible.

3.6.1 Methods

Participants

Two hundred participants were advised by another human and seventy five participants were advised by an AI. Participants were recruited from Prolific. The prescreening requirements were the same as in Experiment 1. Participants received $3.00 for their participation and $1.50 for good performance (i.e., if their performance was within the top 30% across all participants). We pre-registered this study at https://osf.io/srhnx.

Materials

The stimuli consisted of the same jar images as Experiment 1.

Procedure

On each trial, participants first submit an independent estimate of the number of objects in the jar as shown. Next, participants are shown the advisor's view of the jar and the advisor's estimate based on their view as shown in Figure 3.1(c). Participants are required to provide a final estimate, with the option to integrate information from the advisor's estimate. Each

participant was yoked with a participant (an advisor) from Experiment 1. The participants in Experiment 3 viewed the same jar as the advisor, but their view of the jar could be different or the same. The views of the participants were counterbalanced so that there were approximately 8 estimates for each jar-view image pair between the participant and the advisor. Each advisor from Experiment 1 was yoked with a total of 2 participants in Experiment 3.

Analysis

Weight of Advice Weight of Advice (WoA) is measured by how much the advisee's estimate shifts towards the advice given, proportionally to the difference between their original belief and the advice (Bailey et al., 2022; Himmelstein, 2022; Hogarth & Einhorn, 1992; Soll & Larrick, 2009). WOA typically has a trimodal distribution (Soll & Larrick, 2009), with spikes at 0 (advice declined) and 1 (advice adopted). WoA (w) can be understood through a simple mathematical relationship involving the initial estimate of the advised (c), the estimate provided by the advisor (p), and the revised estimate of the advisee after receiving the advice (r). The revised estimate of the advisee is a weighted average of the advisee's initial estimate and the advisor's estimate:

$$\begin{aligned} r &= wp + (1 - w)c \\ w &= \frac{r - c}{p - c} \end{aligned} \tag{3.1}$$

WoA measures how much the advisee's estimate shifts towards the advisor's estimate relative to their initial estimate. If the advisee ignores the advisor's advice, WoA is 0, and if the advisee fully adopts the advisee's estimate, WoA is 1. WoA thus provides a precise way to assess the influence of advice on an individual's decision-making process. We utilize WoA as a metric to infer participants' reliance on the advisor's estimate.

Bayesian Regression Models Bayesian linear regression models were used to compare different models of how individuals determine the weight of advice based on their own and the advisor's viewpoint. All models encoded the independent variables as categorical variables. We used the BayesFactor package (Morey et al., 2018) in R to implement the models and to calculate the Bayes factor, a metric for assessing the strength of evidence in Bayesian Statistics. Bayes Factors allow us to quantify the evidence supporting each model relative to others.

3.6.2 Results

Participants exhibited mean absolute errors of .13, .16, and .38 in independently estimating cylinders, disks, and spheres in the jar when advised by a human. In the AI advisor condition, participants had mean absolute errors of .12, .15, and .38 when estimating the number of cylinders, disks, and spheres respectively. These results are similar to the performance of participants in Experiment 1, where the mean absolute errors were .12, .16, and .43 for the same shapes. Notably, when participants had access to an advisor's estimate, in the human advisor condition, their errors reduced to .10, .11, and .36 and in the AI advisor condition, their errors reduced to .09, .10, and .36 for cylinders, disks, and spheres. The average weight given to the human advisor's advice was .38, .39, and .39, while the average weight given to the AI advisor's advice was .49, .46, and .49 for each of these shapes.

	Human Advisor		AI Advisor	
Shape	Advisee View	Advisor View	Advisee View	Advisor View
Disk	1.7×10^{36}	3.5×10^{21}	7.4×10^{14}	6.1×10^{18}
Sphere	6.2×10^{2}	7.4×10^{3}	6.2	1.5×10^{3}
Cylinder	8.2	4×10^{4}	2×10^{1}	2×10^{1}

Table 3.1: Bayesian Linear Regression Analysis on Weight of Advice. The table compares the Relative Favorability model with the two baseline models (Advisee and Advisor View models) across human and AI advisors, and three object types: Disks, Spheres, and Cylinders. The values reflect Bayes Factors of the Relatively Favorability model against the Advisee and Advisor View models.

Relative Favorability of Views Modulates Weight of Advice

The key question in this study is whether participants take into account the advisor's view relative to their own view when weighting the advisor's estimate. We investigate this question by examining three regression models that aim to predict the Weight of Advice (WoA) based on various types of information about the advisee and advisor perspectives.

The *Relative Favorability* model allows for the possibility that the WoA is dependent on the relative favorability of the advisor and advisee views. The model represents relative favorability based on the preference probability derived from Experiment 2. We divided the preference probabilities into three categories: less than .33 (advisee view is preferred over advisor view), between .33 and .66 (no clear preference for advisor or advisee view), and greater than .66 (advisee view is preferred over advisor view). The goal of this model is to determine how people's preferences for the advisor's point of view influenced the weight they assigned to the advisor's estimate.

The second and third models both function as null models. The *Advisee View* model assumes that only the advisee view predicts the WoA and ignores the advisor view (e.g., the advisee may rely more on the advisor if their view is less favorable regardless of whether the advisor has a favorable view). Similarly, in the *Advisor View* model, the only determinant of WoA is the advisor's point of view and ignores the advisee's point of view.

Each of the three regression models was applied separately to each object shape. The results of this analysis are presented in Table 3.1. The Bayes Factors represent the relative evidence of the Relative Favorability model over the two null models (Advisor and Advisee View). For both the Disks and Spheres, the results provide substantial evidence (BF $>$ 100) for the Relative Favorability model, suggesting participants weigh advice based on the Relative Favorability of different views rather than solely their own or the advisor's view. For Cylinders, the Bayes factor of 8.2 for the comparison between Relative Favorability and Advisee View

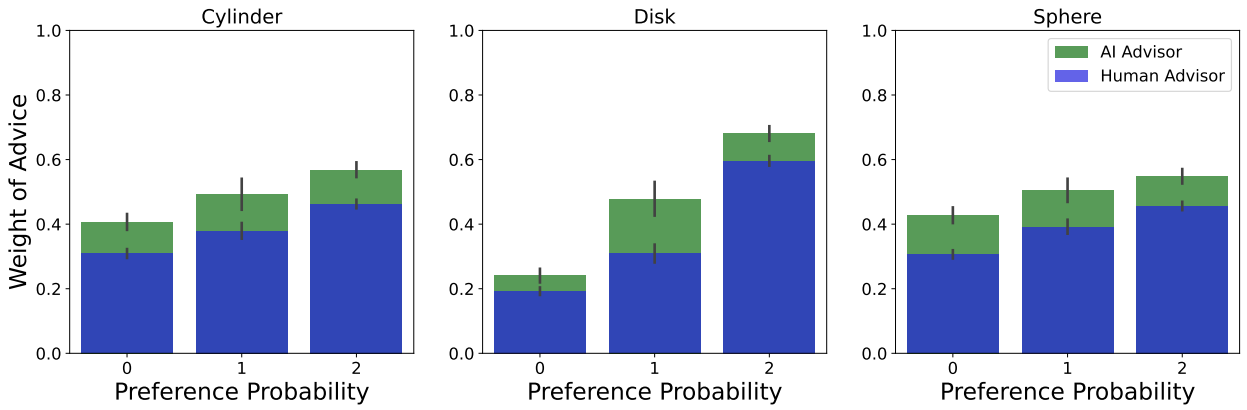


Figure 3.5: People's observed weight of advice across different preference probabilities for both human and AI advisors. The figure shows that people's WOA increased as the advisor's view became more favorable compared to their own. Error bars show standard error of the mean.

suggests more modest evidence in favor of Relative Favorability. In contrast, The Bayes factor of 4×10^4 for Relative Favorability' against Advisor View indicates very strong evidence for Relative Favorability.

In summary, our findings reveal that participants tend to base their advice-taking decision on the relative favorability of views instead of solely relying on their view or the advisor's view of the jar.

Difference between Observed and Optimal Weight of Advice

In this section, we investigate to what extent participants are optimal in their weighting of advice. To excel in the advice-taking task, an ideal participant would need to adjust their weight of advice to minimize their estimation error. The optimal WoA would also have to take into account the combination of advisor and advisee views.

To formalize the analysis of optimal WoA, we introduce some notation. For each trial i in a collection of n advisor-advisee view combination trials, let t_i be the true number of objects in the jar, p_i be the advisor's estimate, c_i be the advisee's initial estimate, and w be the weight

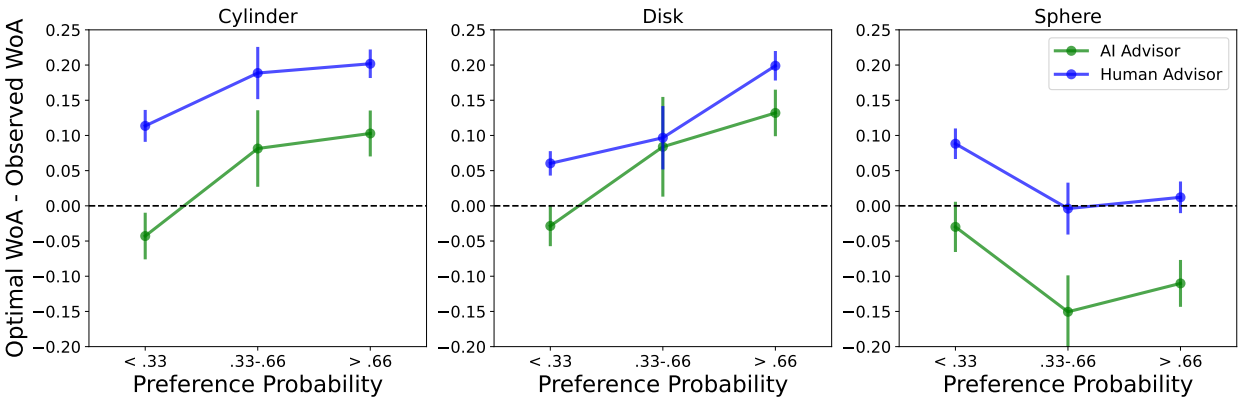


Figure 3.6: Deviation of people's observed weight of advice from the optimal weight of advice across different preference probabilities for both human and AI advisors. The figure shows two trends: 1) Increasing observed WOA with the advisor's view becoming more favorable, and 2) consistent underestimation of the advisor's advice (egocentric discounting), especially for the human advisor, leading to suboptimal decision-making.

of advice. The participant's objective is to minimize:

$$\min \left(t_i - (w_i p_i + (1 - w_i) c_i)\right)^2$$

which gives us a closed-form expression for the optimal weight of advice for trial i, w_i^{opt} (for a detailed derivation, see Appendix A.9):

$$w_i^{opt} = \frac{t_i - c_i}{p_i - c_i} \tag{3.2}$$

.

Figure 3.6 illustrates the comparison between participants' observed and optimal weight of advice for different preference probabilities for both human and AI advisors. Our findings reveal two distinct, consistent patterns across all shapes.

Relative Favorability of the Views Matters. There is a clear correlation between the relative favorability of the advisor's view and the WOA. As the advisor's view becomes more favorable relative to the advisee's view of the jar, observed WOA increases. This is evident in the upward trend of WOA values across preference bins.

Prevalence of Egocentric Discounting. A notable discrepancy can be seen across all preference probabilities: the observed WOA consistently falls short of the optimal WoA. This indicates that individuals tend to assign less weight to the advisor's estimate than would be considered optimal. This reflects a bias where participants are inclined to favor their estimates over those of the advisor, even though it negatively impacts their performance. However, note that this discounting is lower for the AI advisor compared to the human advisor.

3.7 Discussion

Collaboration often involves seeking and incorporating advice from others. However, research on advice-taking consistently reveals that people tend to under-utilize the advice they receive (Bailey et al., 2022; Bonaccio & Dalal, 2006; Kämmer et al., 2023). Traditional experimental studies in this domain have largely focused on scenarios where the advisee and advisor engage in repeated interactions (Schultze et al., 2017) and have access to identical information (Yaniv & Choshen-Hillel, 2012). In contrast, our research takes a novel approach by investigating how individuals process and integrate advice when they are aware that their advisor has access to distinct information. The central idea of our study is that people are able to evaluate the relevance and utility of the information available to the advisor for a specific task, enabling them to make informed adjustments in their reliance on advice.

In this work, we conducted three experiments to examine participants' performance in a simple, yet challenging visual counting task. This task, inspired by everyday experiences,

involves estimating the number of objects in a jar from various perspectives. From Experiment 1, we learned that participants were better at the estimation task when shown some views of the jar compared to others. This is intuitive – some perspectives of the jar obstruct critical information. For example, a top view of the jar with spheres obscured the depth of the objects, impacting the accuracy of people's estimates. Experiment 2 confirmed that people's intuitions are correct - people preferred views of the jar which led to lower estimation errors in Experiment 1. Finally, in Experiment 3, we presented either a human or an AI advisor's estimate and view of the jar as additional information that participants could use to improve their estimates. Our analysis revealed that participants adjusted their initial estimates to incorporate the advisor's estimate, with the degree of adjustment or weight of advice increasing when the advisor had a more favorable view of the jar. This suggests participants' nuanced understanding of how different perspectives can influence the accuracy of estimates. Peoples' shift toward the advisor's estimate was higher for the AI advisor compared to the human advisor. This suggests that people expected AI advisors to perform better on this task compared to humans. This observation is consistent with our earlier discovery in Chapter 2, which showed an overestimation of AI capabilities by individuals in the absence of feedback on their assessments.

Our findings also align with previous advice-taking research, revealing a persistent trend of 'egocentric discounting'. Despite recognizing the value of the advisor's differing perspective, individuals exhibit a bias toward their own information. However, this discounting of advice is less pronounced for AI advisors. Our results highlight an opportunity for further optimizing advice-taking behavior. (Yaniv & Choshen-Hillel, 2012) show that prompting participants to predict another person's perspective reduces egocentric discounting. They propose using perspective-taking as a corrective measure for egocentric discounting. (Kämmer et al., 2023) also explored factors contributing to egocentric discounting, highlighting that decision-makers know the reasons for which they hold their own opinions, but they may lack insight into their advisor's opinions. This information asymmetry can contribute to egocentric discounting.

Our study reinforces these insights, showing that providing information on why advisors hold their opinions can mitigate egocentric discounting.

Understanding the dynamics of advice-taking and egocentric discounting can significantly improve decision-making processes, ultimately enhancing the quality of collective judgments and facilitating more effective collaboration. While our work contributes to a deeper understanding of these dynamics, several aspects remain unexplored. One key limitation of our study is the absence of feedback from participants during the experiment. In real-life scenarios, feedback plays a crucial role in refining one's understanding of both the advisor and the task, allowing for adjustments in mental models based on actual outcomes. Future research should delve into whether and how feedback — either after each trial or in aggregate —affects individuals' strategies in weighting advice. Additionally, our study does not examine the interaction between the overall proficiency of the advisor at the task and their access to differential information. Therefore, one important empirical extension, which we leave for future research, is to look at how people account for advisor reputation when incorporating advice, especially in scenarios with varying levels of information availability.

