SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of Psychology

PhD program in Psychology, Linguistics and Cognitive Neuroscience, 38th Cycle

Curriculum in Mind, Brain and Behavior

Enhancing Human-AI Collaboration: Improving Advice Integration from Artificial Agents

Introzzi Luca

Registration number: 764633

Tutor: Carlo Reverberi

Co-tutor: Marcello Gallucci

Coordinator: Simona Sacchi

ACADEMIC YEAR 2024/2025

Abstract

One of the core tenets in cognitive science is that the human mind is an information processing system with several well-known limitations. Our cognitive architecture constrains the range of data we can gather from the environment, and moreover the use of information we collect, especially the advice from others, is generally adaptive yet suboptimal. Recent developments in computer science have delivered increasingly capable decision support systems, mostly through various forms of Artificial Intelligence, which can be used to augment human information processing by compensating limits in perception, attention, workload, and decision making. However, limitations in decision making processes are especially hard to tackle. Given that humans are and still will be the ultimate decision makers in many domains, it is important to understand how to make Human-AI Collaboration effective the most. Here we studied four possible areas of intervention: competence of the adviser, decision stakes, individual variability, design of the decision context. We also studied how advice integration might work between two large multimodal models, to gain further insights on how AI might be used to augment human decision making, and to understand the potential of AI-AI Collaboration. Our results show first that human decision makers adopt advice from others and that adoption improves final performance if the advice is of good quality. We show moreover that advice taking remains suboptimal, despite improvements, and that it augments decisions without actually reaching a synergy with the adviser on average: final assisted decisions are more accurate as compared to the decisions taken alone, but they are not on average superior to the decisions of the best evaluator available. Second, the competence of the adviser leads to better improvements in advice taking as compared to variations in decision stakes in the form of economic incentives, which instead tend to increase response times without noticeable improvements in the accuracy of the final judgment. Third, both a concurrent decision design and a sequential decision design lead to comparable improvements in the final performance, with comparable reliance on an AI adviser. Fourth, we bring evidence of the effects of inter-individual variability in advice integration and of clustering in performance and reliance when interacting with AI advisers. Finally, we evidence similarities and differences in performance and reliance between human minds and Artificial Intelligence. Our work contributes to the literature in Human-AI Collaboration by investigating areas of interventions for making cooperation more effective, highlighting the extent to which performance benefits can be gained with them. It contributes also to the early development of the AI-AI Collaboration field, towards a future in which AI systems will be increasingly autonomous and endowed with agentic capabilities.

Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 The Emerging Potential of Human-AI Collaboration

The field of Artificial Intelligence (AI) has been growing enormously in recent years. The development of parallel distributed processing systems, in the form of neural networks (Rumelhart et al., 1986), allowed us to solve problems previously too complex to tackle and to build, for the first time, "machines who think" (McCorduck, 2018). Great advancements in Computer Vision came from the development of convolutional neural networks, which for the first time matched and, in some cases, even surpassed human ability in image recognition (Cireşan et al., 2012; He et al., 2015; Krizhevsky et al., 2012). Deep Reinforcement Learning allowed to develop AI systems capable of surpassing human abilities in various fields such as chess, Go, and video games (Schrittwieser et al., 2020; Silver et al., 2016, 2018), and training without any human-generated data led to even better results (Silver et al., 2017). As of today, the most outstanding result of the field has been the development of the protein structure database with AlphaFold, an AI system capable of computing the folded tridimensional structure of more than 200 million proteins starting from their amino acid sequence (https://alphafold.ebi.ac.uk/). This success is astonishing compared to what human intelligence has achieved alone: we managed to compute the structure of about two hundred thousand proteins in five decades, whereas AI managed to compute it in a few years, and for a thousand times more proteins than what we did, a problem considered unsolvable in practical times before the advent of AI (Jumper et al., 2021). Neural networks are so effective in handling such a variety of problems because, mathematically, they can be understood as universal approximators (Barron, 1993; Cybenko, 1989), although they are not likely suitable for any class of problems (Pendurkar et al., 2022).

Another problem that AI has successfully tackled is the comprehension of natural language. The development of the Transformer architecture (Vaswani et al., 2017) gave birth to conversational agents, such as ChatGPT, Gemini and Claude, that rapidly became adopted by millions of users worldwide for everyday tasks and became known as *Large Language Models* (LLMs) and *Large Multimodal Models* (LMMs). Whereas conventional neural networks have been used for analyzing the content of images or videos, the Transformer has been used to *generate* content, leading to the so-called Generative AI. While the standard neural network architecture is trained to solve complex but specialized problems, such as image classification, the new architectures are more generalist and can be applied to wider classes of problems and more variable inputs. They can be trained on all sorts of languages, either natural or artificial (such as programming languages), and they have become widely adopted for coding, writing, planning or just conversations out of curiosity. These AI models show remarkable abilities, and several authors have started to study them with tools from psychology and cognitive science (Binz & Schulz, 2023; Li et al., 2024; Loconte et al., 2023; Sartori & Orrù, 2023; Wei

et al., 2022; Yang et al., 2023). Theory of Mind, the ability to understand one's mental states and the states of others, seems to have emerged in LLMs as their development progressed (Gandhi et al., 2023; Kosinski, 2023; Leslie et al., 2004). In particular, artificial reasoning is the ultimate frontier. Recent efforts have led to AI capable of handling mathematical formal reasoning (DeepMind, 2025; Trinh et al., 2024) and to the adoption of machine learning in theoretical science (Davies et al., 2021; Douglas, 2022). Furthermore, LLMs have been upgraded with *Chain-of-Though* reasoning (Wei et al., 2023) and with *Tree of Thought* reasoning (Yao et al., 2023), new methods for recursion recently led to *Hierarchical Reasoning Models* (G. Wang et al., 2025) and to *Tiny Recursive Models* (Jolicoeur-Martineau, 2025), supervised learning approaches able to compete in performance with much larger earlier LLMs at a fraction of their parameter size (about 7 million as compared to hundred billions). These developments are endowing AI with gradually improving abilities to systematically reason on human prompts, the last iteration of the long-standing dream of building "machines who think".

It is increasingly realistic to consider our contemporary AI models as artificial minds with their own mental processes, as agents that will work with humans in many fields. The outstanding success of the Transformers in dealing with vast amounts of data and learning from them is currently fueling the next iteration of the architecture, where large databases of behaviors are used to improve movements of robotic effectors (Team et al., 2025), building the first *Large Behavior Models* (LBM). Current Transformer-based AI is showing remarkable abilities in automating tasks with increasing autonomy from human activity. Whereas the first generation of LLMs was able to complete autonomously tasks that required a human several minutes to be completed, relative to a fixed success measure, current models can now work in autonomy for tasks that would take a human about one hour to complete, according to the same metric, and recent evidence shows that this improvement rate increased in 2024 as compared to previous years. An extrapolation of the current trend would imply that these AI agents in the next five years will be able to complete autonomously tasks that would take a human a full month of work to be completed, at least in some domains (Kwa et al., 2025). The trend is clear: our artificial machines are increasingly autonomous, increasingly capable of tackling longer, more complex and more variegated tasks. Their penetration in our work environments and daily lives, given the current trends, is inevitable. The question is more relative to what shape our interaction with them will take.

Whereas AI is likely to perform many tasks previously done uniquely by humans, and with increasing ability, its adoption for the near future will be framed more not as substitution but as cooperation. Medicine is the realm in which the adoption of AI as a collaborative agent has become evident and instructive (Introzzi, Zonca, et al., 2024). Even well before the advent of AI, decades ago, it was already known that simple models for statistical prediction could perform equally well or even better than predictions by expert clinicians (Dawes et al., 1989; Meehl, 1954). In more recent years, conventional neural networks have been used for medical decision making, for the detection and characterization of a variety of pathological conditions (Chen et al., 2022; Litjens et al., 2017). One intriguing possibility is to use LLMs to assist clinical reasoning (Brügge et al., 2024), and especially to test the potential of current *Large Reasoning Models* (Mondillo et al., 2025). Several studies have investigated their performance in medical realms, but most of them have evaluated the performance of LLMs alone and not as assistants of humans. These studies tentatively indicate that LLMs encode clinical

knowledge and perform well in answering medical questions and problems (Ho et al., 2024; Liu et al., 2024; Rossettini et al., 2024; Scaioli et al., 2023; Singhal et al., 2023). Consequently, it would be interesting to explore the potential of LRMs as assistants for decision-making, and to study them not as standalone systems but in conjunction with humans. For ethical and legal reasons, in fact, it is extremely unlikely that AI will perform any diagnostic task without human assistance and without humans making the final decisions, at least for now. Moreover, even if AI's knowledge in the medical domain seems promising, its integration in the clinical workflow is at present really far from productive and effective (Hager et al., 2024). Inevitably, AI adoption will take the form of collaboration between artificial agents and human decision-makers.

Given the inevitability of cooperation, it is crucial to understand if assistance by AI leads to improvements in decision making and problem-solving. Teaming of humans and AI can sometimes lead to outcomes superior to the performance of each member taken alone, as has been shown in colonoscopy (Reverberi et al., 2022; Rondonotti et al., 2022). Moreover, thinking abilities in humans and in LLMs do not entirely overlap, and more recent LLMs have gotten increasingly better at answering questions and problem-solving, with previous limitations progressively disappearing (Yax et al., 2024). However, superior outcomes are not always granted: LMMs are still not superior to the best clinicians (Singhal et al., 2023), Human-AI collaboration in many cases leads to inferior outcomes (Hager et al., 2024; Vaccaro et al., 2024), and whether the abilities of LRM constitute "reasoning" or not is debated (Lawsen, 2025; Shojaee et al., 2025). On top of that, while the AI field has experienced clear progress in hardware, software and architectures, these improvements have not been equally matched by comparable progresses in the "human factors", and consequently the interaction is often suboptimal. Current challenges for achieving effective Human-AI collaboration include understanding how human decision-makers represent themselves and the AI agents, with their respective strengths and limits, in a comprehensive "theory of artificial mind" (Introzzi, Zonca, et al., 2024); the effect of design choices about the timing of assistance and quantity of information presented (Steyvers & Kumar, 2024); the effect of information asymmetry and capability asymmetry (Hemmer et al., 2024). Indeed, the work presented here is intended as an attempt to understand suboptimality in how humans use advice for decision making, if it is possible to improve it, and how much can we gain from interacting with artificial agents. Building on experimental and empirical knowledge, we hope that our work will help future research in Human-AI collaboration.

1.2 A Cognitive Analysis of Human-AI Collaboration

Multiple perspectives are viable to frame Human-AI Collaboration. Intuitively, it is possible to understand AI as a tool in our hands to perform tasks, the latest available in a long successful series of technological developments, and the main challenge consists in learning how to "handle" it. This account provides us just a basic and partial understanding of how AI might benefit us, mainly because it focuses almost exclusively on the intrapsychic level of the human side of the interaction. It might be helpful in interacting with some AI systems to some extent, those that are "narrow artificial intelligences" (Cristianini, 2025) trained and specialized for a restricted number of tasks, like in the case of computer vision, but it becomes

less and less useful as time passes and the field progresses towards more general forms of AI, capable of handling multiple different tasks and to be prompted on the spot, with the capacity to interact with us in our natural language. Another common intuitive understanding of AI frames it "like us": artificial beings with which we can converse, ask questions, write code, plan travels, and discuss topics. Especially with the development of LRMs, and the wave of enthusiasm around them, it has become commonplace to read or hear opinions about a near future in which AI will evolve to match our abilities or even surpass them. However, also this account is partial and limited. The issue with this view mirrors the previous one: it focuses too much on AI without taking into sufficient consideration its differences from the human mind. The possibility of AI developing human-level or even superior abilities is far from unrealistic (Cristianini, 2023, 2024, 2025), and yet our AI products are not exactly like us, they do not think and reason in the same ways as we do. Their workings, from this point of view, might resemble what already happened in the past with flight: our airplanes with jet engines fly without flapping their wings, realizing a functional process with different means and with different physical laws as compared to biological organisms.

We should consider the possibility that AI is not just another tool among many, waiting for us to use it, but a system with complex inner workings, starting to manifest the hallmarks of an intelligent agent in a psychological sense: a system that not only we can use and that not only reacts to external stimuli, but that can adapt to them to some extent, and that can use its knowledge to predict outcomes and, in the near future, act in the world, next to human beings. Not all AI systems manifest these hallmarks, and not all in the same way, but the rapid evolution of the field suggests that they might appear soon in most of them. Their differences in respect to the human mind allow us to adopt them as assistants, but they will not likely be exactly the same as a human adviser, such as an expert colleague. We need to understand their workings, their "mental" properties, how their advice is considered by human decision makers. For these reasons, in this work, the perspective of cognitive science is adopted. We study the Human-AI team as an instance of any Decision Maker-Adviser team: as an information processing system composed by two agents, each of them with its own characteristics, capable of reacting to stimuli, collecting knowledge, and using it for adaptation, prediction and action in the external world. In cognitive science, "mind" is not a substance but a set of processes realized by a physical system. The two agents are separate and, in principle, might have different processes to analyze information coming from the environment. Their different processes might make them differently specialized, with different strengths and weaknesses ("cognitive blind spots"), possibly partially overlapped to some extent, so that a combination of the two might in part build on reciprocal strengths to the extent that they might compensate for their respective weaknesses. In this sense, AI (or any device giving advice to a human decision maker) cannot be understood simply as a tool extending our reach to the environment. They are fundamentally different from other devices that we developed in the past to "extend our mind", as Clark and Chalmers, 1998 described them, such as books, libraries or even computers. Our AI devices are constantly developed and refined so that they can learn from the environment and from us: we adapt to and learn from them, but they too adapt and learn from us. They can understand our preferences, our tastes, the nuances of our writings and speech, and they can tailor themselves to them. We always had the challenge of studying and understanding how to use our tools, but this is the first time in which our tools have the possibility to study and understand us. In the future,

they will adapt and learn from themselves too, as soon as "AI-AI Collaboration" will rise with the capability to actually manipulate the external world and not just to converse with us.