Chapter 4

Cognitive Modeling of Human-AI Interaction: From Assisted Image Classification to Autonomous Driving

4.1 Abstract

Cognitive modeling is a powerful tool for understanding the hidden cognitive states of humans, enabling predictions about their future actions, beliefs, and knowledge. While much of the past research in cognitive science has concentrated on decision-making models for individuals working alone or with others, recent efforts have shifted towards developing cognitive models to understand human decision-making processes in the context of human-AI interaction (Kumar et al., 2021; Tejeda et al., 2022). This chapter discusses two distinct applications of cognitive modeling in human-AI collaboration:

Latent Reliance on AI in Classification Tasks: We introduce a cognitive model designed to infer humans' latent strategies of relying on AI assistance in an image classification task.

This model circumvents the need for direct queries about their decision-making process. The first employs a concurrent paradigm, presenting AI assistance alongside the task, while the second uses a sequential approach, where participants first make an independent decision before receiving AI input. The model accurately reflects the reliance strategies observed in these experiments, offering a structured method to understand and predict human-AI reliance without direct interrogation. This advancement could broaden the research scope in human-AI cooperation significantly.

Perception of Risk in Autonomous Vehicles (AVs): Despite AVs being designed for safer and more efficient traffic management than human-operated vehicles, it's crucial to pinpoint which AV decisions are deemed unsafe or risky by drivers. Our research explores this through a driving simulator experiment, involving participants and two types of AVs — cars and sidewalk mobility devices, each mimicking the participant's driving style. We develop a cognitive model to assess drivers' perceived risk in various scenarios, enabling us to measure and compare perceived risks across different situations and mobility types. The findings reveal that drivers sense a higher risk in scenarios where the AV adapts to their preferred driving style. Moreover, the perceived risk varies significantly between the two types of mobility. This model's ability to quantify perceived risk and its variation offers valuable insights for designing AVs that are more attuned to human perceptions and preferences.

4.2 Assisted Image Classification: Modeling Latent Reliance Decisions

To investigate AI-assisted decision-making, researchers have designed a variety of workflows. Some workflows require the human to provide an independent decision first, then display

the AI's advice which the human can then use to update their final decision (Chong et al., 2022; Poursabzi-Sangdeh et al., 2021; Yin et al., 2019). Other workflows present AI advice alongside the prediction problem and the human can decide to follow the advice or ignore it (Rajpurkar et al., 2020; Sayres et al., 2019). Finally, a few studies force individuals to spend time thinking about the decision problem by artificially delaying the presentation of AI advice (Buçinca et al., 2021; Park et al., 2019) or making AI advice available only when it is requested (Kumar et al., 2021; Liang et al., 2022). We focus on two of the aforementioned workflows of AI-assisted decision-making and refer to them as paradigms; a detailed illustration can be found in Figure 4.1. We term the first as a *sequential* paradigm, where AI advice is displayed only after the human provides an independent judgment and the human can choose to revise their initial judgment. We term the second as a *concurrent* paradigm where AI advice is displayed concurrently with the prediction problem.

The sequential paradigm provides direct insights into the human's reliance on the AI based on two human judgments: the initial independent judgment and a final judgment after receiving the AI advice. This paradigm makes it easier for experimenters to disentangle the influence of AI advice on the human's decision. However, in many real-world applications, the human user does not independently make a decision before AI assistance is provided since providing the AI's recommendation immediately simplifies the workflow and can save time. The concurrent paradigm offers an alternative setting to study AI-assisted decision-making.

A major drawback of the concurrent paradigm is the fundamental ambiguity in data interpretation — it is unclear as to how one can assess the usefulness of the AI decision aid to the human user. Since there is no initial human judgment available before AI advice is offered, there is no direct empirical observation about any changes the human is making in their decision-making. Any observed agreement between the human and the AI, in the concurrent paradigm, could arise because the human changed their judgment and took the AI's advice or the human already arrived at the same judgment independent of the AI. How, then, do we

assess the impact of AI assistance on the human's decision?

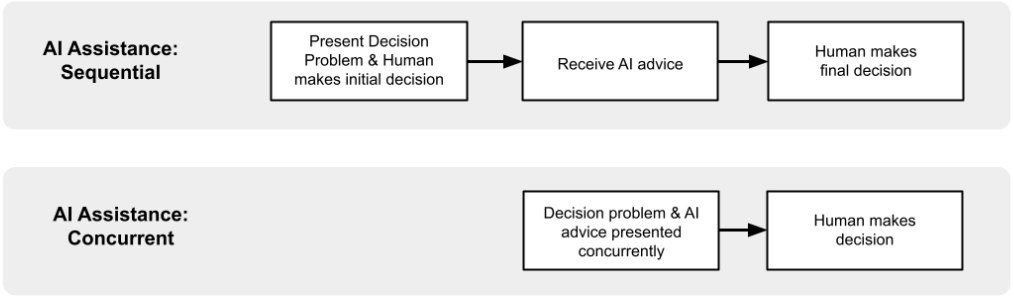


Figure 4.1: Illustration of the sequential and concurrent paradigms for AI-assisted decision-making.

We develop a computational cognitive model for AI-assisted decision-making in the concurrent paradigm. The cognitive model provides a principled way to infer the latent reliance of a human on the AI assistant even though there are no direct observations of switching behaviors when a person is presented with the AI advice. We empirically validate the computational model by collecting empirical data from a behavioral study using both the sequential and concurrent paradigms. The data from the sequential paradigm offers a comparison to the concurrent paradigm and provides a test to assess the merit of the computational framework. We demonstrate that the model's predictions of reliance behavior in the concurrent paradigm are qualitatively similar to the reliance behavior observed in the sequential paradigm. In addition, we demonstrate that the model can generalize to held-out trials in the concurrent paradigm.

4.2.1 Cognitive Model: Inferring Latent Reliance

Based on the paper 'AI-assisted decision-making: A cognitive modeling approach to infer latent reliance strategies' (Tejeda et al., 2022). This work was led by H. Tejeda. M.S., H.T., and P.S. designed research; M.S. and H.T. performed research; M.S., H.T., and A.K. analyzed

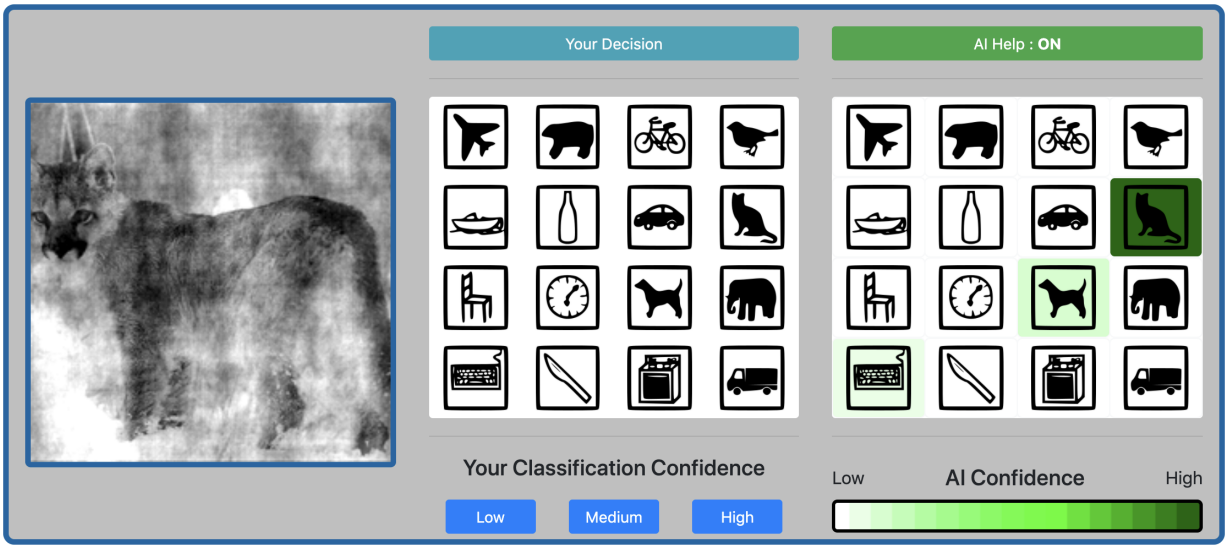


Figure 4.2: Illustration of the behavioral experiment interface in the AI assistance condition.

data; and M.S., A.K., H.T., and P.S. wrote the paper.

The main goal of the computational model is to draw inferences about the latent advice-taking policies. The policy can be influenced by several factors, such as the confidence state of the participant, the confidence scores of the AI as well as the overall accuracy of the AI. We develop a hierarchical Bayesian model to draw inferences about the policies not only at the population level but also at the level of individual participants. In the first part of the model, a Bayesian Item-Response model (Fox, 2010) is applied to the no-assistance condition to infer individual differences in ability as well as differences in difficulty across items (i.e., prediction problems). In the AI-assistance part of the model, these latent person and item parameters are used to explain the observed prediction from a participant which depends on their (unobservable) unaided prediction and the advice-taking policy that determines the likelihood that a participant switches to the AI prediction or stays with their own prediction. Figure 4.3 visualizes the graphical model of the computational model that explains the human predictions with and without AI assistance.

Modeling Human Decisions before Assistance

The computational model for human predictions without AI assistance is based on a Bayesian Item-Response model (Fox, 2010). The Item-Response model makes it convenient to model individual differences in accuracy as well as differences in item difficulty (where items refer to the individual images participants have to classify). To model the human predictions, we use a three-parameter IRT model to capture the probability $\theta_{i,j}$ that a correct response is made by person i on item j:

$$\log\left(\frac{\theta_{i,j}}{1-\theta_{i,j}}\right) = s_j a_i - d_j \tag{4.1}$$

The person parameter a_i is an ability parameter that determines the overall performance of the person across items. The item parameter d_j captures differences in the item difficulty while the item parameter s_j captures discrimination: the tendency of an item to discriminate between high and low ability individuals.

In a typical IRT model, the probability of making a correct response, θ, is used to sample the correctness of an answer. However, for our model, we code the responses from individuals in terms of the predicted label. Let $x_{i,j}$ represent the prediction by person i for item j in the absence of AI assistance. Each prediction involves a choice from a set of L labels, i.e. $x \in \{1, \ldots, L\}$. Let z_j represent the true label for item j. We assume that person i produces the correct label z_j on item j with probability $\theta_{i,j}$ and otherwise chooses uniformly from all other labels, as follows:

$$p(x_{i,j} = m) = \begin{cases} \theta_{i,j} & \text{if } z_j = m \\ (1-\theta_{i,j})/(L-1) & \text{if } z_j \neq m \end{cases} \tag{4.2}$$

Various model extensions could be considered that allow for response biases such that some labels are preferred a priori over others.

Participants not only make a prediction but also express a confidence level, $r_{i,j}$, associated with their prediction. In the experimental paradigm, confidence levels are chosen from a small set of labels, $r_{i,j} \in \{\text{low}, \text{medium}, \text{high}\}$. In the model, we assume that predictions associated with higher accuracy on average lead to higher confidence levels, but that at the item level, the mapping from accuracy to confidence is noisy. To capture the noisy relationship between accuracy and confidence, we use a simple generative model based on an ordered probit model:

$$r_{i,j} \sim \text{OrderedProbit}(\theta_{i,j}, v_i, \sigma_i) \tag{4.3}$$

In this generative model, normally distributed noise with standard deviation σ_i is added to the probability of being correct $\theta_{i,j}$. The resulting value is then compared against a set of intervals defined by parameters v_i, and the interval that contains the value determines the resulting confidence level. Changes in v_i can lead the participant to different uses of the response scale (i.e., using one particular confidence level relatively often) while σ_i determines (inversely) the degree to which accuracy and confidence are related. Note that the parameters σ and v are person-specific to allow for individual differences in the confidence-generating process.

Modeling Human Decisions after Advice

In the model for human decisions in the presence of advice, let $y_{i,j,k}$ represent the observed prediction made by person i on item j after AI advice is considered from AI algorithm k. We include a dependence on the type of AI algorithm as our empirical paradigm will present AI advice from different AI algorithms. In the advice-taking model, we assume that the participant initially makes their own prediction $x_{i,j}$ independent of the AI advice but that their final decision $y_{i,j,k}$ can be influenced by the AI advice. Note that in the no-assistance

condition, the independent predictions $x_{i,j}$ and associated confidence levels $r_{i,j}$ are directly observable, but they are latent in the AI assistance condition of the concurrent paradigm. However, we can use the IRT model in the previous section to simulate the counterfactual situation about the prediction and confidence level that a person would have made if AI advice was not provided. Specifically, we can use the generative model in Equations 4.8-4.3 to generate predictions for $x_{i,j}$ and $r_{i,j}$ on the basis of information about the participant's overall skill (a) as well as information about the difficulty of the particular item (d_j)[1].

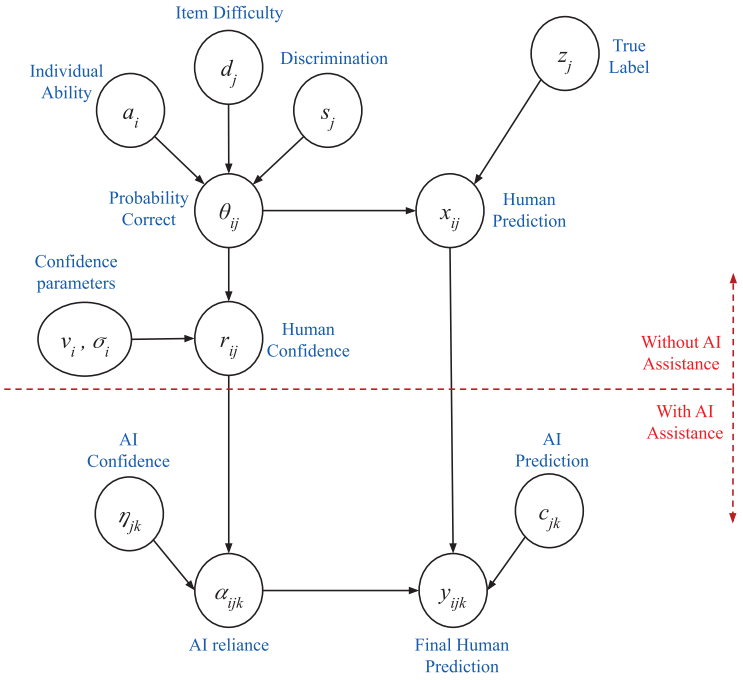


Figure 4.3: Graphical model for the AI-assisted decision-making model. In the condition without assistance, r_{ij} and x_{ij} , and z_j are observed. In the condition where AI assistance is provided, r_{ij} and x_{ij} are latent and y_{ijk}, z_j, c_{jk} and η_{jk} are observed. For visual clarity, plate notation is omitted.