Our first step is to develop a *cognitive analysis* of Human-AI Collaboration: understanding the anatomy of factors affecting it and organizing them into a taxonomy. Much work of our research group is currently dedicated to studying a closely linked topic: use of social information, integration of advice and metacognitive competence. The Human-AI team indeed is one specific instance of the more general interactive context in which a Decision Maker is in charge of a decision and receives an opinion coming from an Adviser, evaluating the extent to which the advice might be integrated in the final decision. Evaluating the characteristics of the decision maker and of the adviser, and the quality of the advice received, are important steps to calibrate proper reliance on one's own opinion and the opinion of others, weighting them accordingly. This weighing process is just one of the multiple processes acting in advice integration, and we developed a taxonomy of influencing factors, which might be more or less influential depending on the task at hand. Building on our previous work on Human-AI Collaboration in colonoscopy (Introzzi, Cherubini, & Reverberi, 2024; Introzzi, Zonca, et al., 2024), we can consider four classes of factors where limitations of the human cognitive system manifest, and where AI (or any other Decision Support System, DSS) might be adopted to ameliorate them: 1) Perception: reduction of visual acuity from central to peripheral vision (convergence, cortical magnification); 2) Attention: limits of exhaustive search with focussed attention; single-feature search and conjunction search; inattentional blindness; change blindness; inhibition of return; 3) Workload: cognitive load in dual tasks from sustained attention and from task interference; decrements in vigilance and fatigue; 4) Decision Making: miscalibration; automation bias; algorithm aversion; egocentric advice discounting; global-to-local integration deficits.

As we will see below, integration of advice is often suboptimal. The work presented here investigated four classes of factors that might influence advice taking, and that consequently might be used to inform interventions to improve Human-AI Collaboration: 1) competence of the adviser in respect to the decision maker: differences in abilities between the two agents involved in an interaction; 2) Motivation: stakes involved in a decision; 3) Individual differences in advice taking and reliance; 4) Design of the decision context: how the interaction is structured by hardware and software, with concurrent or sequential decisions.

1.2.1 Perceptual Factors

Perception is the process by which the nervous system constructs a multidimensional model of the world starting from the bidimensional sensory input from the eyes. Light from the external world is focussed on the retinas and transduced into electrical ion currents flowing mainly to the visual areas in the occipital cortex. The process of vision is implemented and distributed throughout the visual system, with early analysis of contrasts already happening in the gangliar cells of the retina, and more complex information extraction, such as color, motion or object identification happening in progressively upper circuits in the visual processing stream (Kandel et al., 2021). The model of the world that perception constructs however is not an exact transposition of the visual stimuli it sees: every model is an *interpretation* of the data it is based on. Multiple processes are involved in such interpretation but they are also affected by several limits. Especially, the visual percept does not have a uniform

resolution across all the visual field: the center, perceived by the fovea, is reconstructed in higher resolution as compared to the periphery. Convergence of neuronal projection from the retinal periphery to the cortical visual areas is higher as compared to foveal neuronal projections, resulting in a magnified representation of the center in respect to the periphery with higher resolution and consequently with higher visual acuity, a phenomenon known as *cortical magnification*. The lower peripheral resolution imposes one of the first *cognitive blind spots* we encounter in the human cognitive system: small targets are more difficult to detect in the periphery of the visual field (Lago et al., 2021; Wolfe, 2021). A second blind spot hinders visual search building up on the previous one: to compensate for lower visual acuity in the periphery, exhaustive search might be used, but it requires time and it might be inefficient: visual search might under-explore some areas while re-exploring previously scanned ones, potentially missing even more targets (Lago et al., 2021), hindering detection with both defective and inefficient search. Furthermore, the visual field is often crowded: multiple objects are present and a target must be searched among many distracting objects that might attract our gaze without being relevant to our aims (Manassi & Whitney, 2018). These multiple limitations reduce the useful field of view for the detection of objects to an angle of about 2.5° degrees centered on the fovea, or about an area of 4.4cm wide 1m away: we look but we fail to see (Wolfe et al., 2022).

Limits in human visual search are a well-known problem that affects the analysis of medical images by clinicians (L. H. Williams & Drew, 2019). This is a class of factors where humans clearly benefit from AI support. For instance, colonoscopy greatly benefits from Computer-Aided Detection (CADe) based on convolutional neural networks, as it has been demonstrated in multiple randomized controlled trials (Attardo et al., 2020; Aziz et al., 2020; Hassan et al., 2021; Huang et al., 2022; Wallace et al., 2022). Indeed, this happens for many reasons, one of them being that artificial neural networks do not suffer from any of the blind spots outlined above for humans. Their neural convergence can be coded so that it is uniform across their visual field (by fixing the size of the convolution kernels and of pooling layers). They do not suffer from losses of visual acuity and from inefficient serial visual search, to the point that CNNs ordinarily outperform visual detection and recognition of most targets by humans (Cireşan et al., 2012; He et al., 2015; Krizhevsky et al., 2012). Moreover, even if CNNs have been designed drawing inspiration from the human visual system, their visual abilities are only partially similar to human ones (Lindsay, 2021; Marblestone et al., 2016) and they are characterized by differences in high-level visual processing (Bracci et al., 2019). These characteristics allow to combine the two agents, human and artificial, with partially different abilities and permits the Human-AI team to benefit from them, provided that their reciprocal strengths are understood in the context of the interaction.

1.2.2 Attentional Factors

Attentional processes are a second class of factors that influence the effectiveness of Human-AI Collaboration. The presence of an object in the visual field is necessary but not sufficient to detect it. Attention enhances perceptual processing, allowing us not just to look but to see objects; however it is affected by several limits, some related to the type of target we are looking for, some instead are due to its functioning independently of the type of target. Some objects indeed are immediately visible if a salient feature distinguishes them from

the background and from other distracting objects: a red brick is easily spotted in a grey wall, and a colored letter O stands out among many other O letters written in black on a screen. Visual search in these situations happens fast because of the pop-out effect of the target. However, the same colored O is more difficult to spot if presented together with other letters O colored differently, or together with many letters Q of the same colour. If just one salient feature characterizes an object among many others, visual search quickly finds it. But if we are looking for an object with a conjunction of several salient features, then visual search becomes increasingly slower as the number of salient features grows, and as the variety of objects sharing at least one salient feature, but not others, grows (Treisman & Gelade, 1980; Wolfe, 2001). When multiple features are both salient and relevant, attention must be focussed intentionally, making visual search guided by the relevant features, but drastically decreasing search speed (Wolfe, 2021). This cognitive bottleneck is further exacerbated by the phenomenon of *inattentional blindness*, where attention overlooks salient but not relevant objects. The phenomenon, popularized as "the invisible gorilla" from a famous experiment (Chabris & Simons, 2010; Most et al., 2001), can pose a health hazard in clinical settings, where attention might be drawn to search some specific lesions but potentially missing out other clinically relevant findings (L. Williams et al., 2021). The similar phenomenon of *change blindness* consists in overlooking changes in objects and scenes (Simons & Chabris, 1999). These two attentional pitfalls imply that, when exploring a complex visual scene, we tend to detect and remember objects that have received sustained and focussed attention, at the expense of other potentially relevant ones: a pure form of "cognitive blindness" and of the blind spots in the human cognitive system. The two forms of cognitive blindness above depend on the characteristics of the target and of the visual scene, but attentional processes might underexplore some areas systematically, independently of the type of targets we are searching for. When a region of space is visually explored and no target is found, after about 300ms, attention is inhibited from re-exploring the same region. This *inhibition of return* may potentially make us miss targets especially when the visual scene changes rapidly and the target, absent in the first exploration phase, might appear subsequently, when gaze will be directed elsewhere.

These limitations can also be tackled by AI with success. CNNs are not affected by inhibition of return, change blindness or inattentional blindness. Indeed, modelling of attention in artificial intelligence improved its abilities in many domains (Vaswani et al., 2017). Moreover, AI can assist human visual search by making a multi-feature visual search faster and more accurate, by transforming the slow, serial conjunction search into a fast parallel pop-out search. This can alleviate the impact of inattentional blindness and change blindness, as it is evident in colonoscopy: small polyps are more difficult to distinguish from healthy colonic mucosa, with the optical video flow constantly updating, and they are easier to miss due to their size and color in respect to the surroundings. Instead. AI for CADe can be used to analyze each single frame from the colonoscope: as soon as a polyp is spotted by the AI, a green square visually highlights it, making it pop out from the optic flow with an immediately salient color contrasting with the mucosal background (Introzzi, Zonca, et al., 2024).

1.2.3 Workload

A third class of limitations stems from the cognitive workload that a task might require to human decision makers. Indeed, many tasks are actually dual-task or multi-task procedures. Colonoscopy provides a perfect example: the clinician must detect colonic polyps, characterize them as pathogenous or not, distinguishing them from distracting objects (water droplets, residuals from not optimal cleaning), evaluate if removing the polyps found, all this while navigating with the colonoscope in the bowel, optimize the visualization by keeping the bowel lumen in the centre, maintain proper orientation in the colon, properly exposing the mucosa to visually inspect it and keeping track of the already exposed areas. Visual inspection overlaps with motor orientation and control. None of these tasks can be performed automatically, they require focussed attention sustained for long bouts of time, and they might interfere even if they do not require the same effectors, due to the limited availability of cognitive resources. Such dual-task situations usually lead to deteriorating performance in at least one of their components, due to resource depletion, time pressure, conflicting needs and task switching, as it has been reported both in clinical settings (Coderre et al., 2010; Jorm & O'Sullivan, 2012; Modi et al., 2020; Sherbino & Norman, 2021) and experimental settings (Pashler, 1994). AI contributes in ameliorating the burden of workload. In the example of colonoscopy, AI not only assists visual search, fastening attentional processes by guiding them, but also by providing assistance to a part of the task (detection and characterization), freeing resources that the human user can then use for other processes (navigation and mucosal exposure). By unburdening the clinician from sustaining a focussed attention in serial search, cognitive resources can now be redirected to other components of the task (Cherubini & East, 2023). Moreover, performance might vary depending on the time of the day, all other factors being equal, an AI assistance can mitigate the impact of these oscillations (Lu et al., 2023).

1.2.4 Decision Making

A fourth broad class of influencing factors concerns higher-order processes in decision making (see Table 1.1). These cognitive tendencies might lead to biases in behavior when interacting with advisers in general, whether human or artificial. Some of these processes are more relative to the self-evaluation of the decision maker (such as miscalibration), some others are more relative to the evaluation of the adviser by the decision maker (such as automation bias or algorithm aversion), some others are referred to the relative evaluation of the decision maker in respect to the adviser (egocentric advice discounting) or to a difficulty of adjusting contingent behavior depending on comprehensive features of the agents involved (global-to-local integration deficit). We present these concepts as theoretically distinguished, but they share some similarities, and they are specific instances of more general limits in human metacognitive abilities.

Metacognition is a set of processes by which humans reflect on their cognitive processes, and it informs us about the quality of our decisions. We will encounter it in detail in chapter 3, relative to inter-individual differences. Metacognition is important to evaluate our performance, especially the connection between our accuracy (defined as percentage of correct responses) and confidence (defined as subjective belief of being correct; see Fleming, 2024; Fleming and Lau, 2014. Part of these processes evaluate the efficacy of one's own decisions

(metacognitive monitoring), or they establish the course of actions to take depending on such efficacy evaluations (metacognitive control) (Nelson, 1990). Another useful distinction is between sensitivity and calibration. *Metacognitive sensitivity* consists in discriminating correct from incorrect decisions and understanding the extent to which our confidence predicts our accuracy (quantitatively, the correlation between the two); metacognitive *calibration* refers to the ability to evaluate the confidence of one's own decisions in respect to one's accuracy (quantitatively, the difference between the two) (Fleming, 2024; Fleming & Lau, 2014; Shekhar & Rahnev, 2021b). Metacognition is affected by several limitations. Multiple sources of *metacognitive inefficiency* are known (Shekhar & Rahnev, 2021a, 2021b, 2024), sometimes not dependent on noise sources influencing the perceptual decisions, sometimes linked to decisional processes, with debated interpretations. For instance, some studies report a positive correlation between perceptual accuracy and metacognitive ability (Kruger & Dunning, 1999), others instead interpret these findings as statistical artifacts (McIntosh et al., 2019). A decision maker with good metacognitive sensitivity expresses confidence judgments that correlate positively with accuracy: high confidence when decisions are accurate, low confidence when decisions are inaccurate, so that the confidence judgments can be used to reliably predict accuracy and they are indicative of correct decisions. *Miscalibration* instead consists in a mismatch between the accuracy of a decision maker in a task and the confidence stated for the decisions. A *calibrated* decision maker has a confidence matching the accuracy: if she is correct in 80% of the trials, and if she judges her probability of being correct at around 80%, then her decisions are calibrated. If her stated mean confidence is higher than the accuracy, then the decision maker is *overconfident*; if the stated mean confidence is lower than the accuracy, the decision maker is *underconfident*. Some subjects manifest low accuracy in judgments coupled with a "double burden" of overconfidence and weak metacognitive sensitivity (Ehrlinger & Dunning, 2003; Koriat, 2012). Calibration and sensitivity are conceptually different: it is possible to be sensitive and still be overconfident, for instance when a decision maker expresses confidence judgments highly correlated with decision accuracy, but with a systematic tendency to overestimate the chances of being correct.

A second family of processes in decision making result in a systematic bias towards or against external advice linked to trust and reliance in it. Sometimes a complacent acceptance of advice manifests, where the decision maker tends to rely on external advice, trusting it to a higher degree than warranted. Such *automation bias* has been reported in studies on Decision-Support Systems (DSS), overweighting the advice coming from AI or from supporting algorithms and undervaluing human judgment (Wickens et al., 2015; Zhang et al., 2020). A similar phenomenon has been reported, an *algorithm appreciation* for information coming from automated procedures instead of human judgment (Logg et al., 2019). Instead, sometimes the opposite process of *algorithm aversion* manifests, when the decision maker has low trust towards advice from DSSs, even if it is reliable, rejecting good advice even when correct (Dietvorst et al., 2015; Jussupow et al., 2024). Multiple factors might lead to these opposite biases. High expertise, high confidence and exposure to DSS failures tend to reduce automation bias, whereas low confidence and high task difficulty increase it (Goddard et al., 2014). Also algorithm aversion is influenced by multiple factors (Dietvorst & Bharti, 2020) and it has been framed not only as "aversion to an algorithm" but also as a "preference for humans" (Morewedge, 2022), especially in tasks where the human individual and social

identity might be relevant (Leung et al., 2018), where evaluation criteria are more ambiguous (Castelo et al., 2019), or in "subjective" tasks where multiple answers might be equally plausible, as compared to "objective" tasks where only one correct answer is available (Riva et al., 2022).