In the advice-taking model, we assume that the participant will stay with their original decision $x_{i,j}$ if it agrees with the AI's recommendation, denoted by $c_{j,k}$. However, when the

[1]Note that in empirical paradigm, each image is presented in both the control condition as well as the AI assistance condition to allow for the estimation of item difficulty parameters for each image.

original decision is not the same as the AI's recommendation, we assume the participant switches to the AI's recommendation with probability $\alpha_{i,j,k}$. Therefore, we can model the probability that the participant chooses label m for their final prediction as follows:

$$p(y_{i,j,k} = m) = \begin{cases} \alpha_{i,j,k} & \text{if } x_{i,j} \neq m \wedge c_{j,k} = m \\ 1 & \text{if } x_{i,j} = m \wedge c_{j,k} = m \\ 0 & \text{if } x_{i,j} \neq m \wedge c_{j,k} \neq m \end{cases} \tag{4.4}$$

The variable $\alpha_{i,j,k}$ determines the tendency of participant i to trust the AI advice from algorithm k related to item j.

Note that in this model, when the participant is provided with AI assistance, the independent prediction $x_{i,j}$ is latent in our experimental paradigm. Instead of explicitly simulating the process of first sampling an independent prediction $x_{i,j}$ and then a final prediction $y_{i,j,k}$, we can simplify the generative process by marginalizing out $x_{i,j}$:

$$p(y_{i,j,k} = m) = \begin{cases} \theta_{i,j} + (1 - \theta_{i,j})\alpha_{i,j,k} & \text{if } z_j = m \wedge c_{j,k} = m \\ \frac{1-\theta_{i,j}}{L-1} + \left(1 - \frac{1-\theta_{i,j}}{L-1}\right)\alpha_{i,j,k} & \text{if } z_j \neq m \wedge c_{j,k} = m \\ \frac{1-\theta_{i,j}}{L-1}(1 - \alpha_{i,j,k}) & \text{if } z_j \neq m \wedge c_{j,k} \neq m \end{cases} \tag{4.5}$$

In this equation, the probability that the participant selects label m is split into three different cases. The first case reflects the probability that the participant makes the correct decision independently (which happened to agree with the AI recommendation) or makes an incorrect decision initially but then adopts the correct AI advice. The second case reflects the probability that the participant initially selects an incorrect decision (which happened to agree with the AI recommendation) or makes another decision different from the AI but then adopts the incorrect AI advice. The third case reflects the probability that the participant makes an incorrect independent decision and decides not to switch to the AI's

recommendation.

Modeling Individual Differences in Advice-Taking

The key latent variable of interest in the model is $\alpha_{i,j,k}$, which determines the willingness of the participant per item to switch to the AI's recommended prediction if it differs from their own prediction. Generally, $\alpha_{i,j,k}$ can depend on many characteristics related to the person, item, and classifier. Here, we will consider functions where α depends on the confidence state of the participant for item j ($r_{i,j}$), the AI confidence score associated with item j ($\eta_{j,k}$), and the type of classifier k:

$$\alpha_{i,j,k} = f(r_{i,j}, \eta_{j,k}, k) \tag{4.6}$$

One way to specify function f is based on a linear model that captures the main effects as well as the interaction between the two putative factors. However, to avoid specifying the exact functional form of f, we will instead simplify the model and treat function f as a lookup table that specifies the α values based on a small number of combinations of participant confidence, AI confidence, and classifier type. Specifically, we create 3 x 4 x 3 lookup table that specifies the α value based on 3 levels of participant confidence ("low", "medium", "high"), 4 levels of AI confidence, and 3 types of classifiers (k). We use a hierarchical Bayesian modeling approach to estimate individual differences in the policy α (See Appendix A.10 for details).

Therefore, we can express the advice-taking policy of person i by $\alpha_{i,k,l}$ where k and l index into the 3 confidence levels of the participant ("low", "medium", "high") and 4 discretized confidence levels of the classifier (see Methods section for details).

To place some constraint on the α parameters, Each participant i has its own advice-taking policy α_i and we assume that these are sampled from a normal distribution on the log-odds

scale:

$$\log\left(\frac{\alpha_{i,j,k}}{1-\alpha_{i,j,k}}\right) \sim \mathcal{N}(\beta_{j,k}, \phi) \tag{4.7}$$

The parameter β represents the advice taking policy at the population level, the tendency across participants to accept AI advice. The standard deviation ϕ captures the spread in individual differences.

4.2.2 Behavioral Experiment

To validate our cognitive model, we investigated human performance with and without AI assistance in two paradigms: the concurrent and sequential paradigms. In this experiment, participants were tasked with classifying noisy images into 1 of 16 categories (see Figure 4.2 for an example of the user interface) with the help of an AI assistant on some trials and without help on other trials. AI assistant performance was a between-subject manipulation whereas image noise was a within-subject manipulation. Participants were instructed to use the AI assistant to the best of their abilities.

There were two experimental manipulations. First, the image noise was varied to produce substantial difference in classification difficulty (Figure 4.4). Second, we varied the overall accuracy of the AI predictions across three conditions: classifier A, classifier B, and classifier C. Classifier A was designed to produce predictions that are, on average, less accurate than human performance. Classifiers B and C were designed to produce predictions that are, on average, as accurate and more accurate than human performance. Each participant was paired with one type of classifier.

The main difference between the two paradigms is that in the concurrent paradigm, participants alternated between blocks of trials where AI assistance was or was not provided. In

the sequential paradigm, there were no alternating blocks. On each trial, the participant first made an independent prediction for an image classification problem and was then allowed to revise their prediction after AI assistance was provided.

Figure 4.4: Illustration of an image under different levels of phase noise. Original images (left) were not used in experiments and are shown only for illustrative purposes.

Stimuli. Following (Geirhos et al., 2019), a subset of 256 images was selected come from the ImageNet Large Scale Visual Recognition Challenge (ILSRVR) 2012 validation dataset (Russakovsky, Deng, Su, Krause, Satheesh, Ma, Huang, Karpathy, Khosla, Bernstein, et al., 2015a). These images were divided equally among 16 classes (chair, oven, knife, bottle, keyboard, clock, boat, bicycle, airplane, truck, car, elephant, bear, dog, cat, and bird). To manipulate the classification difficulty, images were distorted by phase noise at each spatial frequency, where the phase noise is uniformly distributed in the interval $[-\omega, \omega]$ (Geirhos et al., 2019). Eight levels of phase noise, $\omega = \{0, 80, 95, 110, 125, 140, 155, 170\}$, were applied to the images, a different noise level for each unique image, resulting in 2 unique images per category per noise level (see Figure 4.4 for examples of the phase noise manipulation).

AI Assistance. A convolutional neural network (CNN), based on the VGG-19 architecture (Simonyan & Zisserman, 2014), pretrained on the ImageNet dataset was used as AI assistance. Three different levels of classifier performance were created by differentially fine-tuning the VGG-19 architecture to the phase noise used in our experiment.

Procedure. In both the concurrent and sequential paradigms, participants were instructed to classify images as best as possible and to leverage AI assistance, when provided, to optimize performance. Each participant was assigned to a single classifier level (A, B, or

C) at the start of the experiment and each was only presented with AI assistance from that particular classifier; 20 participants were assigned to each classifier level in concurrent paradigm, and 25 participants to each classifier level in the sequential paradigm. When AI assistance was turned on, a grid of the 16 category options was shown with the same layout as the participant response options. Each of the 16 categories was highlighted based on a gradient scale associated with the probability that the AI classifier assigned to the category. The darker the hue of the highlighted category, the more confident the classifier was in that selection. In instances where the classifier was extremely confident in a single category, there would only be one category highlighted with an extremely dark hue. However, in instances where the classifier was not confident in a classification, there would be multiple categories highlighted with low hue levels. Participants were initially given no information about the accuracy of the classifier. Although, at the end of each trial, feedback was provided. In the feedback phase, the correct response option was highlighted in blue. If the participant was incorrect, the incorrect response was highlighted in red.

In the concurrent paradigm, there were 256 trials on which participants were presented a unique image randomly selected from the set of 256 images. The classification trials were separated into 4 four blocks where each block consisted of 48 consecutive trials in which AI assistance was turned on, and 16 consecutive trials without AI assistance.

In the sequential paradigm, there were 192 trials total. On each trial, participants were first tasked with classifying an image on their. After selecting their initial classification decision and submitting their response by selecting a confidence level, participants were provided AI assistance. With AI assistance turned on, participants then made a final classification decision for the image shown and submitted their responses by selecting their confidence level. Once a final classification was made, participants were provided feedback.

4.2.3 Results

Across all classifier conditions, human performance improves with AI assistance, especially at intermediate levels of noise. The accuracy is similar across the concurrent and sequential paradigms. The average human accuracy with AI assistance for classifiers A, B, and C is 57%, 62% and 68% respectively in the concurrent paradigm and 56%, 61%, and 65% respectively in the sequential paradigm. A Bayesian independent samples t-test showed no evidence for a difference in performance for any the classifiers (i.e., all Bayes Factors < 1)[2]. That these results are consistent and very similar in both the concurrent and sequential experiments suggests that the experimental advice-taking paradigm does not produce important differences in how humans rely on and integrate AI assistance.

4.2.4 Model-based Analysis

We used a Markov Chain Monte Carlo (MCMC) procedure to infer model parameters for the graphical model as illustrated in Figure 4.3 (See Appendix A.10 for details). Generally, the model is able to capture all the qualitative trends in the concurrent paradigm. We focus our analysis on β, the advice-taking policy at the population level. We illustrate the inferred policies and compare the results against the empirically observed strategies from the sequential advice-taking paradigm.

Inferred Advice-Taking Policies

Figure 4.5, top row, shows the inferred advice-taking policy β as a function of classifier confidence, participant confidence and classifier. These policies represent the behavior of an average participant at the population level of the model. Overall, the probability of taking AI

[2]Bayes factors were computed using JASP (JASP Team, 2022) with the default priors that came with the software.

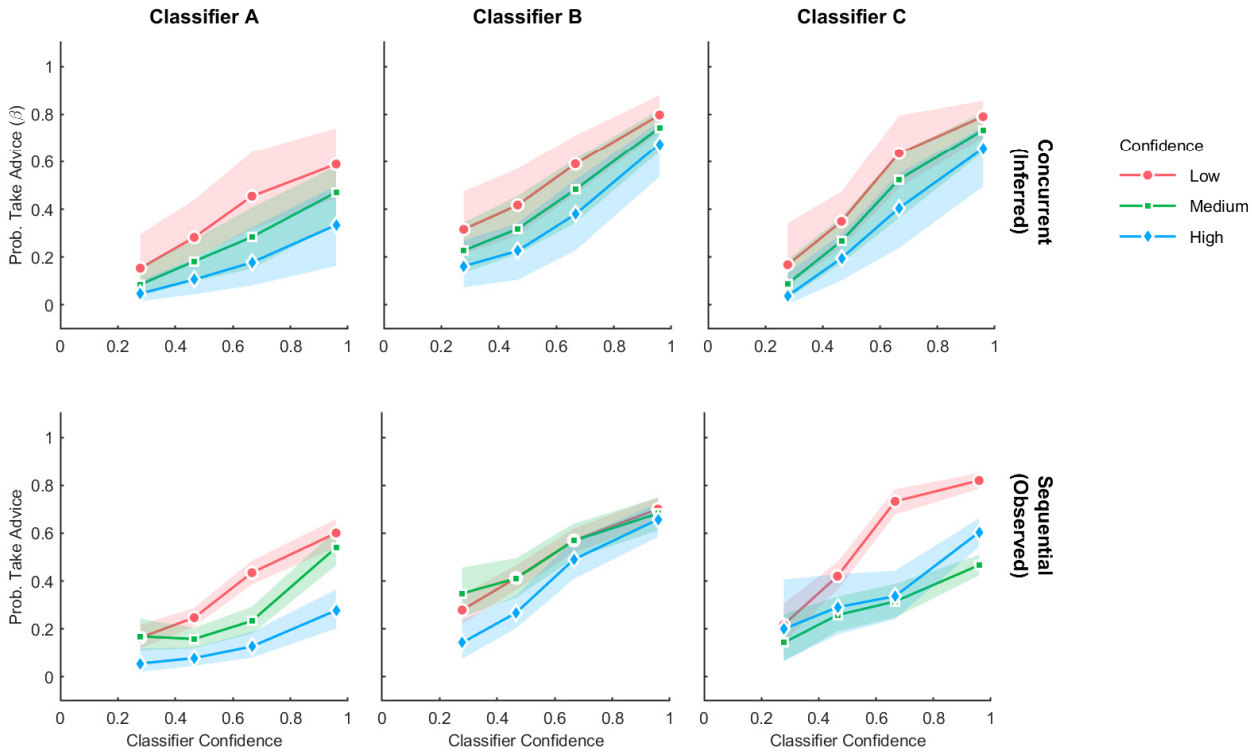


Figure 4.5: Advice-taking policies inferred from the advice-taking behavior in the concurrent paradigm (top row) and observed in the sequential paradigm (bottom row). The policy determines the probability of taking the AI advice as a function of human confidence (colors), classifier confidence (horizontal axis), and type of classifier (columns). The colored areas in the top row show 95% posterior credible intervals. The colored areas in the bottom row reflect the 95% confidence interval of the mean based on a binomial model. The inferred advice taking parameters (β) are converted from log odds to probabilities in this visualization.

advice differs substantially across classifiers. Across the different levels of classifier accuracy, advice is more likely to be accepted from high-accuracy classifiers.

Figure 4.5, bottom row, shows the empirically observed reliance strategies for the sequential paradigm. This analysis focuses on the subset of trials where the initial prediction from the participant differs from the AI prediction (which is not yet shown) and then calculates the proportion of trials where the participant switches to the AI prediction. Importantly, even though some quantitative differences can be observed between the reliance strategies in the two paradigms, the qualitative patterns are the same. Thus, the results from the sequential paradigm provide a key validation of the cognitive model. The latent strategies uncovered by the cognitive model in the concurrent paradigm are very similar to those observed in the sequential paradigm.

4.2.5 Discussion

Appropriate reliance on AI advice is critical to effective collaboration between humans and AI. Most research on AI-assisted decision making has focused on gaining insight into the human's reliance on AI though empirical observations based on trust ratings and comparisons of observed accuracy and final decisions by humans and AI. Instead of using empirical measures to assess reliance, we developed a cognitive modeling approach that treats reliance as a latent construct. The modeling framework provides a principled way to reveal the latent reliance strategy of the individual by using a probabilistic model of the advice-taking behavior in the concurrent paradigm. It can be used to infer the likelihood that a human would have made a correct decision for a particular item independently even when their independent decision is not directly observed. The model can make this inference because it assumes that people, at the same levels of skill, will likely make the same prediction. The model allows us to investigate the difference between agreement with the AI and switching to AI

advice (two metrics often used to assess trust) without explicitly asking the human to respond independently to each problem. In order to apply the model, empirical observations are needed that assess people's independent decisions without the assistance of an AI.

We showed that the AI reliance strategy inferred by the cognitive model based on the concurrent paradigm is qualitatively similar to the AI reliance strategy observed in the sequential paradigm. This demonstrates that a latent modeling approach can be used to investigate AI-assisted decision-making.

4.3 Autonomous Driving: Modeling Perception of Risk

Trust is essential to the functioning of our society. Whether it is at our workplace, at home, or on the roads, everyday we implicitly trust others to more or less do what we expect them to do. We trust other drivers on the road to follow the rules, we trust our coworkers to work on tasks assigned to them, we trust our loved ones to look out for us, and so on. This trust is based on a combination of our past experiences and some assumptions about the world we inhabit. Without trust, the efficacy and efficiency of our day-to-day life would be severely impaired. Trust is an expectation of or confidence placed in or reliance on the other, i.e, trust is always relative to an 'other' - we trust in someone or something. It is important to any form of collaborative work and implies there is risk associated with the task at hand and uncertainty associated with the trustee.