Preference for humans outlined above is defined as a general tendency to prefer information produced by human beings instead of artificial systems. But a third family of processes in decision making is specifically related to the decision maker themselves. In particular, a specific preference of one's own opinion might lead to underestimating external advice, leading to a phenomenon called *egocentric advice discounting* (or *ego bias* in short), generally interpreted as an overestimate of one's chances of being correct (Bailey et al., 2022; Bonaccio & Dalal, 2006; Krueger, 2003; Morin et al., 2021; Yaniv, 2004; Yaniv & Kleinberger, 2000). Ego bias, initially discovered in interactive decision making with humans, might also affect decision making in Human-AI teams. Ideally, advice should be integrated by considering multiple sources of available information relative to the source of the opinions to be integrated (as we will see in Chapter 2). All else being equal, if the decision maker and the adviser are equally accurate, with an equal historical record of past accuracy, their opinions should be weighted equally; if the two are calibrated and express confidence judgments highly predictive of their accuracy, high confidence from the adviser and low confidence in the decision maker should indicate to weight the advice of the former more than the advice of the latter. This is not what happens: generally humans show adaptive yet suboptimal integration of advice (Zonca et al., 2025). Opacity of the advice source might worsen integration: for the human user it might be difficult to form expectations about future performance of an AI adviser (Cadario et al., 2021), and even good evaluation of adviser's accuracy and confidence do not ameliorate advice discounting. In this sense, recent results indeed show that suboptimal integration of advice might stem both from egocentric advice discounting and from a *global-to-local integration deficit*: an impairment of translating a global assessment of the competence of an adviser into trial-by-trial adjustments of the decision-maker's weighting and integration process (Zonca et al., 2025).

Considered together, cognitive blind spots in decision making are difficult to tackle. AI can successfully compensate for limitations in perception, attention and workload, but humans are and will remain the ultimate decision makers, at least in the foreseeable future, and we need to understand how to improve interactive decision making with artificial agents in order to make our interaction truly collaborative. Biases are difficult to remove: simply informing humans about biases does not necessarily lead to debiasing their judgment (Cassam, 2017), and it might also unfortunately lead to the paradoxical effect of enhancing overconfidence, by reinforcing the misleading conviction of being now less vulnerable to them (Croskerry et al., 2013). Showing explicit information comparing the decision maker and the adviser, by describing the quality of their estimates (such as with accuracy, confidence and the correlation between the two), does not lead to substantial improvements (Zonca et al., 2025). Interventions relative to explainability of AI systems and DSSs seem to have limited effects (Vaccaro et al., 2024). We decided then to investigate four different promising directions where knowledge might be gained on how to improve the use of advice. These directions will be presented shortly below, and then fully developed in the following dedicated chapters.

Table 1.1: Cognitive limitations in human decision making which bias integration of advice.

Class of Factors	Cognitive Blind Spots
Perception	Reduced visual acuity in peripheral vision (convergence and cortical magnification)
Attention	Slow conjunction search Inhibition of Return
Workload	Dual task load and interference Inattentional Blindness Change Blindness Sustained Attention
Decision-Making	Miscalibration Automation Bias, Algorithm Appreciation Algorithm Aversion Egocentric Advice Discounting Global-to-Local Integration Deficit

1.3 Modulating Effectiveness of Human-AI Collaboration

1.3.1 Relative competence of adviser and decision maker

As we will see in more detail in Chapter 2, decision makers and decision-support systems can be described by a basic set of behavior attributes characterizing their abilities: accuracy (the percentage of correct responses), confidence (the estimate of the probability of being correct in a decision) and confidence predictivity (the correlation between accuracy and confidence). The relative abilities of human participants and advisers (whether other humans or artificial agents) can be studied by using these three traits, from which we can also infer the extent to which the advice is adopted by the decision maker. Previous studies by our research group have shown that human decision makers can integrate advice and gain performance benefits from it, and that they can adapt to the behavioral characteristics of the adviser, even if integration and adaptation are suboptimal (Zonca et al., 2025). Results also show that integration of advice benefits the decision maker with advisers with varying traits, with the largest benefit coming from interaction with an adviser that was more accurate than the participant (all other traits being equal), and with no benefits when the adviser was less accurate. In these studies, the seven artificial agents used as advisers varied each one by just one trait in respect to the participant (more or less accurate, more or less confident, etc). The fact that most advisers led to performance benefits led us to consider the possibility of studying the effect of advisers whose competence is comprehensively superior to the abilities of the decision maker. As we have seen in the introduction, progress in developing more capable AI systems tends to go in this direction, and these will probably be the kind of advisers that we will encounter in the near future. Previous research has indicated that decision makers

might underweight valuable advice even if coming from very accurate advisers (Dietvorst & Bharti, 2020; Dietvorst et al., 2015), but our methodology allows us to study advice integration in greater detail. We wanted to deepen this line of research and to understand if (sub)optimality of advice integration might be modulated by differences in abilities between adviser and decision makers.

1.3.2 Motivational factors

Motivation is a cognitive process by which humans initiate and persist in goal-directed behavior. It is typically conceptualized in two forms: intrinsic motivation stems from the pleasure or interest in the activity itself, whereas extrinsic motivation stems from external sources of engagement, incentives such as expected rewards, punishments, recognitions, whether psychological or economical (APA Dictionary of Psychology, 2018). Extrinsic motivation is typically used to incentivize behaviors and improve commitment to a task, but benefits on performance are not necessarily granted. Research in individual decision making has shown that stakes alone, in the form of monetary incentives, are not guaranteed to enhance performance: they increase time spent on a task and their benefits might depend on cognitive ability (Awasthi & Pratt, 1990) and task complexity (Camerer & Hogarth, 1999). However, other studies indicate that high stakes scenarios might be linked to higher weighing of external advice and lower egocentric discounting (Løhre & Halkjelsvik, 2024) and might improve metacognitive abilities to some extent (Lebreton et al., 2018). Limitations in the design of such experiments however limit interpretation of results. We addressed part of these limitations in our study presented in Chapter 2.

1.3.3 Individual Differences in Advice Taking

The fact that individuals differ in their mental abilities is the core tenet of the psychology of individual differences. Psychologists and cognitive scientists have extensively operationalized such differences as *traits*, and measured them psychometrically in order to understand their structure and variability, and to gain knowledge on how to potentially modify them. For instance, the study of personality traits informs psychotherapy for the development of more effective interventions; the study of intelligence originated from the attempt to develop tailored training programs for children at school. We naturally expect human subjects to vary in the decision making processes outlined above, such as miscalibration or egocentric discounting, and individual differences are known to affect decision making, as we will see in Chapter 3. We wanted to investigate whether traits known to influence individual decision making, such as cognitive ability, personality, metacognition and self-esteem, might also affect decision-making with artificial assistants. We also wanted to investigate if the effect of global variables (characterizing the functioning of the individual as a whole) on advice integration might be mediated by local variables (specifically measuring metacognitive inefficiencies).

1.3.4 Design of the decision context

The fourth factor we considered is the design of the decision context. Much research in decision making adopts the Judge-Adviser System (JAS, Sniezek and Buckley, 1995): the

decision maker ("judge") gives an initial estimate of a quantity or decision relative to a stimulus; then advice is given to the decision maker, who is in charge of a second, final decision. This is a *sequential design*: evaluations from the decision maker and from the adviser follow in series. However other decision paradigms are possible, for instance a *concurrent design* in which the advice is already present at the moment of the decision. In Chapter 4 we investigated the possibility that the two designs might induce different biases. The information coming first might act as anchor for the decision, so that a sequential design, where the first information available is the evaluation by the decision maker, might foster egocentric advice discounting more as compared to a concurrent design, where the advice is present before a decision takes place and might consequently be weighted comparably more. Our investigations into this factor are presented in Chapter 4 and 5.

Chapter 2
The Effect of Competence and Stakes on the Integration of Advice

2.1 Introduction

The use of advice from others is important for decision-making (Morin et al., 2021): it enhances knowledge acquisition, lowers exploration costs, and reduces the costs of individual trial-and-error learning (Burke et al., 2010; Kendal et al., 2018). To use advice properly, the decision maker should adapt their behavior considering its quality, such as the extent to which it is reliable, and considering the characteristics of the adviser and the decision-maker themselves. Indeed, humans can regulate reliance on advice based on several important characteristics, such as expertise of the advisor (Bonaccio & Dalal, 2006; Boorman et al., 2013; Sniezek et al., 2004), its past accuracy (Behrens et al., 2008), confidence (Bahrami et al., 2010; Koriat, 2012, 2015; Sniezek & Buckley, 1995), and access to relevant information (Vélez & Gweon, 2019). However, such integration, and consequently the use of advice, is often suboptimal. In individual decisions, humans often overweight confirming evidence and underweight evidence contradicting their opinions (Molleman et al., 2020) and they tend to seek information confirming the beliefs they already possess (Hart et al., 2009); they tend to rely more on their evaluation than the evaluation of others, preferring their own opinion, the phenomenon called *egocentric advice discounting* or *ego bias* (Bailey et al., 2022; Bonaccio & Dalal, 2006; Krueger, 2003; Morin et al., 2021; Yaniv, 2004; Yaniv & Kleinberger, 2000). In joint decisions with other humans, they fail to integrate past accuracy and confidence (Mahmoodi et al., 2015), and similar phenomena also occur in interaction with artificial agents (Mahmud et al., 2022). Moreover, a second parallel stream of limitations happens in human metacognitive ability, especially in the form of poorly informative confidence signals (Shekhar & Rahnev, 2021a, 2021b) and overconfidence in its three facets of overestimation, overplacement and overprecision, and in particular they overestimate their abilities and their chances of success (Moore & Schatz, 2017).

Previous studies reported here examined how humans integrate others' opinions, considering factors such as accuracy and confidence, suggesting suboptimal information use. However, they tended to consider relevant variables separately, simplifying the decision-making scenario. Instead, multiple cues are generally available to the human subject, and an understanding of their use and their interaction is needed. In fact, a third stream of potential limitations in humans might lie in the ability to adapt to advisers with different characteristics and in different decision contexts: experiments conducted by our research group (Zonca et al., 2025) indicate that suboptimality emerges even if complete information about the adviser and explicit trial-by-trial feedback are given to the decision maker. We wanted to deepen our understanding of the effects of this third stream of limitations, and especially of the role

of stakes in (sub)optimality. Many decisions involve stakes: gains or losses contingent on the outcomes, which might act as rewards or punishments, positive or negative. Previous studies in human decision making show mixed findings: in some cases they do not diminish reliance on suboptimal heuristics (Kunreuther et al., 2002), but in other cases increased stakes might enhance performance (Deutscher et al., 2018), increase reliance on advice (Løhre & Halkjelsvik, 2024), with recent evidence suggesting that performance pressure by high stakes might influence decision making with AI (Haduong & Smith, 2024). Our understanding of the effect of stakes in interactive contexts, however, remains limited because no study has manipulated decision stakes directly so far, nor has suboptimality been directly quantified. With the present study, we present an experimental paradigm for an interactive context in which decision stakes are manipulated directly, the decisor possesses all the complete information about themselves and the advisor that would allow a theoretically optimal decision, and Bayesian modelling is used to quantify optimal advice integration and to compare it to observed behavior. Specifically, we study whether stakes can modulate the weighting of advice and its integration.

2.2 Research Questions

In our study, the Decision Maker (DM) completes a series of decisions, first alone, then in interaction with an Adviser (AD). Information relative to their performance indicators is fully disclosed to the DM. Correct decisions lead to gains and incorrect decisions to losses, quantified in monetary terms. Each trial is characterized by a relatively large ("high stake") or relatively small ("low stake") consequence, fully known to the DM before each decision. The experimental paradigm is based on the Judge-Adviser Task. In each trial:

1. The DM sees a perceptual stimulus and makes an initial decision alone, and states her confidence in the accuracy of that decision (a metacognitive assessment);

2. Then the DM receives the AD's opinion with its stated confidence;

3. Then the DM is in charge of a second decision, in which the judgment on the perceptual stimulus might be adapted consequently to the advice received, and the confidence in this final decision is required;

4. In the end, feedback about the correct response is given, and the monetary gain/loss is shown.

With this setup, all the relevant information is provided, and it's not needed to acquire it; the benefits depend only on the accuracy of the decision. A rational DM should integrate the advice using all available information, calibrating its reliance on relevant information from two sources. The first source is the self-awareness of DM's competence, which is in turn based on judgment accuracy, confidence, past accuracy, past average confidence, and past reliability of the confidence (the extent to which the confidence judgment is predictive of its accuracy). The second source is the assessment of the AD's competence, which is based on a

similar set of cues: its judgment accuracy, its confidence, its past accuracy, its past average confidence, and the past reliability of its confidence.

We established two main manipulations for this task: the AD's characteristics (matched to the DM or superior to the DM) and the stake level (Low vs High). We formulate these research questions accordingly.

Research Question 1: Adviser influence on DM's opinion
Do DMs assign different weights to the adviser's opinion depending on the stakes involved?

Research Question 2: Optimality of advice integration
Do decision-makers integrate advice more optimally when the stakes are higher?

Research Question 3: Decision-maker's accuracy enhancement
Do DM's accuracy improve more when the stakes are higher?

Research Question 4: Decision-maker's final confidence
Do DMs express higher confidence in their final decisions when the stakes are higher?

Research Question 5: Time taken to integrate advice
Do DMs spend more time integrating advice when the stakes are higher?

Research Question 6: Stakes and adviser type
Are the previous effects modulated by the adviser type?

Research Question 7: Structure of stake variation
Does the effect of stakes on advice integration optimality differ depending on whether stakes vary across trials or are presented in blocks?

2.3 Methods

The experiment is composed of three sections, each with its own task. The first section is meant to evaluate the participant's characteristics, which are then used to tailor the agents for the second section and the main experimental task. Finally, in the third section, the participant's general cognitive ability is evaluated. We use the tasks developed for the previous studies on advice integration (Zonca et al., 2025), but we modify them to study the effect of stakes.

2.3.1 Preregistration

The two experiments on the effect of stakes on advice integration have been pre-registered on OSF:

1. Experiment 1: intermixed design (link: https://osf.io/5sx4b/).

2. Experiment 2: bloc design (link: https://osf.io/xv8ph/).

2.3.2 Experimental task 1: The Individual Decision Task (IDT)

The IDT is our first task. We use it to train participants and collect data on their performance, which will be used to tune the agents' behavior in the main interactive task. It consists of a series of trials in which the participant is required to make a perceptual decision and state their confidence in it. In each trial, the participant is required to estimate the direction of motion of dots appearing on the screen, some of which move coherently in one direction, whereas the others move in random directions (Random Dot Kinematograms, RDK). The *coherence level* of the RDK is the percentage of points moving in the same direction and is changed trial-by-trial to vary sensory uncertainty.

The trial begins with a black circular area appearing at the center of the screen, with a green fixation point in its center (see Fig. 2.1). After 800ms, the dots appear in the black area, moving for 500ms, and then disappear. The fixation point now turns red, and the participant can express her evaluation of the motion direction by clicking the mouse on a point in the black area. The segment joining the fixation point and the clicked point appears as a line that crosses the black circle at its diameter, indicating the direction the participant estimated for the given RDK. After expressing the perceptual decision, the participant is required to express the *confidence* in it, defined as the probability of giving an accurate answer. *Accuracy* in our experiment is measured with the angular distance of the response from the true direction: participants are told that we consider a response accurate within an angle of 20 degrees from the true direction (10° clockwise and 10° counterclockwise). Confidence is expressed as an integer number in percentage points from 0% to 100%, where the probability of a completely random response is 11%. A scale from 0 to 100 appears, with the 11% reference point ticked, and the participant expresses confidence by selecting a point on the confidence line. The trial finishes with the confidence assessment; then, the next trial begins with the black circular area reappearing.

The higher the coherence level, the easier it is to evaluate the motion direction, all else being equal. The coherence levels are adjusted to regulate participants' performance across the three blocks of 30 trials that compose the IDT. In the first block, the coherence levels are set to 30%, 40%, and 50%, and at the end of the first block, the participant's accuracy is computed as the ratio between accurate judgment and total number of trials; if accuracy is below 0.25, the coherence levels are increased to 35%, 45% and 55% for the second block. If the performance is still too low, with the accuracy below the same threshold of 0.25 even after the second block, then a further increase to 40%, 50%, and 60% is done for the third block. The coherence levels for the third block will then be used for the second task (the ADT, see below). Moreover, data about the characteristics of each participant in the second and third blocks of the IDT are used to build the agents for the ADT, tailoring the agents to each individual subject, discarding the first block to avoid potential contamination from learning effects.

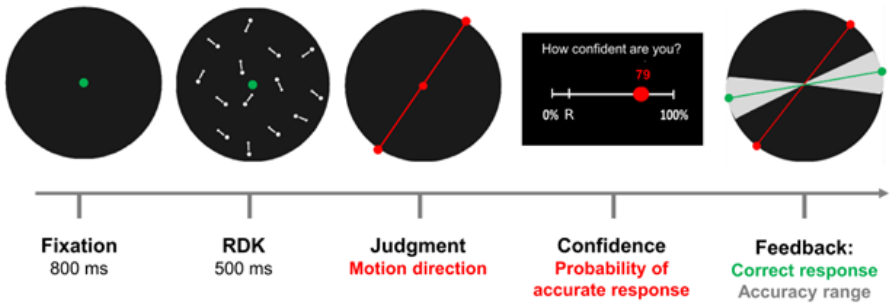

Figure 2.1: The event flow of one trial in the IDT. The black area with the green fixation point appears first, then after 800ms the RDK is projected for 500ms. The fixation point now turns red, and the participant can indicate their judgment of the motion direction on the black circle, which appears as a red segment, and then express their confidence in the judgment. Finally, feedback is given with the correct response in green and a grey area indicating the angular range for the accurate response. Courtesy of Joshua Zonca.

2.3.3 Experimental task 2: The Advised Decision Task (ADT)

The second section of our study adopts an interactive task composed by several blocks (see the section below on "Decision stakes and experimental paradigm") where the DM (the participant) is still initially required to make perceptual decisions and express confidence, like in the IDT, but then interacts with two ADs (artificial assistants), one for each block giving its own assessment of direction and confidence on the same stimuli (see Figure 2.2 for the event flow of a trial). The ADT is the main experimental task and two factors are manipulated: 1) the *Stake level* (low vs. high); 2) the *Adviser type*, that could be an *Alter Ego* (A) of the subject, or an *Enhanced Agent* (E) which is more accurate, with confidence calibrated to its accuracy and more reliable than the subject. In each block, the participant interacts with a specific AD with stable characteristics, assigned for that block, whereas the stakes vary trial-by-trial. Characteristics of the DM and of the AD (mean accuracy, mean confidence, confidence predictivity) are stated at the beginning of each block (see the paragraphs below on the design of the behavior of the agents).

The trials begin with the black circular area and green fixation point on screen for 800ms, then the RDK for 500ms, as in the ADT. The participant expresses the perceptual judgment and confidence in it exactly as in the IDT. After the expression of DM judgment and confidence, the stake for that trial appears on screen, written in light blue or orange, with the association between stake level and color pseudo-randomized across subjects (the colors have been selected to avoid confusion with other colors already present in the ADT). After 500ms, the black circular area is plotted again on screen, with the magnitude of the stake above it, with the DM judgment and with the advice provided by the artificial agent (evaluated motion direction and confidence in it). The participant is now in charge of the final choice, expressed by a mouse click to generate a segment, as usual, and the expression of confidence in the final choice is required on the same 0-100% scale. Finally, feedback with the true motion direction and accuracy ranges is provided, together with the consequence (stake gained or lost). Proceeding to the next trial is initiated by another mouse click, allowed

between 1 and 5 seconds after the feedback, and occurs automatically after 5 seconds have elapsed. The accuracy of the final judgment is used to compute the monetary incentive for that trial.

At the end of each block, the following questions are asked to the participant:

1. Think back to YOUR initial answers (red answers) in the block you just faced. How much do you think was YOUR percentage of accurate initial responses?

2. Think back to the agent's initial responses (blue responses) in the block you just faced. How much do you think was ITS percentage of accurate initial responses?

3. Think back to YOUR confidence ratings in the block you just faced. How much do you think was the average value of YOUR confidence judgments?

4. Think back to the agent's confidence ratings in the block you just faced. How much do you think was the average value of ITS confidence judgments?

5. Think back to all the trials in which YOU have provided an ACCURATE initial judgment. How much was the average value of YOUR confidence judgments in these trials?

6. Think back to all the trials in which YOU have provided an INACCURATE initial judgment. How much was the average value of YOUR confidence judgments in these trials?

7. Think back to all the trials in which the AGENT has provided an ACCURATE initial judgment. How much was the average value of ITS confidence judgments in these trials?

8. Think back to all the trials in which the AGENT has provided an INACCURATE initial judgment. How much was the average value of ITS confidence judgments in these trials?

9. How much did the final decision matter to you when the STAKE was 1/-0.5€, on a scale from 0 to 10?

10. How much did the final decision matter to you when the STAKE was 20/-10€, on a scale from 0 to 10?

11. Imagine playing another experimental block with the same AGENT you encountered in this block. In this hypothetical block, you always gain 1€ for accurate responses and lose 0.5€ for inaccurate ones. The final payment would be the average earnings across all trials. How much would you be willing to pay (in €) for the possibility of receiving, in this hypothetical block, feedback on the agent's response and confidence?

12. Imagine playing another experimental block with the same AGENT you encountered in this block. In this hypothetical block, you always gain 20€ for accurate responses and lose 10€ for inaccurate ones. The final payment would be the average earnings across all trials. How much would you be willing to pay (in €) for the possibility of receiving, in this hypothetical block, feedback on the agent's response and confidence?

Questions 1-8 were previously used in the ADT to collect participants' self-assessments and their assessments of the AD characteristics. We added questions 9-12 to assess participants' perceptions of the stakes.

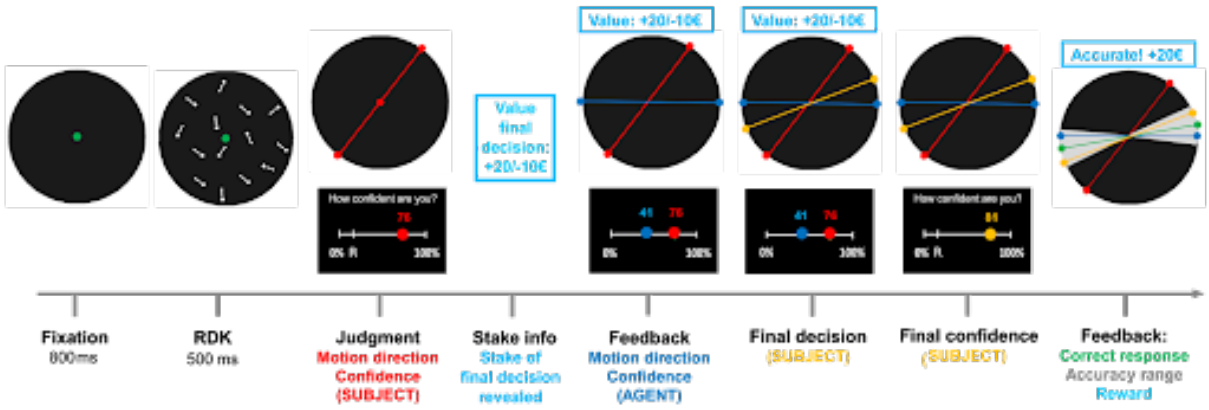

Figure 2.2: The event flow of one trial in the ADT. The black area with the green fixation point appears first, then after 800ms the RDK is projected for 500ms. The fixation point now turns red, and the participant can indicate their judgment of the motion direction on the black circle, which appears as a red segment, and then express their confidence in the judgment. The stake info appears next, along with the AD advice and its confidence, and the DM judgment and its confidence. The DM is in charge of the final decision and expresses the final confidence in it. Finally, feedback is given with the correct response in green and a grey area indicating the angular range for the accurate response. Courtesy of Joshua Zonca.

2.3.4 Experimental task 3: Assessment of Cognitive Ability

The third section of the experiment consists of testing participants' general cognitive ability, under the reasonable assumption that higher intelligence might improve the integration of advice and its components. We use the Raven's Advanced Matrices Test (RAPM III, Raven and Court, 1998) in the short time-limited version of 20 minutes by Hamel and Schmittmann, 2006. It consists of 36 items, each a 3x3 matrix of shapes, ordered according to one or more logical rules; 8 are shown to the participant, and one is missing. The task for the participant is to understand the ordering rules and find the missing shape among 8 possible alternatives. The number of correctly solved items determines the task's incentive.

2.3.5 Decision Stakes: two experimental paradigms

The first factor we manipulate is the *Stake*, a value associated with the outcome of the choice: accurate judgments lead to gains, inaccurate judgments lead to losses, where we intend "accuracy" relative to angular distance from true motion direction as defined above. We selected two stake levels: high and low. In high-stakes trials, the gain is 20€, and the loss is 10€; in low-stakes trials, the gain is 1€, and the loss is 0.5€. Stakes are balanced across

agent type and coherence levels. The overall payment will take into account all ADT trials (see the section on monetary incentives below). We designed two experiments. In Experiment 1, stakes vary trial-by-trial (experiment *stakes-intermixed*; link to OSF preregistration: https://osf.io/5sx4b/). In Experiment 2, stakes vary between blocks (experiment *stakes-blocks*; link to OSF preregistration: https://osf.io/xv8ph/).

Given the trial-by-trial randomization of the stake level in Experiment 1, participants might learn a strategy for high-stakes trials that also applies to low-stakes trials. For instance, they might be closer to optimality in integrating advice in high-stakes trials, and maintain the same level of integration also in low-stakes trials. To account for the possible presence of *spillover effects*, the first 18 trials in each ADT block will have a fixed low stake level, and we will treat them as a baseline. The ADT of Experiment 1 has 2 blocks of 78 trials. We manipulate the Stake level (low vs high) and the Adviser type (Alter Ego vs Enhanced Agent), with the two blocks corresponding to the agent used in each.

Experiment 2 was designed based on preliminary findings from Experiment 1, in which participants tended not to adapt from one trial to another but did so over longer sequences of trials. Moreover, findings suggested the potential presence of *order effects*, with optimality of advice integration varying depending on which stake level was encountered first. To account for these longer behavioral tendencies, we decided to design an experiment in which the stakes vary across blocks. This version includes an *Order* factor with two levels (low-stake first vs. high-stake first), and we use it to test interactions with the main Stake factor in the primary analyses for RQ1-RQ6 (see the section on secondary analyses below). The ADT of Experiment 2 has 4 blocks of 39 trials. We manipulate the Stake level (low vs high) and the Adviser type (Alter Ego vs Enhanced Agent) in a 2x2 design: The first two blocks are always with the same Stakes level, whereas the last two blocks have the other Stakes level; the order is counterbalanced (LLHH or HHLL). For each Stakes level, one block is played with the Alter Ego and one block with the Enhanced agent, with counterbalanced order (AEAE or EAEA).

2.3.6 Designing the behavior of the agents

Data from the IDT are used to estimate their behavior on three dimensions: *subject mean accuracy* (sMA: the proportion of accurate trials), *subject mean confidence* (the mean of the expressed confidence ratings), and *subject confidence predictivity* (sCP). In particular, sCR is expressed by two values: the point-biserial correlation between accuracy and confidence, trial-by-trial, and the Pearson correlation between confidence and absolute angular error, trial-by-trial. These data are used to build the advisers for the ADT, and they are modelled according to the characteristics of each individual subject (i.e. each participant will have their own agent, with specific characteristics, generally differing between participants).

The second factor we manipulate is the *Adviser type*. Analogously to human participants, an artificial agent can be described by three characteristics: the *agent mean accuracy* (aMA: the proportion of accurate judgments), the *agent mean confidence* (aMC: the mean of its confidence ratings) and the *agent confidence predictivity* (aCP: the correlation between accuracies and confidence ratings). These dimensions are used to build the two agents used in our experiments (see Table 2.1):

1. Alter Ego: adviser whose characteristics are matched to the subject it is interacting with.
2. Enhanced Agent: adviser with superior accuracy, calibrated confidence (equal to its mean accuracy) and high confidence predictivity (selected at r=0.5, to be compared to the mean r=.22 of human subjects in our Advice Taking 2 experiment: Zonca et al., 2025).