The risk perception of drivers has been extensively researched to enhance safety in transportation. We know that a variety of driver-specific factors such as age (Kington et al., 1994), experience (Ivers et al., 2009), socio-economic status (Machado-León et al., 2016), and personality types (Sjöberg, 2003) affect drivers' risk perception. While understanding the driver-specific factors influence the subjective perceived risk of mobility users, it is also

essential to understand how scenario-specific factors manifest in the drivers' perception of risk. Furthermore, most previous investigations of risk perception have been limited to car mobility (Charlton et al., 2014; Cox et al., 2017; Kington et al., 1994; Machado-León et al., 2016). However, the choices of autonomous mobility types are rapidly evolving. In addition to cars, drivers' may soon have access to a variety of semi-autonomous sidewalk mobility types such as e-scooters (Stocker & Shaheen, 2017). Hence it is important to investigate how drivers' perception of scenarios differs across mobility types. Our work focuses on two questions: 1) which scenarios are generally perceived as high risk by drivers? and 2) how does the type of mobility interact with the risk perception of scenarios? Specifically, we contrast the scenarios perceived as risky when interacting with an autonomous car versus an autonomous sidewalk mobility (such as an e-scooter).

When interacting with an AV, drivers must constantly appraise situational risk and adjust their trust in the AV to make decisions of whether to take over control from the AV or to allow the AV to continue driving. Driver intervention can be interpreted as an indication of a lack of trust, or high perceived risk, or a combination of both. Several studies suggest that drivers tend to trust AVs that drive in a way that resembles their own driving style (Griesche et al., 2016; Natarajan et al., 2022). In contrast, (Lehsing et al., 2019) show that drivers favor defensive AV driving in general. Most previous work attributes a driver's decision to override the AV to driver-specific traits or AV-specific features. We hypothesize that interventions such as braking by the driver may be influenced by scenario-specific risk in addition to the AV's driving style and the driver's trust in the AV. To tease apart the contribution of driver-specific trust and scenario-specific risk, we develop a computational model that infers trust and perceived risk from drivers' braking behavior.

Trust combined with situational attitudes such as perceived risk modulates drivers' intention to rely on automation lee2004trust. Both trust and perceived risk are commonly measured in human-subject experiments via surveys, or by intermittently probing participants to indicate

their perceived risk of a situation he2022modelling, machado2016socio. Instead of using such empirical measures, we directly infer trust and perceived risk from the intervention behavior of humans being driven by an AV. We posit that a driver's tendency to intervene in an AV's driving in a given situation depends on two factors: 1) the driver's trust in the AV's driving capabilities, and 2) the risk associated with the situation. We develop a model that treats driver trust and scenario-specific perceived risk as latent constructs. This modeling framework provides a principled way to understand which scenarios and AV actions are perceived as risky. It also allows us to examine how risk perception differs across two different kinds of mobility, namely car and sidewalk mobility, where the latter is largely unexplored in the literature. We qualitatively contrast scenarios that are perceived as high risk in the two mobility types.

(a) Autonomous car.

(b) Autonomous sidewalk mobility.

Figure 4.6: Driving simulator setup.

4.3.1 Methods

Experiment Setup

We conducted an in-person driving simulator experiment that tasked participants with supervising an AV (Mehrotra et al., 2023). Each participant supervised three AV drives

Car Mobility			Sidewalk Mobility		
Scenario	Aggressive Decision	Defensive Decision	Scenario	Aggressive Decision	Defensive Decision
Yellow light ahead	maintaining speed	stopping	Yellow light ahead	maintaining speed	stopping
Heavy traffic ahead	turning at gap	waiting for safe turn	Crowd ahead	continuing at gap	waiting for safe turn
Pedestrians stopped	continuing	stopping	Pedestrians stopped	continuing	stopping
Left turn lane blocked	crossing lines to pass	waiting	Sidewalk blocked	crossing lines to pass	waiting

Table 4.1: Illustrative examples of *proactive* scenarios and corresponding aggressive and defensive AV decisions for the car and sidewalk mobility.

across two mobility types: car and sidewalk. The experiment was conducted using a high-fidelity driving simulator based on Unreal Engine 4.24 epicgames2019unreal with AirSim shah2018airsim that consisted of a custom city where the automated driving was simulated by replaying a past researcher's drive via the "Wizard of Oz" technique wang2017marionette. The environment was shown to participants using a StarVR headset with a 210-degree FoV. A motion base (MB-200 6-degree of freedom motion base by Cosmate Co., Ltd.) was used to allow for a higher fidelity simulation where participants could feel the typical forces experienced in a vehicle. A car platform was used for all car mobility drives, and a scooter platform was used for all sidewalk mobility drives (see Figure 4.6).

Experiment Design

Drive Types The experiment consists of a tutorial drive, "proactive" drives, and standard drives, all autonomous drives and are described below.

Tutorial drive. The tutorial consisted of a simple drive through an empty urban area with no other cars or pedestrians, which lasted for approximately 3 minutes.

Proactive drives. To manipulate trust without changing automation reliability, a set of events was created in which the automation performed "proactive" maneuvers while maintaining safe driving behavior. The goal of these events was to create situations in which

the AV had multiple options for safe actions to perform. For each event, two actions were designed; one represented an "aggressive" action while the other represented a "defensive" action. Table 4.1 and Figure 4.7 give examples of proactive scenarios and the AV's response to these scenarios for both car and sidewalk mobility. Each proactive drive had 8 events and each drive was approximately 12 minutes long. Since all car events were not directly applicable to sidewalk mobility, equivalent events were created to match the car events as closely as possible.

We hypothesize that scenarios where the AV makes proactive decisions would be perceived as high risk by participants. However, we expect that participants would not intervene in the proactive decisions of the AV as the AV's driving style is matched to the participant's preferred driving style as described in Section 4.3.1.

Standard drives Standard drives involved no proactive events to serve as prerequisite trust-building drives (before proactive drives). They involved the automated vehicle navigating through multiple intersections in an urban area. Similar to proactive drives, standard drives included the presence of other cars and pedestrians throughout the urban area; however, there were no ambiguous scenarios that required advanced decision-making by the automated vehicle. Each drive lasted for approximately 8 minutes.

Intervention by drivers. Participants were instructed to monitor the AV's driving and they could indicate their intent to intervene in the AV's driving by braking or accelerating for both the car and sidewalk mobility. Braking and accelerating intents were captured using the brake and throttle foot-pedals for the car and the brake and throttle hand-levers for the sidewalk mobility, respectively. Participants were informed that their braking or accelerating intentions would not cause any changes in the drive but it would be used as feedback by the AV.

(a) (b)

Figure 4.7: Example proactive car event and sidewalk event. The green arrow shows the path that will be traversed. (a) Car event where the car crosses the double yellow line in order to get into the turn lane when there is stopped traffic ahead. (b)Sidewalk mobility event where the sidewalk mobility drives onto the road outside the crosswalk lines because of stopped pedestrians ahead.

In our current analysis, we divide each drive into traffic scenarios of ~ 10s each, and limit our analysis to *braking* interventions by drivers. We perceive braking as an indication that participants may have realized the limitations of automation or they may not be comfortable with the AV actions. If the driver brakes for the duration of a scenario, then it is considered a braking instance.

Experiment Procedure

Forty-eight participants recruited from a university campus completed the in-person study at the simulator room in Honda Research Institute's USA Inc, San Jose. Participants were matched with a defensive or aggressive AV based on a survey: they were asked to choose between two driving videos based on the video's resemblance to their own driving style. The two videos were designed to recreate aggressive and defensive driving styles. Once the participant was assigned to a defensive or aggressive AV group, the participant was then randomly assigned to monitor an autonomous car or an autonomous sidewalk. For each of the two mobility types, the participant first monitored a tutorial, then saw either 1) a standard drive followed by a proactive drive, or 2) two proactive drives. Participants monitored the two mobility types in two separate sessions conducted 1 hour apart.

This experiment allows for many interesting investigations of drivers' braking behavior in different drive-type combinations. For instance, comparing braking in standard and proactive drives, or the effect of interacting with one mobility type, say car, before switching to the sidewalk mobility. However, here we present preliminary work where we restrict our analyses to drivers' interaction with the AV in the first proactive drive they monitored.

4.3.2 Cognitive Model: Inferring Perceived Risk

An individual taking a risk is considered to be exhibiting behavioral trust mayer1995integrative. Conversely, braking by a driver may be interpreted as a lack of trust in the AV when faced with a risky scenario. We posit that driver's intent to brake during any traffic scenario is based on two key latent variables: 1) the driver's trust in the AV's driving capabilities, and 2) the risk associated with the traffic scenario. While trust is a complex multi-dimensional construct bhattacharya1998formal, here we model it as one-dimensional – all variations in trust can be characterized by changes along a single overall trust scale.

Our framework is based on the item response theory model (IRT, van2013handbook, fox2010bayesian). Specifically, we use a basic Rasch model rasch1993probabilistic. IRT models have been extensively used in the education and cognitive science communities to model performance differences across people and the problems they encounter.

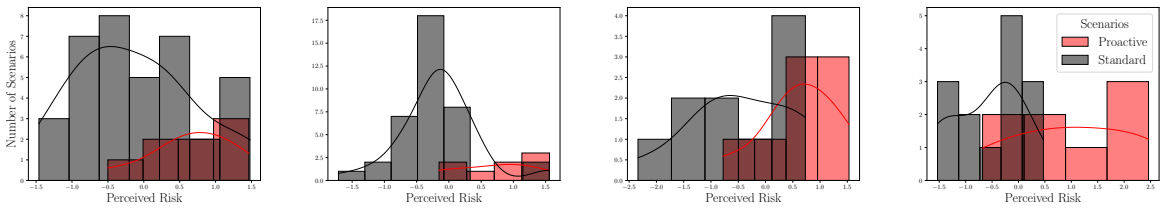
We modify the basic IRT model to account for individual differences in trust and differences in risk associated with traffic scenarios. Let $x_{i,j}$ be a binary indicator of braking by driver i in scenario j. We model $x_{i,j}$ by combining two latent factors, the trust t_i of each driver i and

the risk r_j associated with situation j:

$$\begin{aligned} \theta_{i,j} &= t_i - r_j \\ p_{i,j} &= \frac{1}{1 + \exp(-\theta_{i,j})} \\ x_{i,j} &\sim \text{Bernoulli}(p_{i,j}) \end{aligned} \tag{4.8}$$

Note that t_i represents an aggregate measure of trust for each driver i. The IRT model allows us to rank drivers in order of their exhibited trusting behavior while accounting for the risk r_j associated with each scenario j.

For t_i, we use a left-skewed normal prior to account for the empirical observation that some drivers never expressed an intention to brake. We posit that drivers who never intended to brake have the highest level of trust in the AV. For risk, r_j, we use a standard normal prior to avoid identifiability issues fox2010bayesian. We used Markov Chain Monte Carlo (MCMC) sampling to infer model parameters and obtain samples from the posterior distribution. We chose the Stan computing environment for posterior inference stan.



(a) Aggressive Car (b) Aggressive Sidewalk Mobility (c) Defensive Car (d) Defensive Sidewalk Mobility

Figure 4.8: Perceived risk inferred by the IRT model in proactive drives.

4.3.3 Results & Discussion

Perceived Risk in Proactive vs Standard Scenarios

We use the IRT model to infer which events are perceived as having high risk by participants. In Figure 4.8, we see the distribution of the inferred risk values for situations where the AV took a proactive decision and for situations where the AV did not take a proactive decision. Higher values on the x-axis correspond to higher perceived risk. For all mobility and driving style combinations, we observe a similar pattern: participants perceived proactive events as having higher risk as compared to standard events.

Most prior research suggests that drivers prefer an AV with a driving style similar to their own ma2020investigating or defensive AV driving in general ekman2019exploring. In line with these previous findings, we expected that participants in our experiment would have fewer braking instances in scenarios where the AV makes a proactive decision. Instead, we observe that both defensive and aggressive proactive actions by the AV lead to high perceived risk and consequently braking by the participants. We posit that participants perceive some scenarios as risky independent of the driving style of the AV. Hence, we must account for drivers' perceived risk in addition to driving style preferences when designing human-aware AV systems.

We used a two-sample Kolmogorov-Smirnov (KS) test to examine if the perceived risk of proactive events is different from the perceived risk of standard events. For both mobility types, and for both aggressive and defensive AVs, the distribution of the perceived risk of proactive scenarios is significantly different from the perceived risk of standard scenarios (Fig. 4.8 (a) KS statistic = .632, p-value = .005, (b) KS statistic = .572, p-value = 0.014, (c) KS statistic = .550, p-value = .095, and (d) KS statistic = .679, p-value = .008).

Perceived Risk in Car vs Sidewalk Mobility

A key feature of our experiment is that participants interacted with two different kinds of mobility types: car and sidewalk mobility. In order to better understand which scenarios are perceived as riskier than others in both these mobility types, we tag the scenarios in all drives as interactions with either pedestrians or cars. For example, if the scenario involves a pedestrian crossing the road then it would be tagged a 'pedestrian' scenario. Whereas if the scenario involves the AV waiting at an intersection for other cars to pass then it would be tagged a 'car' scenario.

The output from the IRT-based model allows us to *qualitatively* contrast which scenarios are perceived as having high risk in the two mobility types. Within the scenarios that are inferred as being high-risk by the model, we observe a pattern: when being driven by an autonomous car, participants perceived events involving pedestrians as higher risk when compared to other events, i.e, drivers were highly likely to override the AV and brake when there was a pedestrian around. In contrast, when being driven by the autonomous sidewalk mobility, events that involved cars were perceived as higher risk than other events. This is a key yet intuitive insight from this analysis: which scenarios are perceived as risky depends on the type of mobility.

Future Directions

In our future work, we aim to study how perceived risk changes as drivers repeatedly interact with the same AV and in similar traffic scenarios. To do this, we have developed an online version of our experiment that will give us access to a larger participant pool and will allow us to have drivers interact with an AV in multiple drives. Note that while our model captures the perceived risk of scenarios, it does not explain how the perceived risk of scenarios changes over time. One important extension of the current work is to model change in perceived risk

when participants are exposed to similar scenarios multiple times.

We present the first (to our knowledge) assessment of drivers' risk perception of traffic events across autonomous car and sidewalk mobility. We develop an IRT-based model to infer the perceived risk of traffic scenarios. We use this model to examine the differences in risk associated with scenarios when the AV makes decisions that are in line with the driver's preferred driving style. We observe a significant difference between drivers' perceived risk in scenarios where the AV takes a proactive decision as compared to scenarios where the AV takes a standard decision. This difference is consistent across two different mobility types—car and sidewalk.

Chapter 5

Conclusion & Future Directions

Effective collaboration with another agent requires humans to develop an accurate mental model of the agent's capabilities. This dissertation examined people's mental models of other humans and AI agents through the lens of cognitive science.