Table 2.1: The characteristics of the two advisers used in our experiments with stakes.

Characteristics	**Alter Ego (AE)**	**Enhanced Agent (EA)**
Mean Accuracy	=	+20%
Mean Confidence	=	= Agent mean accuracy
Confidence Predictivity	=	r = 0.5

2.3.7 Generating the behavior of the agents

The behavior of the agents is described by their aMA, aMC, aCR, and these values are built by generating distributions with specific properties. Basically, to build an agent, we generate distributions of perceptual judgments (and consequently accuracies) and confidence ratings, and then we sample from them based on a desired confidence predictivity. The set of judgments and confidences respecting the specified aCR becomes the agent assigned to the subject for the block in execution. The aim is to build agents whose behavior is as close as possible to the desired agent as described in Table 1. Two ordered vectors of absolute angular errors and of confidence ratings are initially generated, and a subset of these, equal to the number of trials for a block, is selected iteratively: for each iteration, mean accuracy, mean confidence and confidence predictivity are computed for an extracted subset and the values are compared to the target aMA, aMC, aCR for the agent type to be built for the block, with margins of tolerance within the shuffling stops and the subset is selected. For the first 15000 iterations, the tolerances for accuracy, the correlation between confidence and accuracy, and the correlation between confidence and absolute angular error are set to 0.025, and the tolerances for the mean confidence and the SD of angular error are set to 2.5. If these margins are not enough, after 30000 iterations, tolerances are increased to 0.05 and 5; after 20000 iterations, tolerances are increased to 0.1 and 10; after 45000 iterations, tolerances are increased to 0.2 and 20. Once a subset is selected, a sign (positive or negative) is randomly assigned to angular errors, which are then used to build the perceptual judgments and confidence ratings the agent uses as advice during the trials.

We wanted the agents' perceptual errors to be similar to human behavior by correlating sensory uncertainty with perceptual accuracy. To do so, correlation with the coherence of the RDKs was obtained by associating the two vectors of perceptual judgments and confidence ratings with a coherence vector containing the three coherence values for the RDKs. The coherence vector was shuffled until a correlation of $r = 0.15 \pm 0.03$ with the perceptual error vectors was found, with the correlation coefficient corresponding to the average correlation among human participants in a pilot study.

2.3.8 Disclosure of behavior

Following what was already done for the Advice Taking experiment 2 (Zonca et al., 2025), at the beginning of each block, we disclose to the participant details about their own behavior and the agent they will encounter in that block (Figure 2.3). We display on screen their respective mean accuracy, mean confidence, and confidence predictivity. This third piece of information is presented in two formats: numerically as the mean confidence in accurate and inaccurate choices, and graphically as the overlap between the distributions of mean confidences, with larger distances indicating higher predictive confidence. Data for the human participant are drawn from the second and third blocks of the IDT. Data for the agent's accuracy and confidence are computed from the subset of trials selected to build the agent for that block and the human participant will encounter; its confidence predictivity is estimated by simulating 1000 trials with the agent's assigned behavioral parameters.

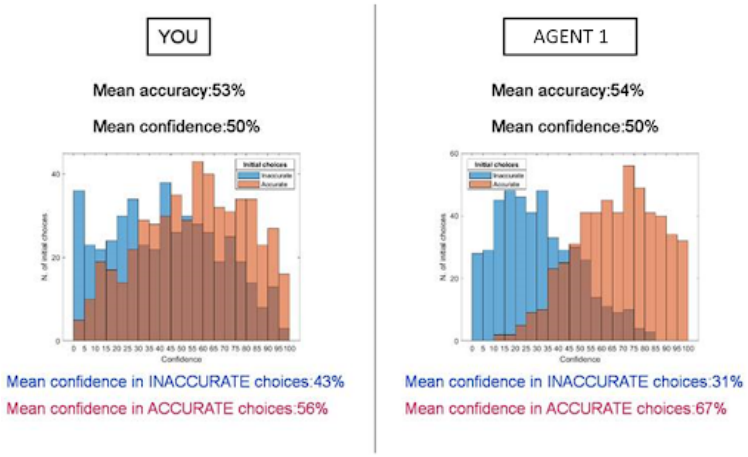

Figure 2.3: Graphical display used at the beginning of each block to disclose the characteristics of the human participant and the adviser of the block. Courtesy of Joshua Zonca.

2.3.9 Monetary incentives

Reimbursement to participants is determined by their performance on the three experimental tasks. In the IDT, participants are rewarded according to:

1. accuracy: reward is a linear function of accuracy (% correct on all trials), converted in euros in the range [0; 5]€, with [0.1; 0.7] as extremes of sMA

2. confidence predictivity: we convert the correlation between accuracy and confidence ratings in euros in the range [0; 5]€, with [0; 0.5] as extremes of sCR

In the ADT, rewards vary depending on performance and of the stake condition. All trials contribute to the payment, with a weighting coefficient determined by the stake value for each trial: the reward is computed as the average reward for high-stakes trials, summed with

the average reward for low-stakes trials. In the RAPM, the reward is proportional to the number of correct responses converted into euros in the range [0;3]€. The final payment is the sum of all rewards plus a show-up fee of 4€, rounded up to the next integer, up to a maximum of 25€. In the event of a negative payoff due to very low performance, the show-up fee guarantees a minimum reimbursement.

2.3.10 Sample size and Power Analysis

Our considerations on sample size follow what was already done for the Advice Taking experiment 2 (OSF preregistration, see section 2.3.1): data from the experiment were used to estimate realistic effect sizes. The *SimR* package was used to run a simulation-based power analysis, with 2 agents instead of the 7 used in previous studies on advice-taking. We use experimental data from 43 participants doing 54 trials for each adviser, with Model 2 used in the simulation (see the section on Analyses). To run the simulation, a dummy variable for *Stakes* is generated and treated as the main effect of interest, with 140 simulated trials. The standardized effect is set as $\beta = 0.15$ (unstandardized: B=0.05), representing a 5% difference in optimality between the two *Stakes* conditions. The two-tailed p-value criterion was set at $\alpha = 0.05$, the statistical power at 0.9, with 10000 iterations. With these values set, the minimum required sample size from our power analysis is 43 participants, achieving 100% power. A slightly lower standardized effect of $\beta = 0.10$, all other parameters being equal, reaches 98% statistical power with the same sample size.

2.3.11 Data exclusion

For these experiments, we adopted the same data exclusion protocol of the previous experiments on advice-taking (Zonca et al., 2025).
Exclusion of participants:

- Participants who did not complete the experiment.
- Initial and final mean accuracy in the ADT. We set a minimum mean accuracy of 0.15 for initial and final decisions in the ADT to consider the participant's data. We plan to discard data of participants who perform below this threshold in both the performance parameters (initial and final decisions). Indeed, we consider lower levels of accuracy for both parameters as a strong indicator of random behavior (chance level: 0.11).

Exclusion of trials: As described in the section "Derived variables", influence and optimal influence variables will not be computed in these cases

- Distance between the two initial responses $< 5^\circ$
- Final/Optimal response outside the minimum angle (α) subtended by the two initial responses, with a tolerance of 10°.

2.4 Analyses

2.4.1 Primary measured variables from the ADT

Judgment accuracy: Binary variable taking value 1 if the response falls within an angle of $\pm$ 10° from the true motion direction (accurate response), and 0 otherwise (inaccurate response).

Judgment angular error: Continuous variable taking as value the absolute angular distance between observed response and true motion direction.

Confidence: Estimated probability of having given an accurate response in a trial, expressed with continuous values from 0 to 100, as evaluated by the participant or the agent.

Final decision accuracy: Binary variable describing accuracy of the final judgment: it takes value 1 if the response falls within an angle of $\pm$10° from the true motion direction (accurate response), and 0 otherwise (inaccurate response).

Final decision confidence: Estimated probability of having given an accurate final response in a trial, expressed with continuous values from 0 to 100, as evaluated by the participant or the agent.

Final decision RT: Continuous variable given by reaction times for the final decision.

2.4.2 Derived variables

Influence: For each trial, the measures above are used to quantify the influence by the agent on the decision maker on each trial. We define influence (I) as the angular shift (s) of the DM final response (fin) from the DM initial response (sbj) towards the agent's advice (ag), divided by the angular distance α between DM initial response and agent advice, considering the narrowest of the angles between them (see Figure 2.4) and a tolerance γ=10°. Operatively:

$$I = \frac{s(\text{fin}\,,\ \text{sbj})}{d(\text{sbj}\,,\ \text{ag})}$$

If $\alpha < 5°$, then the response is discarded and the influence is not computed, since such angular distances make the initial response and the advice nearly indiscernible visually to the participant. If the final response is outside α and outside the margin of tolerance γ, that final response is discarded. If the final response falls outside α but within the tolerance γ, then the influence is assigned 0 if the final response is closer to the initial response, or it is assigned 1 if it is closer to the agent's advice. A graphical illustration is rendered in Figure 2.4.

Miscalibration: We define calibration as the match between confidence and accuracy of a perceptual judgment. A mismatch between the two is miscalibration. Numerically, it is computed by subtracting the participants' average accuracy from their average confidence,

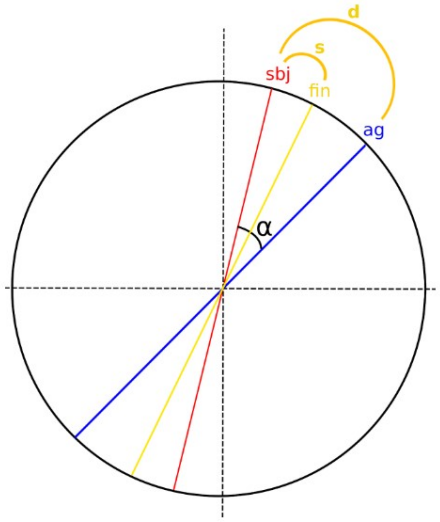

Figure 2.4: Definition of influence of the adviser on the decision maker in a trial. Courtesy of Joshua Zonca.

across all trials, yielding an individual index of confidence calibration. Positive values indicate overconfidence; negative values indicate underconfidence.

Miscalibration (final decisions): Miscalibration index, computed by considering only final decision confidence and final decision accuracy. Positive values indicate overconfidence; negative values indicate underconfidence.

Individual confidence predictivity: The correlation between the trial-by-trial participant's confidence ratings and judgment accuracy.

Individual confidence predictivity (final decisions): The correlation between the trial-by-trial participant's confidence ratings and the accuracy of the final decisions.

Δ mean accuracy: The difference between the agent's and participant's mean accuracy.

Δ mean confidence: The difference between the agent's and participant's mean confidence.

Δ confidence predictivity: The difference between the agent's confidence predictivity (point-biserial correlation between trial-by-trial confidence and accuracy) and the participant's confidence predictivity.

Augmentation index: Mean difference, over all trials, between the accuracy of the human final decision, taken after interacting with the adviser, and the accuracy of the initial decision taken alone, before the interaction.

Synergy index: Mean difference, over all trials, between the accuracy of the human final

decision and the accuracy of the more accurate adviser.

2.4.3 Bayesian model of optimality

Bayesian network

We define *optimality* as the behavior that maximises the probability of an accurate response given all the information available at the moment of the decision. In the ADT, the participant's final decision for each trial is the observed behavior to be analyzed and compared with a model of optimal behavior. We formalize optimal behavior with a Bayesian model that, given all available information in a trial and the characteristics of the human participant and the adviser, outputs the stimulus value that maximizes the likelihood of the observed response. From this, the optimal influence for each trial can be derived. The model is a directed acyclic graph built with GeNIe ("GeNIe Modeler", n.d.). Variables are represented with nodes, and probabilistic relationships between variables are represented with arcs. At its core, it is a probabilistic signal-detection model in which the stimulus direction is the signal, and the perceptual error, with its distribution, is the noise (the structure is shown in Fig. 2.5). The core has been expanded to include confidence distributions. The participant and the agent are the two detectors who estimate the signal and are characterized by the three dimensions of mean accuracy, mean confidence, and their correlation, confidence predictivity. The following nodes compose the model:

1. Individual properties: mean accuracy, mean confidence, confidence predictivity of the human participant, and of the adviser.

2. Evidence: information known to the participant before the final decision, which consists of the initial perceptual judgment of the participant, the confidence estimate, the perceptual judgment by the agent, and the agent's confidence in the judgment.

3. Objective properties of the stimulus: coherence of the moving dots and their direction of motion

4. Error: distributions of perceptual angular errors for the participant and for the adviser. They depend on the accuracy of the participant/adviser and on task difficulty. Together with the stimulus's objective properties, they build the distributions of participants' and the adviser's judgments. Together with mean confidence and confidence predictivity, they determine the distributions of confidence of the participant and of the adviser.

The objective properties of the stimulus in each trial are unknown to the participant until the end of feedback. Indeed, the task consists of expressing the direction of motion that would maximise the evidence given all the individual properties.

The distributions of perceptual errors depend on mean accuracy and task difficulty, for both the participant and the adviser. The distribution of mean accuracy is organized in N bands, whereas the distribution of task difficulty is given by the three levels of motion coherence. Before the stimulus's appearance, all judgments are equiprobable, and the distribution of the signal (the moving dots in the RDK) is uniform; also, priors of mean accuracy

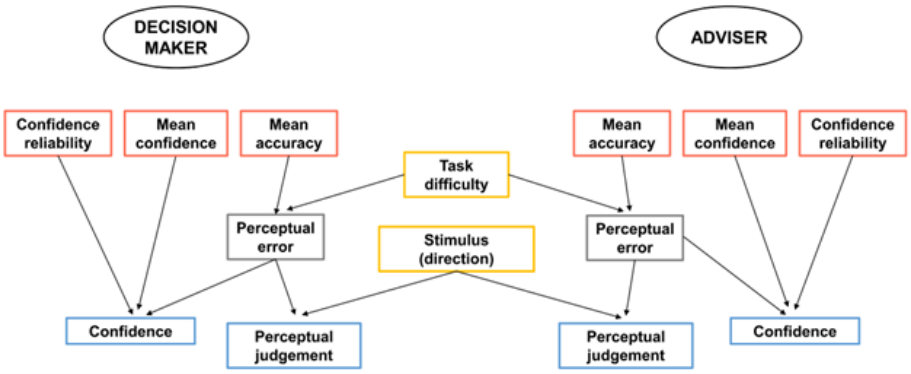

Figure 2.5: Formal Bayesian model of optimality. Rectangles represent nodes (variables), whereas arrows are arcs (relations between variables). Red rectangles represent individual properties; blue rectangles represent available evidence; yellow rectangles represent objective properties of the stimulus; grey rectangles represent error distributions. Courtesy of Joshua Zonca.

and of task difficulty are uniformly distributed, representing maximal uncertainty of the characteristics of the stimulus and of the accuracies of the participant and of the adviser. The prior distributions of perceptual errors will be estimated by fitting experimental data to the model. After the signal appears, its value is added to the noise distributions, generating the probability distributions for the perceptual judgments of the detectors. Given these two distributions and the noise distributions, the model allows the inverse distribution of objective stimulus directions to be obtained, thereby maximising the probability of observing the two perceptual responses. The individual characteristics are communicated in advance to the participant, and their nodes remain set to the values presented at the beginning of each block. Consequently, the model does not need to update them based on the feedback. The distributions of confidence are generated by the estimate of perceptual error for each trial, by metacognitive ability to estimate confidence which is correlated to the errors (confidence predictivity) and by a tendency to give lower or higher confidence estimates on average (mean confidence node), which will remain set, as they are communicated to the participant at the beginning of each block.