In Chapter 2, we developed a theoretical framework to understand the mental computations that people employ when assessing the knowledge of other agents. The empirical results and model predictions revealed that people's evaluation of the other person's performance (a theory of mind computation) is linked to their evaluation of their own performance (a metacognitive computation). We showed that people assess others' abilities relative to their own, and this comparison is informed by their experience in the task. We also applied this model to investigate people's perceptions of an AI's ability in a trivia task. This revealed that people perceived a larger capability gap between themselves and AI when compared to other humans. The perception of difference in capability persisted despite feedback about the AI's actual performance.

Chapter 3 investigated if people's utilization of advice improves when they know the advisor has access to different information compared to them. Through three experiments, we

examined how individuals integrate advice in an estimation task, where the advisee and the advisor have access to different perspectives of the same problem. We found that people can evaluate the relevance and utility of the information available to the advisor for a specific task and make informed adjustments in their adoption of advice. However, despite the difference in the favorability of the information available, people consistently undervalued advice and exhibited egocentric discounting. Understanding these dynamics of advice-taking and egocentric discounting is critical to improving decision-making processes and facilitating more effective collaboration.

Chapter 4 presented two applications of cognitive modeling to uncover people's latent reliance on AI assistance. First, we presented a cognitive model to predict people's decisions to rely on AI assistance in an assisted image classification task. The model allowed us to infer people's reliance strategy on the AI's assistance in the concurrent paradigm. This strategy closely matched people's strategy observed in the sequential paradigm. Our results demonstrate that latent cognitive modeling techniques can be an effective tool to explore AI-assisted decision-making. Second, we conducted in-person driving simulator experiments where people intervened in an autonomous vehicle's driving if they perceived its decisions to be risky. We presented a cognitive model to infer people's perceived risk during this interaction. The model revealed a clear pattern: participants in autonomous vehicles perceived situations with pedestrians as higher risk, often choosing to manually override the vehicle's controls to brake in their presence. Conversely, when interacting with the autonomous sidewalk mobility, scenarios involving cars were deemed higher risk compared to other situations. A nuanced understanding of how perceived risk differs across mobility types and how perceived risk evolves will provide critical insights for the design of human-aware mobility.

This thesis provides insights into human-AI interaction by examining human perceptions, evaluations, and interactions with both humans and AI. The findings lay the groundwork for leveraging cognitive science to guide the development of AI systems that enhance human

capabilities and the design of interactions mindful of human cognitive processes. Future research should aim to deepen our understanding of these mental models and their influence on human interactions with AI. We now outline a few promising directions for future work.

5.1 Future Directions

5.1.1 Improving the Assessment of Mental Models

Understanding people's mental models of AI will require new research in several directions. First, not much is known at the moment about the long-term changes in human beliefs about AI (Glikson & Woolley, 2020). Longitudinal studies will need to be conducted to understand the changes in mental models over time. Do they become more accurate over time? In addition, methods such as cognitive modeling will be useful to make inferences about the latent content of people's mental models, including decision-making strategies and beliefs that cannot be directly assessed with behavioral measures (e.g., (Chong et al., 2022; Tejeda et al., 2022)). Finally, as the human mental model of the interaction with the AI encodes the perceived differences between one's capabilities and the capabilities of the AU, it will be useful to leverage insights from psychology research on metacognition to understand how people estimate their own self-confidence (Koriat & Levy-Sadot, 1999) and their performance relative to others (Moore & Cain, 2007).

5.1.2 Designing Intelligent Interactions

Designing intelligent interactions between humans and AI presents a promising avenue for research. We know that humans are capable of improving their knowledge when they engage deeply with the advice they receive (Gajos & Mamykina, 2022). How can we then design

algorithmic assistance that not only aids decision-making but also helps humans upskill? Moreover, it has been observed that forcing humans to deliberate about their decisions can significantly reduce overreliance on AI advice (Buçinca et al., 2021). However, demanding extensive deliberation may not be appropriate or efficient in every decision-making scenario, especially when there is a dual goal of high accuracy and swift decision-making. When should AI assistance encourage users to engage in extended deliberation? Future work may leverage cognitive modeling to identify optimal moments to prompt users (Callaway et al., 2023) for deeper engagement with AI input, prioritizing deliberation when human error is more likely.

5.1.3 Enabling Adaptive AI-Assistance

Previous empirical investigations have demonstrated that providing more information regarding AI does not always increase performance. Given the limited cognitive resources and time that might be available to process AI recommendations, the AI needs to adjust its output by providing explanations at the right level of detail. Overloading users with information can hinder decision-making processes. (Poursabzi-Sangdeh et al., 2021; Schaffer et al., 2019). Therefore, AI systems must be designed to adapt to the cognitive limitations of the human DM (Cummings, 2017). The questions of what, when, and how much information should be presented to a human DM highlight the need to develop theoretical frameworks that infer the impact of AI aids on human cognition and observed performance. Such frameworks are now starting to emerge in the context of explainable AI (Chen et al., 2022). In addition, theories and computational models drawn from psychology can be leveraged to better understand human cognition when collaborating with an AI (Rastogi et al., 2022). For example, in situations in which decisions must be made quickly or in which varying degrees of mental effort are required to process the AI's output, theories of rational resource allocation (Becker et al., 2022; Lieder & Griffiths, 2020; Lieder et al., 2018) could be used to identify when people might disregard AI predictions if the perceived gains do not warrant the associated

costs in terms of time and mental effort.

5.1.4 Interactions with Generative AI

This thesis focused on AI agents based on discriminative machine learning. However, recent advances in generative AI have markedly changed the capabilities of AI assistance. Tools like Dall-e can create spectacular images from text prompts and large language models (LLMs) like ChatGPT are capable of responding to queries in human-like text. LLMs are being deployed across diverse fields, including public health (Ali et al., 2023), coding (Zambrano et al., 2023), and education (Whalen, Mouza, et al., 2023). For example, LLMs are notorious for generating responses that, while convincing, may be inaccurate or nonsensical (Huang et al., 2023; Jo, 2023). This raises concerns about the reliability of these models. Recent research (Steyvers et al., 2024) has begun to look at how humans understand and interpret responses from these LLMs. However, several questions remain unexplored. Can humans differentiate between factual information and 'LLM hallucinations'? How can LLMs best communicate their internal uncertainty to human users? When should LLMs elaborate on their responses? How does the LLM response's emotional valence impact the user's trust?

Bibliography

Aboody, R., Dunham, Y., Jara-Ettinger, J., et al. (2021). I can tell you know a lot, although i'm not sure what: Modeling broad epistemic inference from minimal action. *Proceedings of the 43rd Annual Conference of the Cognitive Science Society*, *43*.

Abraham, H., Lee, C., Brady, S., Fitzgerald, C., Mehler, B., Reimer, B., & Coughlin, J. F. (2017). Autonomous vehicles and alternatives to driving: Trust, preferences, and effects of age. *Proceedings of the transportation research board 96th annual meeting*, 8–12.

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, *22*(3), 37–51.

Ali, S. R., Dobbs, T. D., Hutchings, H. A., & Whitaker, I. S. (2023). Using chatgpt to write patient clinic letters. *The Lancet Digital Health*, *5*(4), e179–e181.

Allahverdyan, A. E., & Galstyan, A. (2014). Opinion dynamics with confirmation bias. *PloS one*, *9*(7), e99557.

Baer, C., Malik, P., & Odic, D. (2021). Are children's judgments of another's accuracy linked to their metacognitive confidence judgments? *Metacognition and Learning*, 1–32.

Bailey, P. E., Leon, T., Ebner, N. C., Moustafa, A. A., & Weidemann, G. (2022). A meta-analysis of the weight of advice in decision-making. *Current Psychology*, 1–26.

Baker, Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 1–10.

Baker, Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-ai team performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *7*(1), 2–11.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? the effect of ai explanations on complementary team performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Becker, F., Skirzyński, J., van Opheusden, B., & Lieder, F. (2022). Boosting human decision-making with AI-generated decision aids. *arXiv preprint arXiv:2203.02776*.

Berke, M., & Jara-Ettinger, J. (2021). Thinking about thinking through inverse reasoning. *43*.

Bevan, W., Maier, R. A., & Helson, H. (1963). The influence of context upon the estimation of number. *The American Journal of Psychology*, *76*(3), 464–469.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34.

Blender. (2022). Blender (version 3.1.0) [computer software].

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes*, *101*(2), 127–151.

Boyd-Graber, J., & Börschinger, B. (2019). What question answering can learn from trivia nerds. https://doi.org/10.48550/ARXIV.1910.14464

Brennan, S., & Williams, M. (1995). The feeling of another s knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of memory and language*, *34*(3), 383–398.

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–21.

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239.

Callaway, F., Hardy, M., & Griffiths, T. L. (2023). Optimal nudging for cognitively bounded agents: A framework for modeling, predicting, and controlling the effects of choice architectures. *Psychological Review*.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825.

Charlton, S. G., Starkey, N. J., Perrone, J. A., & Isler, R. B. (2014). What's the risk? a comparison of actual and perceived driving risk. *Transportation research part F: traffic psychology and Behaviour*, *25*, 50–64.

Chen, C., Feng, S., Sharma, A., & Tan, C. (2022). Machine explanations and human understanding. *arXiv preprint arXiv:2202.04092*.

Chesney, D., Bjalkebring, P., & Peters, E. (2015). How to estimate how well people estimate: Evaluating measures of individual differences in the approximate number system. *Attention, Perception, & Psychophysics*, *77*, 2781–2802.

Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (2022). Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of ai advice. *Computers in Human Behavior*, *127*, 107018.

Cox, J. A., Beanland, V., & Filtness, A. J. (2017). Risk and safety perception on urban and rural roads: Effects of environmental features, driver age and risk sensitivity. *Traffic injury prevention*, *18*(7), 703–710.

Craik, K. J. W. (1952). *The nature of explanation* (Vol. 445). CUP Archive.

Cummings, M. L. (2017). Automation bias in intelligent time critical decision support systems. In *Decision making in aviation* (pp. 289–294). Routledge.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114.

Dunning, D. (2011). The dunning–kruger effect: On being ignorant of one's own ignorance. In *Advances in experimental social psychology* (pp. 247–296). Elsevier.

Dunning, D., & Helzer, E. G. (2014). Beyond the correlation coefficient in studies of self-assessment accuracy: Commentary on zell & krizan (2014). *Perspectives on Psychological Science*, *9*(2), 126–130.

Fleming, S. M. (2021). *Know thyself: The science of self-awareness*. Basic Books.

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological review*, *124*(1), 91.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in human neuroscience*, *8*, 443.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. Springer Science & Business Media.

Frith, C. D., & Frith, U. (1999). Interacting minds–a biological basis. *Science*, *286*(5445), 1692–1695.

Fussell, S. R., & Krauss, R. M. (1991). Accuracy and bias in estimates of others' knowledge. *European Journal of Social Psychology*, *21*(5), 445–454.

Gajos, K. Z., & Mamykina, L. (2022). Do people engage cognitively with ai? impact of ai assistance on incidental learning. *27th international conference on intelligent user interfaces*, 794–806.

Geirhos, R., Medina Temme, C., Rauber, J., Schütt, H., Bethge, M., & Wichmann, F. (2019). Generalisation in humans and deep neural networks. *Thirty-second Annual Conference on Neural Information Processing Systems (NeurIPS 2018)*, 7549–7561.

Gentner, D., & Stevens, A. (1983). Mental models. hillsdale, nj: Lawrence erlbaumassociates. *Inc. GentnerMental models1983*.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660.

Gneiting, T., & Walz, E.-M. (2021). Receiver operating characteristic (roc) movies, universal roc (uroc) curves, and coefficient of predictive ability (cpa). *Machine Learning*, 1–29. https://doi.org/10.1007/s10994-021-06114-3

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, *19*(1), 121–127.

Gopnik, A., & Astington, J. W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child development*, 26–37.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Mit Press.

Greene, W. H., & Hensher, D. A. (2010). *Modeling ordered choices: A primer*. Cambridge University Press.

Grgić-Hlača, N., Engel, C., & Gummadi, K. P. (2019). Human decision making with machine assistance: An experiment on bailing and jailing. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–25.

Griesche, S., Nicolay, E., Assmann, D., Dotzauer, M., & Käthner, D. (2016). Should my car drive as i do? what kind of driving style do drivers prefer for the design of automated driving functions. *Braunschweiger Symposium*, *10*(11), 185–204.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of educational psychology*, *56*(4), 208.

Hartig, J., & Höhler, J. (2009). Multidimensional irt models for the assessment of competencies [Assessment of Competencies]. *Studies in Educational Evaluation*, *35*(2), 57–63. https://doi.org/https://doi.org/10.1016/j.stueduc.2009.10.002

Himmelstein, M. (2022). Decline, adopt or compromise? a dual hurdle model for advice utilization. *Journal of Mathematical Psychology*, *110*, 102695.

Himmelstein, M., & Budescu, D. V. (2023). Preference for human or algorithmic forecasting advice does not predict if and how it is used. *Journal of Behavioral Decision Making*, *36*(1), e2285.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive psychology*, *24*(1), 1–55.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Ignatiev, A. (2020). Towards trustable explainable ai. *IJCAI*, 5154–5158.

Ivers, R., Senserrick, T., Boufous, S., Stevenson, M., Chen, H.-Y., Woodward, M., & Norton, R. (2009). Novice drivers' risky driving behavior, risk perception, and crash risk: Findings from the drive study. *American journal of public health*, *99*(9), 1638–1644.

Jameson, A., Nelson, T. O., Leonesio, R. J., & Narens, L. (1993). The feeling of another person s knowing. *Journal of Memory and Language*, *32*(3), 320–335.

Jansen, R. A., Rafferty, A. N., & Griffiths, T. L. (2021). A rational model of the dunning–kruger effect supports insensitivity to evidence in low performers. *Nature Human Behaviour*, *5*(6), 756–763.

Jansen, R. A., Rafferty, A. N., & Griffiths, T. (2020). A rational model of sequential self-assessment. *CogSci*.

JASP Team. (2022). JASP (Version 0.16.2)[Computer software]. https://jasp-stats.org/

Jenkins, A. C., Macrae, C. N., & Mitchell, J. P. (2008). Repetition suppression of ventromedial prefrontal activity during judgments of self and others. *Proceedings of the National Academy of Sciences*, *105*(11), 4507–4512.

Jenness, A. (1932). The role of discussion in changing opinion regarding a matter of fact. *The Journal of Abnormal and Social Psychology*, *27*(3), 279.

Jo, A. (2023). The promise and peril of generative ai. *Nature*, *614*(1), 214–216.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

Jost, J. T., Kruglanski, A. W., & Nelson, T. O. (1998). Social metacognition: An expansionist review. *Personality and Social Psychology Review*, *2*(2), 137–154.

Kämmer, J. E., Choshen-Hillel, S., Müller-Trede, J., Black, S. L., & Weibler, J. (2023). A systematic review of empirical studies on advice-based decisions in behavioral and organizational research. *Decision*.

Kausel, E. E., Culbertson, S. S., Leiva, P. I., Slaughter, J. E., & Jackson, A. T. (2015). Too arrogant for their own good? why and when narcissists dismiss advice. *Organizational Behavior and Human Decision Processes*, *131*, 33–50.