Model training

The arcs of the model represent probabilistic relations between the variables, and they are estimated by fitting experimental data to the model. We used the parameter space from the simulated agents from the experiment "Advice Taking 1" (https://osf.io/myg3c), which includes a larger number of agents as compared to the ones used here. Further analyses in the future will consider a more limited parameter space, more refined to reflect only the two agents used in this study. Final decisions by the participants are not included in the training data because the aim is to estimate the optimal behavior given all information available before the final decision happens. Nodes for individual properties represent the distribution of human participants and artificial agents relatively to their mean accuracy, mean confidence and confidence predictivity, discretized in quantiles; between-participant variability and the objective properties of the stimulus (Difficulty and Direction) determine the probability

distribution of the hierarchically dependent nodes (Error and Evidence) based on the causal relationships between them as described in Figure 5. These probability distributions are derived either from actual human-generated data or from simulated data to extend the parameter space.

Optimality in the model
After training, the model can be used to estimate the optimal response for each trial, according to the following steps:

1. For each trial, the distributions of Perceptual error and Objective stimulus properties start as noninformative priors

2. Evidence data for the trial and Individual properties are fed to the model, and the posterior distribution for perceptual error and objective properties are estimated

3. The estimated probability distribution of the Stimulus provides the most probable angular direction of the perceptual stimulus that might have generated the Evidence, given all the information available to the human participant. This angular direction is defined as the optimal response

4. Feedback on the true direction is provided to the model

5. After the feedback has been received, the model updates the probability distribution of the Individual properties nodes for the participant; Individual properties for the agent instead remain fixed for all the trials of the block with that agent (matching the values disclosed to the participant). These distributions will be used as prior for the next trial.

Optimality and observed use of advice
The model provides estimates of optimal response for each trial. Optimal influence is then derived analogously to the observed influence for the trial, but using the model response rather than the participant's observed response. Once optimal influence for a trial is obtained, we can compute a measure expressing the distance between optimal decision and observed decision, called *optimality index* for that trial:

$$O = 1 - |I_{\text{OPT}} - I_{\text{OBS}}|$$

Values tending to 0 indicate increasing suboptimality, whereas values tending to 1 indicate optimality (where the difference between optimal and observed influence would be low or negligible). Finally, we can derive an individual optimality index as a measure of the average (sub)optimality level for each subject. This index is computed as the mean of the subject's optimality indices across all trials.

2.4.4 Models of the Research Questions

***Research Question 1**: Adviser influence on DM's opinion*
RQ1: Do DMs assign different weights to the adviser's opinion depending on the stakes

involved? A linear mixed model is estimated, with *Influence* (I) trial-by-trial as dependent variable, with *Stake* (S: high vs. low) as independent factor, and a random intercept (μ_0) at participant level.
Model 1: $I = \beta_0 + \beta_1 S + \mu_0$

***Research Question 2**: Optimality of advice integration*
RQ2: Do decision-makers integrate advice more optimally under high-stakes conditions? A linear mixed model is estimated, with *Optimality index* trial-by-trial as dependent variable, with *Stake* (high vs. low) as independent factor, and a random intercept (μ_0) at participant level.
Model 2: $O = \beta_0 + \beta_1 S + \mu_0$

***Research Question 3**: Decision-maker's accuracy enhancement*
RQ3: Do DM's accuracy improve more (after advice) under high-stakes conditions? A logistic mixed model is estimated, with trial-by-trial *Accuracy* (A) as the dependent variable, with *Stakes* (S: Low, High), *Decision* (D: Initial, Final), and their interaction as independent variables, and a random intercept (μ_0) at participant level.
Model 3: $A = \beta_0 + \beta_1 S + \beta_2 D + \beta_3 S \cdot D + \mu_0$

***Research Question 4**: Decision-maker's final confidence*
RQ4: Do DMs express higher confidence in their final decisions under high-stakes conditions? A linear mixed model is estimated, with *final Confidence* (fC) trial-by-trial as dependent variable, with *Stake* (S: high vs. low) as independent factor, and a random intercept (μ_0) at participant level.
Model 4: $fC = \beta_0 + \beta_1 S + \mu_0$

***Research Question 5**: Time taken to integrate advice*
RQ5: Do DMs spend more time integrating advice when the stakes are higher? A linear mixed model is estimated, with trial-by-trial *Final decision RTs* ($fRTs$) as the dependent variable, with *Stakes* (S: Low, High) as the independent factor, and a random intercept (μ_0) at the participant level.
Model 5: $fRTs = \beta_0 + \beta_1 S + \mu_0$

***Research Question 6**: Stakes and adviser type*
RQ6: Are the previous effects modulated by the adviser type? Models from 1 to 5 above will be extended by adding *Adviser* (AD: Alter Ego, Enhanced Agent), and their interaction with *Stakes* (S: Low, High) as independent variables.
Model 6: $Y_{1\text{-}5} = \beta_0 + \beta_1 S + \beta_2 AD + \beta_3 S \cdot AD + \mu_0$

***Research Question 7**: Optimality and structure of stake variation*
RQ7: Does the effect of Stakes on the optimality of advice integration depend on the structure of the stakes variation (block vs. intermixed)? Results from Experiment 1 and Experiment 2 will be compared by extending the Models 1 to 5 above with *Treatment* (T: Exp.1 vs Exp.2) and its interaction with *Stakes* (S: Low, High) as independent variables.
Model 7: $Y_{1\text{-}5} = \beta_0 + \beta_1 S + \beta_2 T + \beta_3 S \cdot T + \mu_0$

***Secondary analyses**: augmentation and synergy*
Following Vaccaro et al., 2024, we consider two possible ways in which interactions with advisers might improve the performance of human decision makers. In *Human augmentation*, adopting advice improves the performance of humans alone, but not to the point of surpassing the performance of the best of the two team members taken alone (which, in our study, is the Enhanced Agent, always more accurate than its user). In *Human-AI synergy*, instead, performance of the team is superior to each of the members taken alone. We compute the augmentation index and mean synergy index for each subject, and their means for our sample. Positive values of augmentation indicate that the final accuracy is on average improved by the interaction with the adviser. Positive values of synergy indicate that performance of the human-agent team is superior to each member taken alone (and that synergy is reached), whereas negative values of synergy indicate that the accuracy of the Enhanced Agents are generally still superior to the accuracy of the human-agent team in final decisions (and consequently that synergy is not reached). A Wilcoxon signed/rank test was used to analyze if the distributions of augmentation and synergy indexes were statistically different from 0.

***Secondary analyses**: fluid intelligence in advice integration*
We hypothesize that higher fluid intelligence (FI) predicts better advice integration: observed behavior should be closer to optimality for participants with higher Raven scores than for those with lower Raven scores. We test the role of Stakes, adviser, Fluid intelligence, and their interactions in modulating optimality in the ADT in a linear mixed model with a random intercept at the participant level. We also explore analogous effects on Final accuracy and Reaction Times.
Model 9a: $O = \beta_0 + \beta_1 S + \beta_2 AD + \beta_3 FI + \beta_4 S \cdot AD + \beta_5 AD \cdot FI + \beta_6 S \cdot FI + \beta_7 S \cdot AD \cdot FI + \mu_0$
Model 9b: $fA = \beta_0 + \beta_1 S + \beta_2 AD + \beta_3 FI + \beta_4 S \cdot AD + \beta_5 AD \cdot FI + \beta_6 S \cdot FI + \beta_7 S \cdot AD \cdot FI + \mu_0$
Model 9a: $fRTs = \beta_0 + \beta_1 S + \beta_2 AD + \beta_3 FI + \beta_4 S \cdot AD + \beta_5 AD \cdot FI + \beta_6 S \cdot FI + \beta_7 S \cdot AD \cdot FI + \mu_0$

2.5 Results: Intermixed Design

For our analyses, we report results for our RQs for each of our main dependent variables: influence, optimality, accuracy, final confidence, reaction times, and egocentric bias. For each of them, we report the model with the effect of Stakes and the extended model from RQ6, which also considers the effect of Agent type. We present the results of Experiment 1 first, followed by those of Experiment 2.

2.5.1 Descriptives

Descriptives for our sample are reported in Table 2.2. One subject was excluded from the analysis, according to our exclusion criteria, due to both initial and final accuracy below the chosen threshold of 0.15.

Table 2.2: Descriptive statistics for our sample.

Descriptive statistic	Value
Sample size	44
Gender	27 female; 15 male; 2 "non-binary"
Age	Mean = 22.27, SD = 2.45

2.5.2 Influence

Model 1 considered the effect of *Stake* on *Influence*. Results from Model 1 show that the fixed effect of *Stake* on *Influence* is not significant: $F(1, 4435) = 0.069, p = 0.793$. The $ICC = 0.097$ indicates that about 9.7% of the total variance is due to between-subject variability. Participants adjusted their estimate towards the agent's estimate with an influence of $I = 0.414$ independently of the *Stake* level (mean influence with low stakes $I = 0.412$, mean influence in high stakes $I = 0.416$).

Results from Model 6a considered the effect of *Stake* and *Agent type* on *Influence*. The fixed effect of *Stake* is not significant $[F(1, 4434) = 0.064, p = 0.800]$. The fixed effect of *Agent type* is significant $[F(1, 4434) = 212.461, p < .001]$. The interaction of the two is not significant $[F(1, 4434) = 3.464, p = 0.063]$. The $ICC = 0.10$ indicates that about 10% of the total variance is due to between-subject variability. Participants adjusted their estimate towards the agent's estimate more with the Enhanced Agent ($I = 0.485$) than with the Alter Ego ($I = 0.347$), independently of the *Stake* level (with Alter Ego in low stakes: $I = 0.336$; with Alter Ego in high stakes: $I = 0.358$; with Enhanced Agent in low stakes: $I = 0.493$; with Enhanced Agent in high stakes: $I = 0.478$). Results for Model 6a are reported below in Figure 2.6.

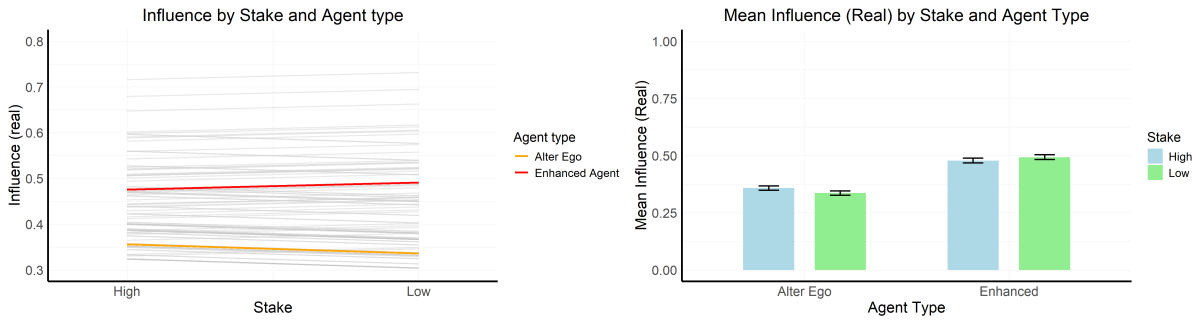

Figure 2.6: Results from the Model 6a for the effect of *Stake* and *Agent type* on *Influence*. On the left: linear mixed model with fixed effects and random intercepts at participant level. On the right: histogram for mean *Influence* depending on the main effects.

2.5.3 Optimality

Model 2 considered the effect of *Stake* on *Optimality*. Results from Model 2 show that the fixed effect of *Stake* on *Optimality* is not significant: $F(1, 4396) = 3.511, p = 0.06$. The $ICC = 0.043$ indicates that about 4.3% of the total variance is due to between-subject variability. The mean optimality of participants was $O = 0.681$, slightly higher in high-stakes trials ($O = 0.689$) than in low-stakes trials ($O = 0.673$).

Results from Model 6b considered the effect of *Stake* and *Agent type* on Optimality. The fixed effect of *Stake* is not significant $[F(1, 4394) = 3.373, p = 0.066]$. The fixed effect of *Agent type* is significant $[F(1, 4393) = 20.421, p =< .001]$. The interaction of the two is not significant $[F(1, 4395) = 3.201, p = 0.073]$. The $ICC = 0.043$ indicates that about 4.3% of the total variance is due to between-subject variability. Mean optimality was marginally higher with the Enhanced Agent ($O = 0.699$) than with the Alter Ego ($O = 0.663$), independently of the *Stake* level (with Alter Ego in low stakes: $O = 0.648$; with Alter Ego in high stakes: $O = 0.678$; with Enhanced Agent in low stakes: $O = 0.663$; with Enhanced Agent in high stakes: $O = 0.701$). Results for Model 6b are reported below in Figure 2.7.

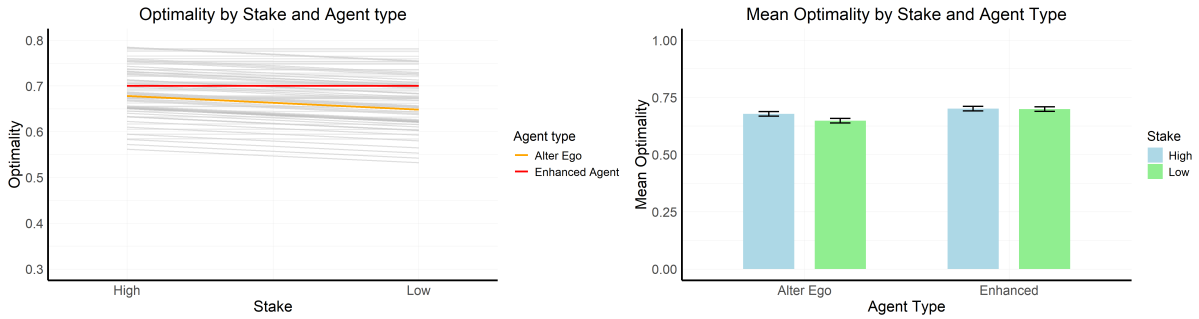

Figure 2.7: Results from the Model 6b for the effect of *Stake* and *Agent type* on *Optimality*. On the left: linear mixed model with fixed effects and random intercepts at participant level. On the right: histogram for mean *Optimality* depending on the main effects.

2.5.4 Accuracy

Model 3 considered the effect of *Stake* on *Accuracy*, considering both the initial and the final decision with the *Decision* factor. Results from Model 3 show that the fixed effect of *Stake* on *Accuracy* is not significant: $z = 1.023, p = 0.306$. The effect of *Decision* is significant: $z = 9.814, p < 0.001$. The interaction between the two is not significant: $z = -0.479, p = 0.632$. The $ICC = 0.269$ indicates that about 27% of the total variance is due to between-subject variability. Mean accuracy of the initial decisions was $iA = 0.370$ for low stakes and $iA = 0.357$ for high stakes. Mean accuracy of the final decisions was $fA = 0.500$ for low stakes and $fA = 0.496$ for high stakes.

Results from Model 6c considered the effect of *Stake*, *Decision*, and *Agent type* on *Accuracy*. The fixed effect of *Stake* is not significant $[z = 0.260, p = 0.794]$. The fixed effect of *Decision* is significant $[z = 3.137, p < 0.01]$. The fixed effect of *Agent type* is significant $[z = -2.675, p < .01]$. Considering interactions, *Stake* by *Decision* is not significant

$[z = -0.164, p = 0.869]$; *Decision* by *Agent type* is significant $[z = 5.638, p < 0.01]$; *Stake* by *Agent type* is not significant $[z = 0.705, p = 0.481]$; *Stake* by *Decision* by *Agent type* is not significant $[z = -0.276, p = 0.782]$. The $ICC = 0.272$ indicates that about 27% of the total variance is due to between-subject variability. Mean initial accuracy was comparable with the two agents, as expected for the initial decision since it happens before encountering the agent (with Alter Ego, $iA = 0.386$, with Enhanced Agent, $iA = 0.340$). Mean final accuracy was higher with the Enhanced Agent ($fA = 0.554$) than with the Alter Ego ($fA = 0.445$). Results for Model 6c are reported below in Figure 2.8.

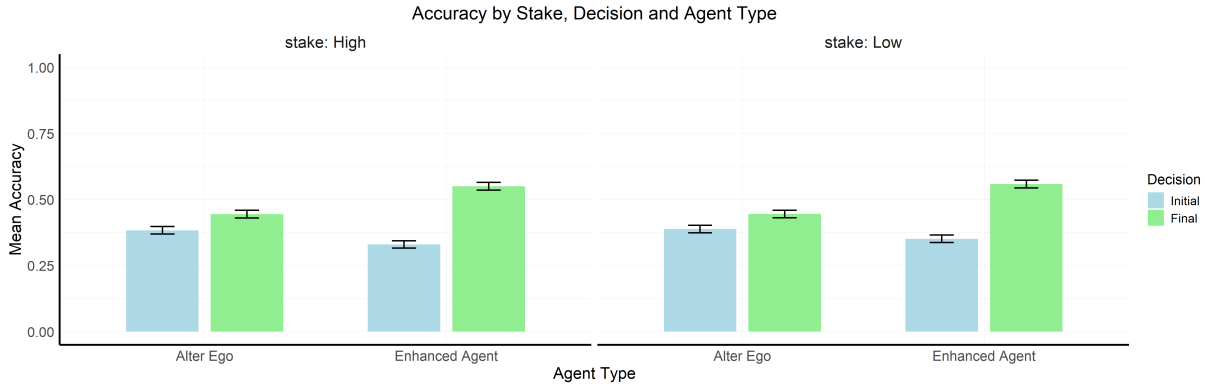

Figure 2.8: Results from the Model 6c for the effect of *Stake*, *Decision* and *Agent type* on *Accuracy*.

2.5.5 Confidence

Model 4 considered the effect of *Stake* on *Final Confidence*. Results from Model 4 show that the fixed effect of *Stake* on *Final Confidence* is significant: $F(1, 4563) = 12.095, p < .001$. The $ICC = 0.428$ indicates that about 42.8% of the total variance is due to between-subject variability. Participants' final confidence was slightly higher in low-stakes ($fC = 58.00$) than in high-stakes ($fC = 56.03$).

Results from Model 6d considered the effect of *Stake* and *Agent type* on *Final Confidence*. The fixed effect of *Stake* is significant $[F(1, 4561) = 12.449, p < .004]$. The fixed effect of *Agent type* is not significant $[F(1, 4561) = 0.002, p = 0.965]$. The interaction of the two is not significant $[F(1, 4561) = 3.206, p = 0.073]$. The $ICC = 0.423$ indicates that about 42.3% of the total variance is due to between-subject variability. Participants expressed higher final confidence in low stakes ($fC = 58.00$) than in high stakes ($fC = 56.03$), and with the Alter Ego ($fC = 57.05$) than with the Enhanced Agent ($fC = 56.98$). The difference between the two was similar independently of the *stake* level: with Low stakes: $fC(AE) = 57.60, fC(EA) = 58.43$; with High stakes: $fC(AE) = 57.59, fC(EA) = 58.43$. Results for Model 6d are reported in Figure 2.9 below.

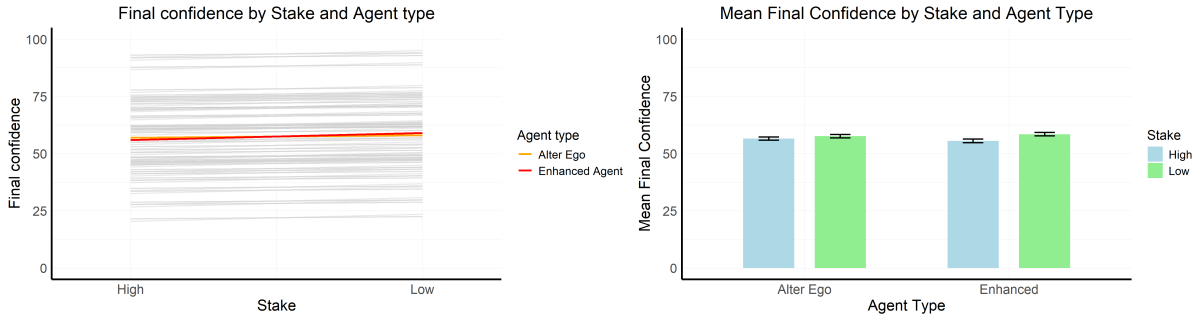

Figure 2.9: Results from the Model 6d for the effect of *Stake* and *Agent type* on *Final Confidence*. On the left: linear mixed model with fixed effects and random intercepts at participant level. On the right: histogram for mean *Final Confidence* depending on the main effects.

2.5.6 Reaction Times

Model 5 considered the effect of *Stake* on *Reaction Times* of the final decision. Results from Model 5 show that the fixed effect of *Stake* on *Reaction Times* is significant: $F(1, 4564) = 100.68, p < .001$. The $ICC = 0.159$ indicates that about 15.9% of the total variance is due to between-subject variability. Mean RTs in high-stakes trials are significantly higher ($fRT = 4.31s$) than in low-stakes trials ($fRT = 3.56s$), with an average gap of 744ms between the two conditions.

Results from Model 6e considered the effect of *Stake* and *Agent type* on *Reaction Times* of the final decision. The fixed effect of *Stake* is significant $[F(1, 4565) = 101.079, p < .001]$. The fixed effect of *Agent type* is significant $[F(1, 4565) = 18.202, p < .001]$. The interaction of the two is not significant $[F(1, 4565) = 0.022, p = 0.881]$. The $ICC = 0.156$ indicates that about 15.6% of the total variance is due to between-subject variability. Mean RTs with the Enhanced Agent were lower ($fRT = 4.103s$) than with the Alter Ego ($fRT = 3.779s$). As evidenced by the non-significant interaction, the differences between RTs in low stakes vs. high stakes were only marginally higher with the Alter Ego as agent: with Alter Ego: $fRT(HS) = 4.15, fRT(LS) = 3.41s, \Delta RT = 736ms$; with Enhanced Agent: $fRT(HS) = 4.48, fRT(LS) = 3.73s, \Delta RT = 751ms$. Results for Model 6e are reported in Figure 2.10 below.

2.5.7 Augmentation and synergy

Our data show that Human augmentation has been reached with both agents. Synergy was reached only with the Alter Ego as adviser. Results, including descriptives and the Wilcoxon signed-rank test, are reported in Table 2.3. Given that the maximum synergy index is positive, some human-agent combinations in our sample achieved Human-AI synergy.

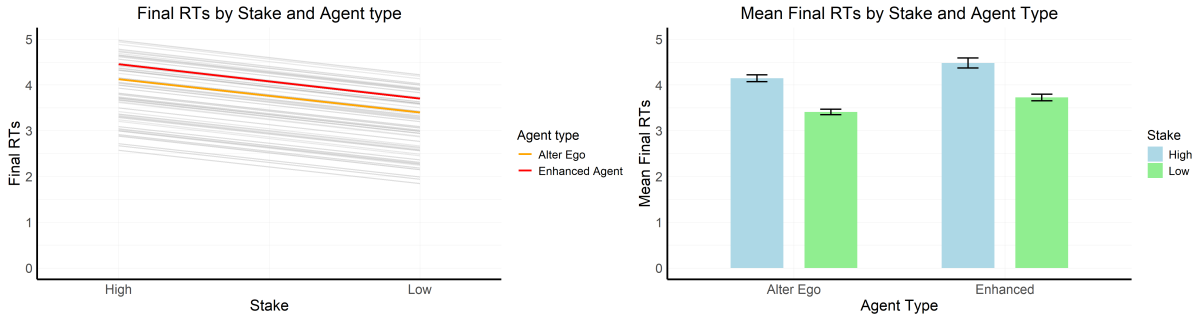

Figure 2.10: Results from the Model 6e for the effect of *Stake* and *Agent type* on *Reaction Times*. On the left: linear mixed model with fixed effects and random intercepts at participant level. On the right: histogram for mean *Reaction Times* depending on the main effects.

Table 2.3: Results for augmentation and synergy in our sample.

Adviser	**Augmentation Index**	**Synergy Index**
Alter Ego	$M(Aug) = 0.061$	$M(Syn) = 0.105$
	$SD(Aug) = 0.080$	$SD(Syn) = 0.102$
	$min(Aug) = -0.115$	$min(Syn) = -0.128$
	$max(Aug) = 0.275$	$max(Syn) = 0.276$
	$U = 738, p < .001$	$U = 886, p < .001$
Enhanced Agent	$M(Aug) = 0.215$	$M(Syn) = 0.013$
	$SD(Aug) = 0.086$	$SD(Syn) = 0.114$
	$min(Aug) = 0.06$	$min(Syn) = -0.208$
	$max(Aug) = 0.423$	$max(Syn) = 0.313$
	$U = 1035, p < .001$	$U = 496, p = 0.400$

2.5.8 Fluid intelligence in advice integration

Model 9a considered the effect of *Stake*, *Adviser type*, and *Fluid Intelligence* on *Optimality*. Only the effect of *Agent type* was significant $[F(1, 4397) = 4.00, p < .05]$.

Analogously, Model 9b considered the effect of *Stake*, *Adviser type*, and *Fluid Intelligence* on *Final accuracy*, with initial accuracy considered as a covariate. The main effect of *Agent type* is significant $[z = 7.326, p < .001]$; the main effect of *Fluid intelligence* is significant $[z = 4.518, p < .001]$; the interaction of *Agent type by Fluid intelligence* is significant $[z = -3.118, p < .01]$; the interaction of *Stake by Fluid intelligence* is significant $[z = -2.391, p < .05]$.

In summary, final accuracy increases with higher fluid intelligence and with the Enhanced Agent; the main effect of stake level on final accuracy is not significant, but the significant interaction indicates that its effect depends on fluid intelligence. The significant interaction between fluid intelligence and agent type indicates that the final accuracy reached with the Alter Ego or with the Enhanced Agent depends on the fluid intelligence score. Considering

initial accuracy as a covariate indicates that the effect of fluid intelligence on final accuracy is not solely due to its effect on initial accuracy.

Model 9c considered the effect of *Stake*, *Adviser type*, and *Fluid Intelligence* on *Final RTs*. The effect of *Agent type* was significant $[F(1, 4565) = 17.868, p < .001]$, and the interaction *Agent type by Fluid Intelligence* was significant $[F(1, 4565) = 10.811, p < .01]$. Results for all models are plotted in Figure 2.11.