Kelley, C. M., & Jacoby, L. L. (1996). Adult egocentrism: Subjective experience versus analytic bases for judgment. *Journal of Memory and language*, *35*(2), 157–175.

Kelly, M., Kumar, A., Smyth, P., & Steyvers, M. (2023). Capturing humans' mental models of ai: An item response theory approach. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1723–1734.

Khashabi, D., Min, S., Khot, T., Sabharwal, A., Tafjord, O., Clark, P., & Hajishirzi, H. (2020). Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*.

Kington, R., Reuben, D., Rogowski, J., & Lillard, L. (1994). Sociodemographic and health factors in driving patterns after 50 years of age. *American journal of public health*, *84*(8), 1327–1329.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of experimental psychology: General*, *126*(4), 349.

Koriat, A. (2000). The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and cognition*, *9*(2), 149–171.

Koriat, A., & Ackerman, R. (2010). Metacognition and mindreading: Judgments of learning for self and other during self-paced study. *Consciousness and cognition*, *19*(1), 251–264.

Koriat, A., & Levy-Sadot, R. (1999). Processes underlying metacognitive judgments: Information-based and experience-based monitoring of one's own knowledge.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, *77*(6), 1121.

Kumar, A., Patel, T., Benjamin, A. S., & Steyvers, M. (2021). Explaining algorithm aversion with metacognitive bandits. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43).

Larrick, R. P., Burson, K. A., & Soll, J. B. (2007). Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not). *Organizational Behavior and Human Decision Processes*, *102*(1), 76–94.

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management science*, *52*(1), 111–127.

Lehsing, C., Jünger, L., & Bengler, K. (2019). Don't drive me my way: Subjective perception of autonomous braking trajectories for pedestrian crossings. *Proceedings of the Tenth International Symposium on Information and Communication Technology*, 291–297.

Leibert, T. W., & Nelson, D. L. (1998). The roles of cue and target familiarity in making feeling of knowing judgments. *The American journal of psychology*, *111*(1), 63.

Liang, G., Sloane, J. F., Donkin, C., & Newell, B. R. (2022). Adapting to the algorithm: How accuracy comparisons promote the use of a decision aid. *Cognitive research: principles and implications*, *7*(1), 1–21.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*.

Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. L. (2018). Rational metareasoning and the plasticity of cognitive control. *PLoS computational biology*, *14*(4), e1006043.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103.

Logg, J. M. (2017). Theory of machine: When do people rely on algorithms? *Harvard Business School working paper series# 17-086*.

Lu, Z., & Yin, M. (2021). Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16.

Lubars, B., & Tan, C. (2019). Ask not what AI can do, but what AI should do: Towards a framework of task delegability. *Advances in Neural Information Processing Systems*, *32*.

Machado-León, J. L., de Oña, J., de Oña, R., Eboli, L., & Mazzulla, G. (2016). Socio-economic and driving experience factors affecting drivers' perceptions of traffic crash risk. *Transportation research part F: traffic psychology and behaviour*, *37*, 41–51.

Mehrotra, S., Hunter, J., Konishi, M., Akash, K., Zhang, Z., Misu, T., Kumar, A., Reid, T., & Jain, N. (2023). Trust in shared automated vehicles - study on two mobility platforms. *102nd Annual Meeting of the Transportation Research Board (TRB)*.

Mei, W., Friedkin, N. E., Lewis, K., & Bullo, F. (2017). Dynamic models of appraisal networks explaining collective learning. *IEEE Transactions on Automatic Control*, *63*(9), 2898–2912.

Michelon, P., & Zacks, J. M. (2006). Two kinds of visual perspective taking. *Perception & psychophysics*, *68*, 327–337.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM*, *38*(11), 39–41.

Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of cognitive neuroscience*, *17*(8), 1306–1315.

Moore, D. A. (2007). Not so above average after all: When people believe they are worse than average and its implications for theories of bias in social comparison. *Organizational Behavior and Human Decision Processes*, *102*(1), 42–58.

Moore, D. A., & Cain, D. M. (2007). Overconfidence and underconfidence: When and why people underestimate (and overestimate) the competition. *Organizational Behavior and Human Decision Processes*, *103*(2), 197–213.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological review*, *115*(2), 502.

Morey, R. D., Rouder, J. N., & Jamil, T. (2018). Bayesfactor: Computation of bayes factors for common designs. r package version 0.9. 12-4.2.

Natarajan, M., Akash, K., & Misu, T. (2022). Toward adaptive driving styles for automated driving with users' trust and preferences. *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 940–944.

National Academies of Sciences, E., & Medicine. (2021). *Human-ai teaming: State of the art and research needs*. The National Academies Press. https://doi.org/10.17226/26355

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological bulletin*, *95*(1), 109.

Nelson, T. O., Kruglanski, A. W., & Jost, J. T. (1998). Knowing thyself and others: Progress in metacognitive social psychology.

Nelson, T. O., & Narens, L. (1980). Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of verbal learning and verbal behavior*, *19*(3), 338–368.

Nicholson, T., Williams, D. M., Lind, S. E., Grainger, C., & Carruthers, P. (2021). Linking metacognition and mindreading: Evidence from autism and dual-task investigations. *Journal of Experimental Psychology: General*, *150*(2), 206.

Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological bulletin*, *125*(6), 737.

Nickerson, R. S., Baddeley, A., & Freeman, B. (1987). Are people's estimates of what other people know influenced by what they themselves know? *Acta Psychologica*, *64*(3), 245–259.

Norman, D. A. (2014). Some observations on mental models. In *Mental models* (pp. 15–22). Psychology Press.

Önkal, D., Gönül, M. S., Goodwin, P., Thomson, M., & Öz, E. (2017). Evaluating expert advice in forecasting: Users' reactions to presumed vs. experienced credibility. *International Journal of Forecasting*, *33*(1), 280–297.

Park, J. S., Barber, R., Kirlik, A., & Karahalios, K. (2019). A slow algorithm improves users' assessments of the algorithm's accuracy. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–15.

Patel, B. N., Rosenberg, L., Willcox, G., Baltaxe, D., Lyons, M., Irvin, J., Rajpurkar, P., Amrhein, T., Gupta, R., Halabi, S., et al. (2019). Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ Digital Medicine*, *2*(1), 1–10.

Paulus, M., Tsalas, N., Proust, J., & Sodian, B. (2014). Metacognitive monitoring of oneself and others: Developmental changes during childhood and adolescence. *Journal of Experimental Child Psychology*, *122*, 153–165.

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., Cavazos, J. G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., et al. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, *115*(24), 6171–6176.

Plummer, M., et al. (n.d.). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–52.

Rajpurkar, P., O'Connell, C., Schechter, A., Asnani, N., Li, J., Kiani, A., Ball, R. L., Mendelson, M., Maartens, G., van Hoving, D. J., Griesel, R., Ng, A. Y., Boyles, T. H., & Lungren, M. P. (2020). CheXaid: Deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. *npj Digital Medicine*, *3*(1). https://doi.org/10.1038/s41746-020-00322-2

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests.* ERIC.

Rastogi, C., Zhang, Y., Wei, D., Varshney, K. R., Dhurandhar, A., & Tomsett, R. (2022). Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW1), 1–22.

Reckase, M. D. (2009). Multidimensional item response theory models. In *Multidimensional item response theory* (pp. 79–112). Springer.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015a). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015b). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Sayres, R., Taly, A., Rahimy, E., Blumer, K., Coz, D., Hammel, N., Krause, J., Narayanaswamy, A., Rastegar, Z., Wu, D., et al. (2019). Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology*, *126*(4), 552–564.

Schaffer, J., O'Donovan, J., Michaelis, J., Raglin, A., & Höllerer, T. (2019). I can do better than your AI: Expertise and explanations. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 240–251.

Schultze, T., Gerlach, T. M., & Rittich, J. C. (2018). Some people heed advice less than others: Agency (but not communion) predicts advice taking. *Journal of Behavioral Decision Making*, *31*(3), 430–445.

Schultze, T., Mojzisch, A., & Schulz-Hardt, S. (2017). On the inability to ignore useless advice. *Experimental Psychology*.

Shimamura, A. P. (2000). Toward a cognitive neuroscience of metacognition.

Shum, M., Kleiman-Weiner, M., Littman, M. L., & Tenenbaum, J. B. (2019). Theory of minds: Understanding behavior in groups through inverse planning. *Proceedings of the AAAI conference on artificial intelligence*, *33*(01), 6163–6170.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sjöberg, L. (2003). Distal factors in risk perception. *Journal of Risk Research*, *6*(3), 187–211. https://doi.org/10.1080/1366987032000088847

Smyth, M. M., Collins, A. F., & Morris, P. E. (1994). *Cognition in action*. Psychology Press.

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 780.

Stan Development Team. (2020). RStan: The R interface to Stan [R package version 2.21.2]. http://mc-stan.org/

Steyvers, M., & Kumar, A. (2022). Three challenges for ai-assisted decision-making.

Steyvers, M., Tejeda, H., Kerrigan, G., & Smyth, P. (2022). Bayesian modeling of human-ai complementarity in image classification. *Proceedings of the National Academy of Sciences*.

Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L., & Smyth, P. (2024). The calibration gap between model and human confidence in large language models. *arXiv preprint arXiv:2401.13835*.

Stocker, A., & Shaheen, S. (2017). *Shared automated vehicles: Review of business models* (International Transport Forum Discussion Paper No. 2017-09). Paris, Organisation for Economic Co-operation; Development (OECD), International Transport Forum. http://hdl.handle.net/10419/194044

Tan, S., Adebayo, J., Inkpen, K., & Kamar, E. (2018). Investigating human+ machine complementarity for recidivism predictions. *arXiv preprint arXiv:1808.09123*.

Tauber, S., & Steyvers, M. (2011). Using inverse planning and theory of mind for social goal inference. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33*(33).

Tejeda, H., Kumar, A., Smyth, P., & Steyvers, M. (2022). Ai-assisted decision-making: A cognitive modeling approach to infer latent reliance strategies. *Under review*.

Thomas, R. C., & Jacoby, L. L. (2013). Diminishing adult egocentrism when estimating what others know. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 473.

Tullis, J. G. (2018). Predicting others' knowledge: Knowledge estimation as cue utilization. *Memory & cognition*, *46*(8), 1360–1375.

Turner, B. M., Steyvers, M., Merkle, E. C., Budescu, D. V., & Wallsten, T. S. (2014). Forecast aggregation via recalibration. *Machine learning*, *95*(3), 261–289.

Vaccaro, A. G., & Fleming, S. M. (2018). Thinking about thinking: A coordinate-based meta-analysis of neuroimaging studies of metacognitive judgements. *Brain and neuroscience advances*, *2*, 2398212818810591.

van der Linden, W. J., & Hambleton, R. K. (2013). *Handbook of modern item response theory.* Springer Science & Business Media.

Wegner, D. M. (1987). Transactive memory: A contemporary analysis of the group mind. In *Theories of group behavior* (pp. 185–208). Springer.

Whalen, J., Mouza, C., et al. (2023). Chatgpt: Challenges, opportunities, and implications for teacher education. *Contemporary Issues in Technology and Teacher Education*, *23*(1), 1–23.

Wright, D. E., Lintott, C. J., Smartt, S. J., Smith, K. W., Fortson, L., Trouille, L., Allen, C. R., Beck, M., Bouslog, M. C., Boyer, A., et al. (2017). A transient search using combined human and machine classifications. *Monthly Notices of the Royal Astronomical Society*, *472*(2), 1315–1323.

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational behavior and human decision processes*, *93*(1), 1–13.

Yaniv, I., & Choshen-Hillel, S. (2012). When guessing what another person would say is better than giving your own opinion: Using perspective-taking to improve advice-taking. *Journal of Experimental Social Psychology*, *48*(5), 1022–1028.

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational behavior and human decision processes*, *83*(2), 260–281.

Yaniv, I., & Milyavsky, M. (2007). Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes*, *103*(1), 104–120.

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. *Proceedings of the 2019 chi conference on human factors in computing systems*, 1–12.

Zambrano, A. F., Liu, X., Barany, A., Baker, R. S., Kim, J., & Nasiar, N. (2023). From ncoder to chatgpt: From automated coding to refining human coding. *International conference on quantitative ethnography*, 470–485.

Zell, E., & Krizan, Z. (2014). Do people have insight into their abilities? a metasynthesis. *Perspectives on Psychological Science*, *9*(2), 111–125.

Zhang, H., & Maloney, L. T. (2012). Ubiquitous log odds: A common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in neuroscience*, *6*, 1.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020). Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305.

Appendix A

Appendix Title

A.1 The ordered probit model

The ordered probit model, $x \sim \text{OrderedProbit}(p, v, \sigma)$ is a generative model that maps a (latent) value p to one of $M+1$ discrete scores $x \in \{0, \ldots, M\}$. In this process, noise is added is added to the latent value resulting in a new latent value, $p' = p + \epsilon$, where $\epsilon \sim \text{N}(0, \sigma)$ and the resulting discrete score is determined by the interval where p' lies:

$$x = \begin{cases} 0 & \text{if } p' \leq v_1 \\ 1 & \text{if } v_1 < p' \leq v_2 \\ 2 & \text{if } v_2 < p' \leq v_3 \\ M & \text{if } p' > v_M \end{cases} \tag{A.1}$$

The ordered vector $v = [v_1, \ldots, v_M]$ represents the transition points between different discrete scores. With this construction, the probability of producing a score $x = k$ conditional on the

latent value p is:

$$P(x = k|p, \sigma) = \Phi((v_{k+1} - p)/\sigma) - \Phi((v_k - p)/\sigma) \tag{A.2}$$

where Φ is the cumulative standard normal distribution and $v_0 = -\infty$.

To simplify the model, we divide the 0-1 range into $M + 1$ equal intervals, (i.e., $v = [1/(M + 1), 2/(M + 1), ..., M/(M + 1)]$). With this construction, when $M = 12$ (as in our experiment), a latent value $p' = 1/12$ will result in a score $x = 1$, $p' = 2/12$ will result in a score $x = 2$, etc. Figure A.1 shows an example of how the latent scores are mapped to scores when $M = 6$. Note that the higher value of the parameter σ (top panel) results in a noisier mapping of latent probabilities to discrete scores.

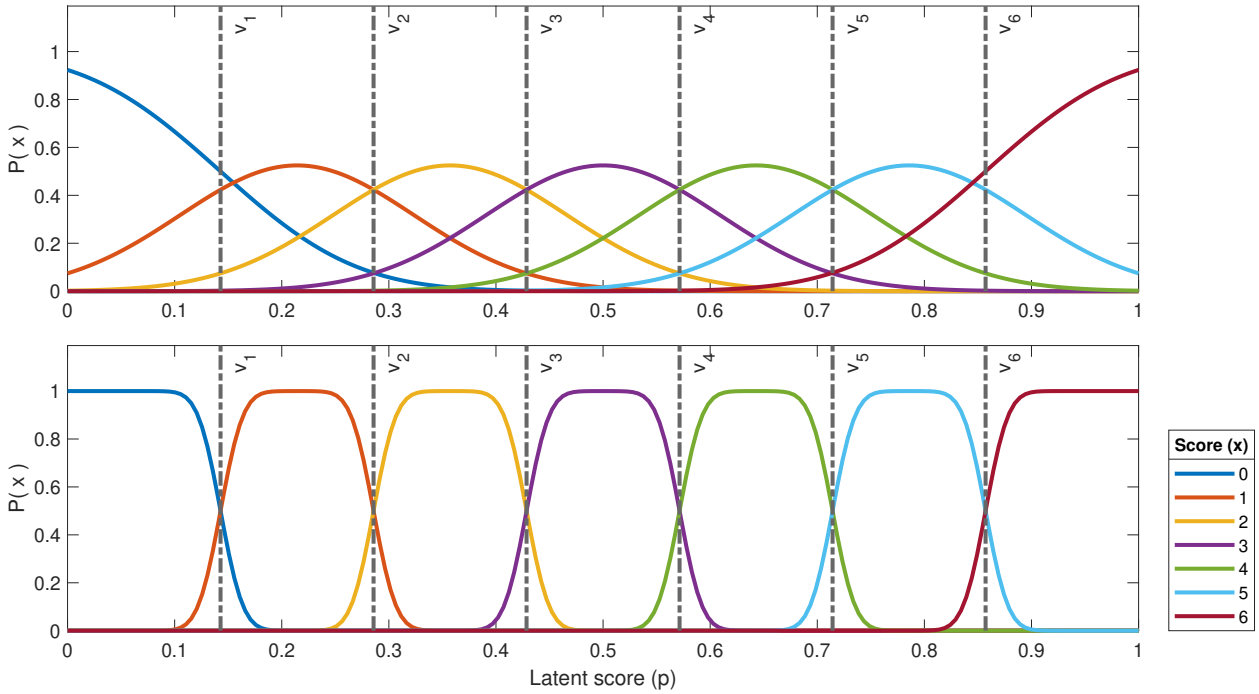


Figure A.1: Illustration of the ordered probit model when $M = 6$. Top and bottom panels are produced with $\sigma = 1/10$ and $\sigma = 1/60$ respectively

A.2 Graphical Models for Self and Other Assessment

Figure A.2 shows the graphical models for the prediction problem corresponding to the three assumptions about the relationship between self- and other assessment. These graphical models illustrate the relationships between the observed and unobserved variables. Note that what is observable or unobserved is all from the perspective of the person reasoning about the other person.

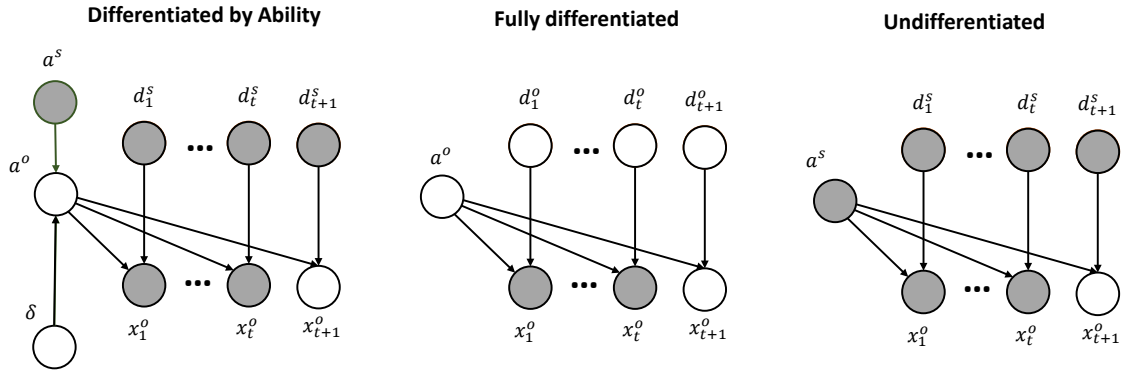


Figure A.2: Graphical models corresponding to three different other-assessment models for predicting the performance of another person. Shaded nodes show information that is known from the perspective of the person reasoning about the other person. Unshaded nodes show latent variables that need to be inferred. The key variable to infer is x_{t+1}^o, the performance of the target person on problem $t+1$.

A.3 Classification Problems

Table A.1 shows a list of the 16 types of classification problems used in the Experiments along with the 4 response options for each classification problem.

#	Category	Response options
1	Bird	Crane (bird), Common redshank, Limpkin, Dunlin
2	Bird	Little blue heron, Oystercatcher, Dowitcher, Great egret
3	Bird	Bustard, Spoonbill, Hornbill, Bittern
4	Bird	Hummingbird, Bald eagle, Vulture, Kite
5	Dog	Shetland Sheepdog, Old English Sheepdog, Rottweiler, Komondor
6	Dog	Lhasa Apso, Airedale Terrier, West Highland White Terrier, Kerry Blue Terrier
7	Dog	Norwich Terrier, Irish Terrier, Scottish Terrier, Norfolk Terrier
8	Dog	Afghan Hound, Ibizan Hound, Norwegian Elkhound, Redbone Coonhound
9	Primate	Macaque, Titi, White-headed capuchin, Guenon
10	Primate	Langur, Black-and-white colobus, Marmoset, Common squirrel monkey
11	Primate	Gorilla, Chimpanzee, Gibbon, Baboon
12	Primate	Ring-tailed lemur, Geoffroy's spider monkey, Howler monkey, Siamang
13	Reptile	Green iguana, Desert grassland whiptail lizard, European green lizard, Carolina anole
14	Reptile	Ring-necked snake, Eastern hog-nosed snake, Vine snake, Worm snake
15	Reptile	Smooth green snake, Night snake, Kingsnake, Saharan horned viper
16	Reptile	Indian cobra, Sea snake, Water snake, Garter snake

Table A.1: List of the classification problems by basic category

A.4 Coefficient of Predictive Ability (CPA)

CPA is a rank-based measure that generalizes the Area under the Curve (AUC) to ordinal and continuous variables. For binary outcomes CPA equals AUC, and for continuous outcomes CPA relates linearly to Spearman's coefficient. We direct the readers to Ref. Gneiting and Walz, 2021 for a detailed discussion on CPA.

Consider data of the form:

$$(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R} \times \mathbb{R}, \tag{A.3}$$

where x_i and y_i are real numbers, for $i = 1, \ldots, n$. Let $z_1 < \cdots < z_m$ denote the $m \leq n$ unique values of $y_1, \ldots, y_n$, and define $n_c = \sum_{i=1}^{n} \mathbb{1}\{y_i = z_c\}$ such that $n_1 + \cdots + n_m = n$. We can reorder and write (A.3) as

$$(x_{11}, z_1), \ldots, (x_{1n_1}, z_1), \ldots, (x_{m1}, z_m), \ldots, (x_{mn_m}, z_m) \in \mathbb{R} \times \mathbb{R}, \tag{A.4}$$

where $x_{i1}, x_{i2}, \ldots, x_{in_i}$ represent the n_i different values of x corresponding to $y = z_i$. This

allows us to compute the CPA as the following

$$CPA = \frac{\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} (j-i)\, s(x_{ik}, x_{jl})}{\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} (j-i)\, n_i n_j}. \tag{A.5}$$

where s is:

$$s(x, x') = \mathbb{1}\{x < x'\} + \frac{1}{2}\mathbb{1}\{x = x'\}, \tag{A.6}$$

A.5 Simulation Details for Tullis (2018)

Tullis, 2018 explores how people use a variety of metacognitive cues to infer the proportion of other people who know the answer to general knowledge questions. This section provides details on the simulation studies we conducted to apply our proposed hierarchical model to the data from Experiments 1 and 2. Since we do not have access to the raw experimental data from the paper, we simulate experimental data for Experiments 1 and 2 and then apply our model to simulate the inference process of others' performance.

To simulate data at the participant level, we randomly generated ability levels, $a_i \sim \mathrm{N}(0, 1)$, for 128 simulated participants who are performing the assessment, as well as 128 other participants to serve as a set of other participants. At the question level we randomly generated the difficulty levels for 40 questions, $d_j \sim \mathrm{N}(\mu_d, \sigma_d)$, where $\mu_d = 1$ and σ_d are simulation parameters that determine overall mean performance and variability in question difficulty. For the self-assessed abilities, we use the same process as in Eq. 2.2, to model the self-assessed abilities, $a_i^s \sim \mathrm{N}(a_i, \sigma_a)$, where parameter σ_a determines the noise in self-assessment. We use the IRT model in Eq. 4.8 to calculate $p_{i,j}$, the true probability of correctly

answering a question for every person i on every question j.

The true probability of being correct (p) is used to generate different knowledge signals, including: feeling of knowing (x^{FK}), response time (x^{RT}), and accuracy (x^{ACC}). We assume feeling of knowing is a random draw from a normal centered around $p_{i,j}$ and with an individual specific variance δ_i:

$$x_{i,j}^{FK} = p_{i,j} + \mathrm{N}(\mu_{FK}, \delta_i), \qquad \delta_i \sim \mathrm{Uniform}(0, \eta) \tag{A.7}$$

Lower values of δ_i correspond to less noise in a participant's feeling of knowing and simulation parameter η determines the degree of noise. To simulate response times, we assume an inverse relationship between RT and $p_{i,j}$:

$$\begin{aligned} x_{i,j}^{RT} &\sim \mathrm{LogNormal}\left(\frac{K}{p_{i,j} + \epsilon_{i,j} + .01}, \nu\right) \\ \epsilon_{i,j} &\sim \mathrm{N}(\omega, \zeta_i), \qquad \zeta_i \sim \mathrm{Uniform}(a, b) \end{aligned} \tag{A.8}$$

where $\epsilon_{i,j}$ is individual-specific noise in response time signals and .01 is added to avoid numerical instabilities. Simulation parameter ν determines the noise in the relationship between RT and accuracy. Figure A.3 shows the RT distribution for different values of $p_{i,j}$. Our assumption results in people having higher RT for problems they have a lower probability of answering correctly and lower RT for problems they have a higher probability of answering correctly.

We model participants' correctness on each problem j as a Bernoulli draw with probability $p_{i,j}$

$$x_{i,j}^{ACC} \sim \mathrm{Bern}(p_{i,j}) \tag{A.9}$$

To simulate the different experimental conditions of Experiment 1 and 2, we follow the logic

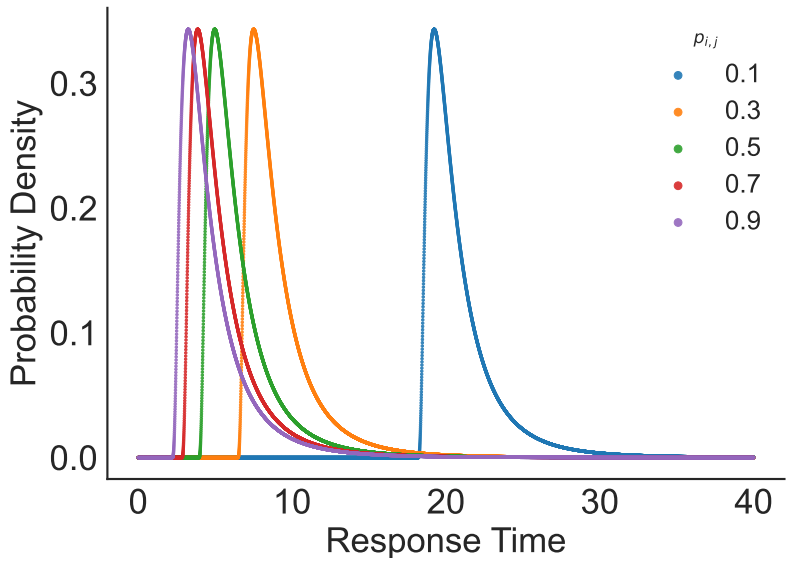


Figure A.3: Simulated response time distributions for different values of $p_{i,j}$ and $K = 2$, $\nu = 2$.

of Table 2.4 that determines which knowledge signals are available in each condition. Next, we apply the hierarchical model of knowledge assessment on the simulated data. Based on the observed knowledge signals x and the long-term self-estimate of ability a^s, the goal for the participant is to infer $x^{o,ACC}$ (which in this setup represents the performance of a randomly sampled person from the population). We used MCMC sampling to infer model parameters for the cognitive model presented in Figure 2.1A with different metacognitive signals x and obtain samples from the posterior distribution of a^o. We used the Stan computing environment for posterior inference Stan Development Team, 2020.

For simulating the experimental data, we use model parameters $K = 2$, $\mu_d = 1$, $\sigma_d = 2$, $\sigma_a = 0.5$, $\eta = .5$, $\nu = 2$, $\omega = .5$, $a = .03$, and $b = .06$. To create a stronger sense of feeling of knowing when participants were asked to answer questions before evaluating the performance of others, we used a value of $\mu_{FK} = .3$. For the answer-after condition, where participants assessed performance before providing their own answers, we used $\mu_{FK} = .5$. As we do not have the raw experimental data available, the goal was not to pursue quantitative model fits and instead show that the model can capture the results from Tullis, 2018 at a qualitative level.

We found that the a wide range of parameter values produce qualitatively similar model predictions. Note that we used the same set of parameters to generate model predictions for all the experiments from Tullis, 2018 and Moore and Healy, 2008 in Appendix A.6.

A.6 Simulation Details for Moore & Healy(2008)

This section provides details on the simulation studies we conducted to apply the hierarchical model to the experiment from Moore and Healy, 2008 . The authors present a synthesis of different ways in which overconfidence has been defined in the literature including the overestimation of one's actual performance and the overestimation of one's performance relative to others. The experimental results show that these forms of overconfidence manifest differently depending on the difficulty of the task. Since we do not have access to the raw data, we simulate data for the experiment presented in the paper, including different levels of difficulty, and apply the hierarchical model to predict how people assess their own performance and place themselves relative to others.

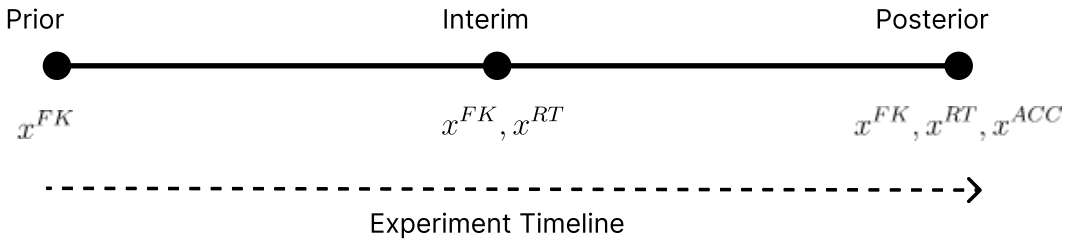


Figure A.4: Timeline of the experiment in Moore & Healy (2008) with the hypothesized metacognitive signals available to participants shown in parentheses.

In the experiment, 82 participants answer 10 questions in 18 categories of trivia questions and predict their own score and the score of 1 randomly selected previous participant (RSPP) at three different stages of the experiment for each category. Figure A.4 shows the timeline of the

experiment and the hypothesized metacognitive signals available to participants when assessing their own performance and the performance of another person. First, participants made prior predictions about themselves and the RSPP before they had any specific information about the quiz they were about to take. Second, they answered 10 quiz questions from a category and then estimated their own scores and the RSPP's score again. This is termed their 'interim' estimate. Next, participants are shown the correct answers to the quiz and asked to make 'posterior' estimates about their performance and the RSPP's performance. Finally, they were given feedback about their own scores and the RSPP's scores.

We focus our model predictions on the interim stage of the experiment. We use the same process used for the Tullis data (Appendix E) with the same simulation parameters (μ_{FK} = .3, μ_d = 1, σ_d = 2, σ_a = 0.5, η = .5, ν = 2) to generate the experimental data for 10 questions and 82 participants. Next, we apply the hierarchical model from Figure 2.1A, Eqs. A.7-A.8 and the same setup as used in Appendix E to obtain the participant's self and other estimates of the number of questions scored correctly out of 10 trivia questions, $\hat{x}^{o,ACC}$ and $\hat{x}^{s,ACC}$. We use a binomial link function to simulate these scores, $x^{ACC} \sim \text{Bin}(10, p_{i,j})$. On the basis of the simulated actual scores ($x^{s,ACC}$ and $x^{o,ACC}$) and the person estimated self and other performance ($\hat{x}^{o,ACC}$ and $\hat{x}^{s,ACC}$), we calculate two empirical measures used by Moore and Healy, 2008. First, we assess the degree of *overestimation*, based on the participant's actual score subtracted from their estimated score, $\hat{x}^{s,ACC} - x^{s,ACC}$. Second, we assess the degree of *overplacement:* which measures whether a participant's assessment of themselves relative to others is in line with the actual observed difference, $(\hat{x}_i^{ACC} - \hat{x}_j^{ACC}) - (x_i^{ACC} - x_j^{ACC})$ where $\hat{x}_i^{ACC}$ is an individual's estimate of their own expected performance, $\hat{x}_i^{ACC}$ is their estimate of another person's expected performance on the same problem, and x_i^{ACC} and x_j^{ACC} refer to the actual scores of the individual and the other person.

A.7 MIRT Model Details and Priors

All models were fit using Stan. We present the details of the multidimensional extension of the hierarchical framework.

To convert latent scores $\theta_{i,j}$ to discrete scores, we used:

$$p_{i,j} = \frac{1}{1 + \exp(-\theta_{i,j})}$$

$$x_{i,j} \sim \text{OrderedProbit}(p_{i,j}, v, \sigma)$$

where v is an array of cutoff points for conversion to discrete scores. We used 11 equally-spaced bins between 0 and 1 (converting into a score between 0 and 12, the number of questions in each problem set).

A.7.1 True Model

Multi-dimensional

$$x_{i,j} = f(\lambda_j \cdot \mathbf{a}_i, d_j, \sigma)$$

$$\sigma \sim \text{Cauchy}(0, 2)$$

$$d_j \sim \text{N}(\mu_d, \sigma_d)$$

$$\mu_d \sim \text{N}(0, 2)$$

$$\sigma_d \sim \text{Cauchy}(0, 5)$$

$$\mathbf{a}_i \sim \text{MVN}_{\text{Cholesky}}(\mathbf{0}, \Sigma_L)$$

$$\Sigma_L = L_{\text{std}} \cdot L_\Omega$$

$$L_\Omega \sim \text{lkj\_corr\_cholesky}(1)$$

$$L_{\text{std}} \sim \text{N}(0, 2.5)$$

One-dimensional

$$Y_{i,j} = f(a_i, d_j, \sigma)$$

$$\sigma \sim \text{Cauchy}(0, 2)$$

$$d_j \sim \text{N}(\mu_d, \sigma_d)$$

$$\mu_d \sim \text{N}(0, 2)$$

$$\sigma_d \sim \text{Cauchy}(0, 5)$$

$$a_i \sim \text{N}(0, 1)$$

A.7.2 Self-Assessment Model

Multi-dimensional

$$x^s_{i,j} = f(\lambda_j \cdot \mathbf{a}^s_i, d^s_j, \sigma^s)$$

$$\sigma^s \sim \text{Cauchy}(0, 2)$$

$$d^s_j \sim \text{N}(\gamma \cdot d_j + \Lambda, \sigma_{d,i})$$

$$\gamma \sim \text{N}(0, 1)$$

$$\Lambda \sim \text{N}(0, 1)$$

$$\sigma_{d,i} \sim \text{Cauchy}(0, 2)$$

$$a^s_{i,k} \sim \text{N}(a_{i,k}, \sigma_{a,i})$$

$$\sigma_{a,i} \sim \text{Cauchy}(0, 2)$$

One-dimensional

$$x^s_{i,j} = f(a^s_i, d^s_j, \sigma^s)$$

$$\sigma^s \sim \text{Cauchy}(0, 2)$$

$$d^s_j \sim \text{N}(\gamma \cdot d_j + \Lambda, \sigma_{d,i})$$

$$\gamma \sim \text{N}(0, 1)$$

$$\Lambda \sim \text{N}(0, 1)$$

$$\sigma_{d,i} \sim \text{Cauchy}(0, 2)$$

$$a^s_i \sim \text{N}(a_i, \sigma_{a,i})$$

$$\sigma_{a,i} \sim \text{Cauchy}(0, 2)$$

A.7.3 Other-Assessment Models

Undifferentiated

Multi-dimensional

$x^o_{i,j} = f(\lambda_j \cdot \mathbf{a}^s_i, d^s_j, \sigma^s)$

Input data: $\mathbf{a}^s_i, d^s_j, \sigma^s$

One-dimensional

$x^o_{i,j} = f(a^s_i, d^s_j, \sigma^s)$

Input data: a^s_i, d^s_j, σ^s

Differentiated by Ability

Multi-dimensional

$x^o_{i,j} = f(\lambda_j \cdot \mathbf{a}^o_i, d^s_j, \sigma^s)$

$a^o_{i,k} = a^s_{i,k} + \delta_{i,k}$

$\delta_{i,k} \sim \mathrm{N}(\mu_{\delta_i}, \sigma_\delta)$

$\mu_{\delta_i} \sim \mathrm{N}(0, 1)$

$\sigma_\delta \sim \mathrm{Cauchy}(0, 2)$

Input data: $\mathbf{a}^s_i, d^s_j, \sigma^s$

One-dimensional

$x^o_{i,j} = f(a^o_i, d^s_j, \sigma^s)$

$a^o_i = a^s_i + \delta_i$

$\delta_i \sim \mathrm{N}(\mu_{\delta_i}, \sigma_\delta)$

$\mu_{\delta_i} \sim \mathrm{N}(0, 1)$

$\sigma_\delta \sim \mathrm{Cauchy}(0, 2)$

Input data: a^s_i, d^s_j, σ^s

Fully Differentiated

Multi-dimensional

$x^o_{i,j} = f(\lambda_j \cdot \mathbf{a}^o_i, d^o_j, \sigma^s)$

$d^o_j \sim \mathrm{N}(\mu^o_d, \sigma^o_d)$

$\mu^o_d \sim \mathrm{N}(0, 2)$

$\sigma^o_d \sim \mathrm{Cauchy}(0, 5)$

$a^o_{i,k} \sim \mathrm{N}(0, 1)$

Input data: σ^s

One-dimensional

$x^o_{i,j} = f(a^o_i, d^o_j, \sigma^s)$

$d^o_j \sim \mathrm{N}(\mu^o_d, \sigma^o_d)$

$\mu^o_d \sim \mathrm{N}(0, 2)$

$\sigma^o_d \sim \mathrm{Cauchy}(0, 5)$

$a^o_i \sim \mathrm{N}(0, 1)$

Input data: σ^s

A.8 Adopt, Revise, Decline

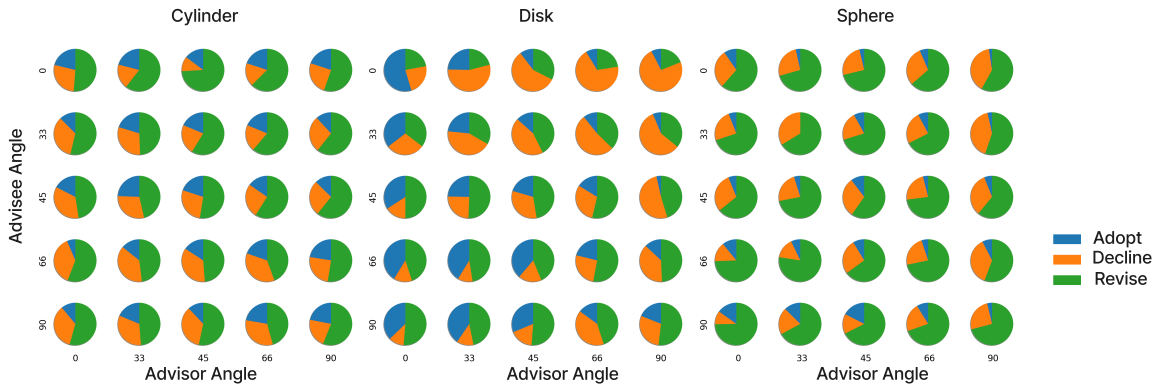


Figure A.5: Proportion of participants' adopt, revise, and decline decisions across shapes and viewing angles

Himmelstein, 2022 proposed a descriptive theoretical model that suggests a discrete 'choosing' stage in advice utilization, where the advisee decides either to decline, adopt, or compromise with the advisor's advice. In line with this framework, we categorized the participants' final estimates in our experiment into three distinct actions: adopting, declining, or revising the advice received during each trial.

We visualize participants' adopt, revise, and decline decisions in Figure A.5, which reveals interesting patterns of how participants responded to advice across different shapes and viewing angles. Specifically for Disks, a notable trend emerges: when an advisee's viewing angle is lower than the advisor's viewing angle, a majority of participants decline the advisor's advice. In contrast, when the advisor's viewpoint is lower than the participant's, the number of participants adopting advice increases substantially as does the number of participants revising advice.

Compared to Disks, Spheres, and Cylinders have a greater proportion of participants who revise their estimates. For cylinders, there is a clear correlation between the advisor's angle and the likelihood of participants either revising or adopting the advice; as the angle of the

advisor increases the proportion of participants who revise and adopt advice increases. For spheres, the proportion of participants who decline advice is greater when the advisor angle is not intermediate. This inclination to decline advice is especially pronounced for 90° angle for the advisor. Across all shapes, a consistent pattern is observed when the viewing angles of the advisor and advisee are aligned: each shape displays a similar distribution of decisions to adopt, decline, or revise the advice.

A.9 Derivation of w_i^{opt}

For each trial i in a collection of n advisor-advisee view combination trials, let t_i be the true number of objects in the jar, p_i be the advisor's estimate, c_i be the advisee's initial estimate, and w be the weight of advice. The participant's objective is to minimize:

$$\min \left(t_i - (w_i p_i + (1 - w_i)c_i)\right)^2$$

Expanding the square, the expression becomes:

$$\min \left(t_i^2 - 2t_i(w_i p_i + (1 - w_i)c_i) + (w_i p_i + (1 - w_i)c_i)^2\right)$$

To find the minimum, we differentiate this expression with respect to w_i and set the derivative to zero. Differentiating yields the following expression:

$$\frac{d}{dw_i}\left(t_i^2 - 2t_i(w_i p_i + (1 - w_i)c_i) + (w_i p_i + (1 - w_i)c_i)^2\right) = 0$$

$$-2t_i(p_i - c_i) + 2(w_i p_i + (1 - w_i)c_i)(p_i - c_i) = 0$$

Rearranging terms, we obtain:

$$w_i(p_i - c_i)^2 = (t_i - c_i)(p_i - c_i)$$

Solving for w_i gives us the optimal weight of advice for trial i, w_i^{opt}:

$$w_i^{opt} = \frac{t_i - c_i}{p_i - c_i}$$

A.10 Assisted Image Classification: Details on Model Inference

We used Markov Chain Monte Carlo (MCMC) to infer model parameters and obtain samples from the posterior distribution, conditioned on the observed data. We chose JAGS for posterior inference (Plummer et al., n.d.). To facilitate posterior inference, the inference procedure was separated into two stages. In the first stage, the observed data $x_{i,j}$, z_j, and $r_{i,j}$ from the no AI assistance condition was used to infer all model parameters related to person and item differences (a_i,d_j,s_j) and confidence generating process (σ_i, v_i). As a result of this stage, we computed posterior predictive distributions for the latent (independent) decisions $x_{i,j}$ and associated confidence levels $r_{i,j}$ for the AI assistance condition. In the second inference stage, the posterior modes of $x_{i,j}$ and $r_{i,j}$ were used as observed data, along with $y_{i,j,k}$, $c_{j,k}$, z_j and $\eta_{j,k}$ to infer the advice-taking model parameters $\alpha_{i,j,k}$. In theory, one does not need to separate the first and second stage of inference and model parameters can be estimated in one joint procedure. We followed this two-stage inference process to facilitate the comparison with the optimization experiments (described in the next section). For both the first and second stage inference process, we ran the sampler with 8 chains with a burn-in of 1000 iterations before taking 50 samples per chain. The chains mixed appropriately. For prior distributions, we used normal priors for the ability and discrimination IRT parameters, consistent with previous Bayesian IRT modeling (Fox, 2010): $a_i \sim \mathcal{N}(0,1)$, $s_j \sim \mathcal{N}(1,1)I(0,)$, where $I(0,)$ denotes truncation a values below zero. Because of the large item differences in the classification task, we use a uniform prior spanning a large range of item differences, $d_j \sim \text{Uniform}(-10,10)$. For the generative process of the confidence levels, we used $\tau_i \sim \text{Uniform}(0,15), \sigma_i = 1/\tau_i$. In addition, we used uniform priors on the two cutpoints needed to produce three levels of confidence, $v_{i,1} \sim \text{Uniform}(0,1)$, $v_{i,2} \sim \text{Uniform}(0,1)$, with the constraint that the cutpoints are ordered (i.e. $v_{i,1} < v_{i,2}$).

Finally, for the advice-taking process, the AI reliance parameter α is treated as a 3 x 4 x 3 lookup table for each individual i where entries are determined by the three confidence levels of the participant ("low", "medium", and "high"), 4 classifier confidence levels (0.00-0.35, 0.35-0.57, 0.57-0.78, 0.78-1.00), and 3 AI classifiers (A, B, and C). The classifier confidence levels were chosen to evenly distribute the observations across bins. Changing notation, the AI reliance parameters can be represented by $\alpha_{i,r,\eta,k}$ where r indexes the participant confidence level and η is the (discretized) AI confidence level. We use a hierarchical Bayesian approach to estimate the individual differences in reliance policies by assuming that these are sampled from a normal distribution on the log-odds scale $\log\left(\frac{\alpha_{i,r,\eta,k}}{1-\alpha_{i,r,\eta,k}}\right) \sim \mathcal{N}(\beta_{r,\eta,k}, \phi)$. The parameter β represents the advice taking policy at the population level, the tendency across participants to accept AI advice. The standard deviation ϕ captures the spread in individual differences. For priors, we use $\beta_{r,\eta,k} \sim \mathcal{N}(0, 3)$. In addition, because there are relatively few "medium" confidence levels, we imposed an order constraint, $\beta_{1,\eta,k} \leq \beta_{2,\eta,k}, \beta_{2,\eta,k} \leq \beta_{3,\eta,k}$ for $\eta = 1, .., 4,$ and $k = 1, ...3$.