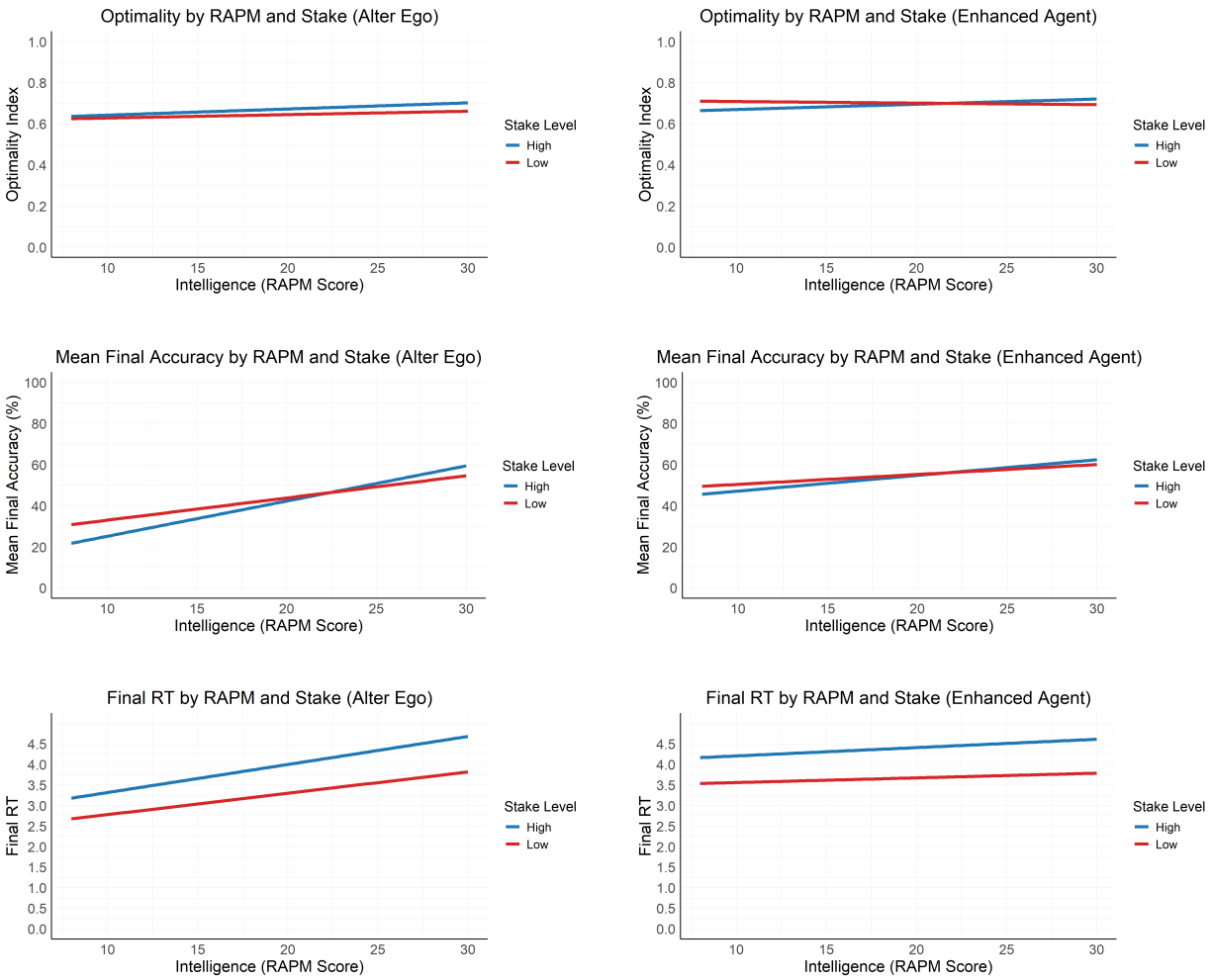

Figure 2.11: Results for our analyses of the combined effects of *Stake* and *Adviser type* on *Optimality*, *Final accuracy* and *Final RTs* depending on the *Fluid intelligence* score (RAPM).

2.6 Results: Block Design

2.6.1 Descriptives

Descriptives for our sample are reported in Table 2.4. Unlike Experiment "Stakes 1", no subjects were excluded for accuracies below the threshold.

Table 2.4: Descriptive statistics for our sample.

Descriptive statistic	**Value**
Sample size	55
Gender	31 female; 24 male
Age	Mean = 23.55, SD = 5.36

2.6.2 Influence

Model 1 considered the effect of *Stake* on *Influence*. Results from Model 1 show that the fixed effect of *Stake* on *Influence* is not significant: $F(1, 8541) = 1.854, p = 0.173$. The $ICC = 0.068$ indicates that about 6.8% of the total variance is due to between-subject variability. Participants adjusted their estimate towards the agent's estimate with an influence of $I = 0.386$ independently of the *Stake* level (mean influence with low stakes $I = 0.381$, mean influence in high stakes $I = 0.391$).

Results from Model 6a considered the effect of *Stake* and *Agent type* on *Influence*. The fixed effect of *Stake* is not significant $[F(1, 8301) = 1.958, p = 0.161]$. The fixed effect of *Agent type* is significant $[F(1, 8300) = 157.9968, p < .001]$. The interaction of the two is marginally significant $[F(1, 8300) = 3.6211, p = .57]$. The $ICC = 0.069$ indicates that about 6.9% of the total variance is due to between-subject variability. Participants adjusted their estimate towards the agent's estimate more with the Enhanced Agent ($I = 0.430$) than with the Alter Ego ($I = 0.341$): influence from the Enhanced Agent increased as stakes increased (from $I = 0.428$ to $I = 0.433$), whereas it instead decreased with the Alter Ego as stakes increased (from $I = 0.353$ to $I = 0.330$). Results for Model 6a are reported below in Figure 2.12.

Model 7a considered the effect of *Treatment* (how stakes trials were presented: intermixed design vs. block design) on *Influence*. None of the effects are significant. The fixed effect of *Stake* is not significant $[F(1, 11893) = 0.118, p = 0.731]$. The fixed effect of *Treatment* is not significant $[F(1, 97) = 2.055, p = 0.155]$. The interaction of the two is not significant $[F(1, 11893) = 0.223, p = 0.633]$. The $ICC = 0.085$ indicates that about 8.5% of the total variance is due to between-subject variability. Mean influence was marginally higher with the Intermixed Design ($I = 0.411$) than with the Block Design ($I = 0.381$), with an opposite trend according to *Stake* level (with Intermixed Design in low stakes: $I = 0.410$; with Intermixed Design in high stakes: $I = 0.412$; with Block Design in low stakes: $I = 0.383$; with Block Design in high stakes: $I = 0.378$). Results for Model 7a are reported below in Figure 2.13.

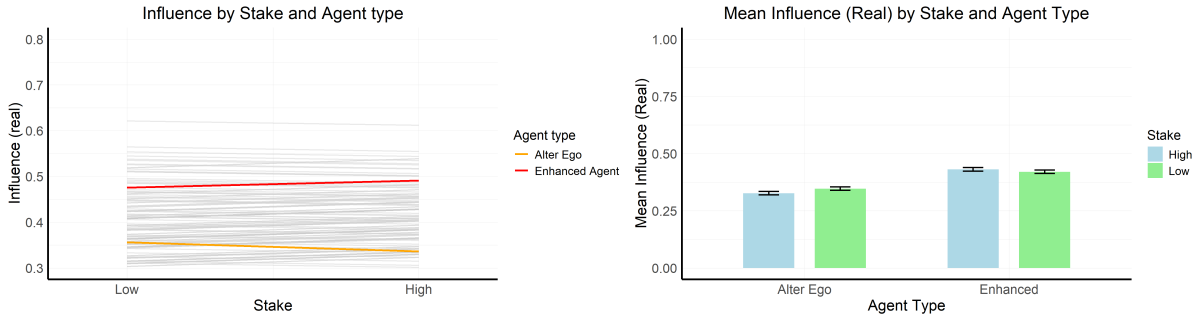

Figure 2.12: Results from the Model 6a for the effect of *Stake* and *Agent type* on *Influence*. On the left: linear mixed model with fixed effects and random intercepts at participant level. On the right: histogram for mean *Influence* depending on the main effects.

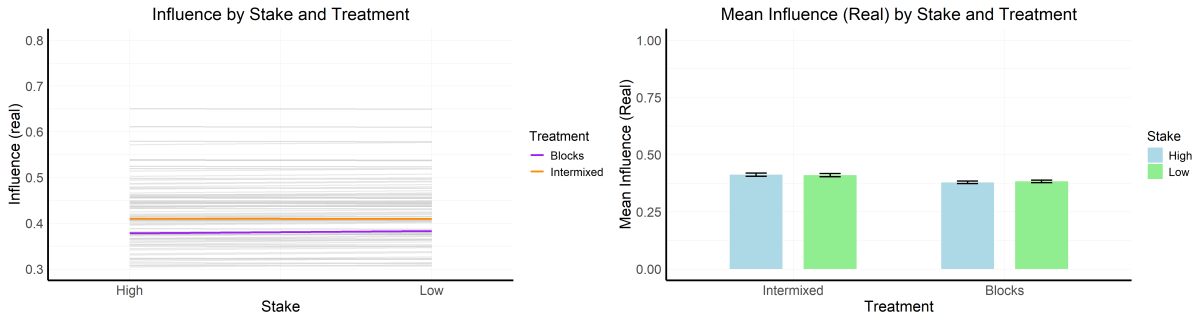

Figure 2.13: Results from the Model 7a for the effect of *Stake* and *Treatment* on *Influence*. On the left: linear mixed model with fixed effects and random intercepts at participant level. On the right: histogram for mean *Influence* depending on the main effects.

2.6.3 Optimality

Model 2 considered the effect of *Stake* on *Optimality*. Results from Model 2 show that the fixed effect of *Stake* on *Optimality* is not significant: $F(1, 8349) = 1.581, p = .209$. The $ICC = 0.038$ indicates that about 3.8% of the total variance is due to between-subject variability. The mean optimality of participants was $O = 0.654$, slightly higher in high-stakes trials ($O = 0.658$) than in low-stakes trials ($O = 0.650$).

Results from Model 6b considered the effect of *Stake* and *Agent type* on Optimality. The fixed effect of *Stake* is not significant $[F(1, 8346) = 1.600, p = .206]$. The fixed effect of *Agent type* is significant $[F(1, 8346) = 17.535, p < .001]$. The interaction of the two is not significant $[F(1, 8346) = 0.010, p = 0.921]$. The $ICC = 0.038$ indicates that about 3.8% of the total variance is due to between-subject variability. Mean optimality was slightly higher with the Alter Ego ($O = 0.668$) than with the Enhanced Agent ($O = 0.640$), independently of the *Stake* level (with Alter Ego in low stakes: $O = 0.665$; with Alter Ego in high stakes: $O = 0.671$; with Enhanced Agent in low stakes: $O = 0.636$; with Enhanced Agent in high stakes: $O = 0.644$). Results for Model 6b are reported below in Figure 2.14.

Model 7b considered the effect of *Treatment* on *Optimality*. The fixed effect of *Stake* is significant $[F(1, 11704) = 9.457, p < 0.1]$. The fixed effect of *Treatment* is not significant

$[F(1, 98) = 0.048, p = 0.827]$. The interaction of the two is not significant $[F(1, 11704) = 0.004, p = 0.949]$. The $ICC = 0.04$ indicates that about 4% of the total variance is due to between-subject variability. Mean optimality was marginally higher with the Intermixed Design ($O = 0.681$) than with the Block Design ($O = 0.679$), increasing by *Stake* level (with Intermixed Design in low stakes: $O = 0.673$; with Intermixed Design in high stakes: $O = 0.690$; with Block Design in low stakes: $O = 0.672$; with Block Design in high stakes: $O = 0.687$). Results for Model 7a are reported below in Figure 2.15.

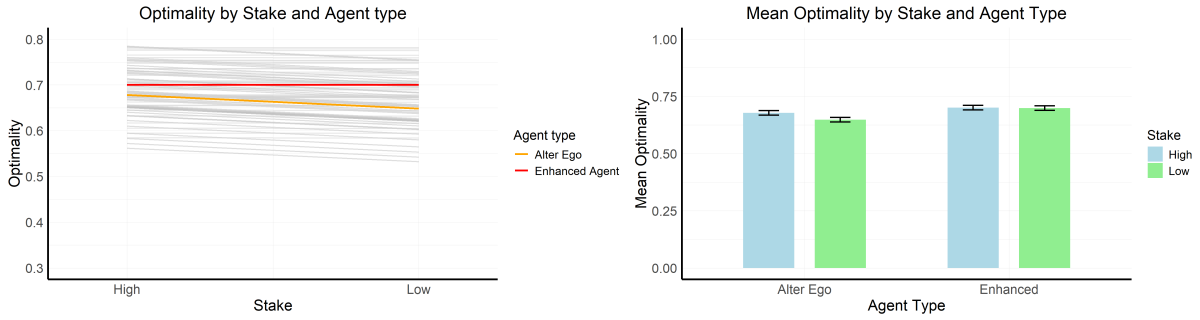

Figure 2.14: Results from the Model 6b for the effect of *Stake* and *Agent type* on *Optimality*. On the left: linear mixed model with fixed effects and random intercepts at participant level. On the right: histogram for mean *Optimality* depending on the main effects.

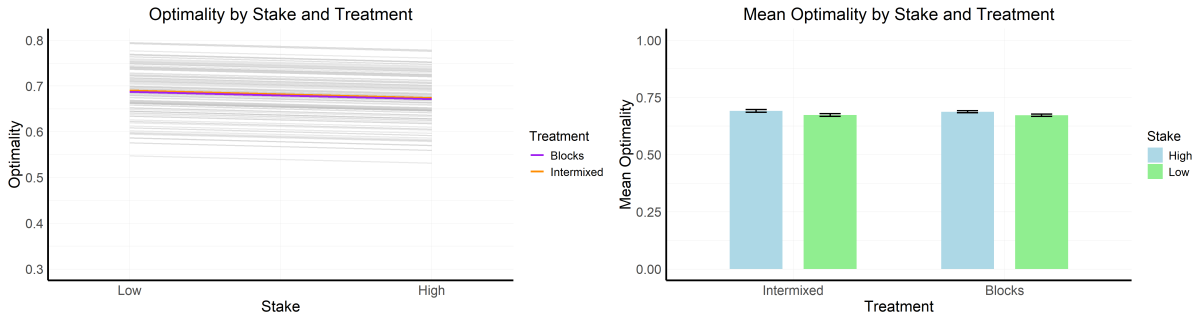

Figure 2.15: Results from the Model 7b for the effect of *Stake* and *Treatment* on *Optimality*. On the left: linear mixed model with fixed effects and random intercepts at participant level. On the right: histogram for mean *Optimality* depending on the main effects.

2.6.4 Accuracy

Model 3 considered the effect of *Stake* on *Accuracy*, considering both the initial and the final decision with the *Decision* factor. Results from Model 3 show that the fixed effect of *Stake* on *Accuracy* is not significant: $z = 0.644, p = 0.520$. The effect of *Decision* is significant: $z = 9.765, p < 0.001$. The interaction between the two is not significant: $z = 0.389, p = 0.697$. The $ICC = 0.340$ indicates that about 34% of the total variance is due to between-subject variability. Mean accuracy of the initial decisions was $iA = 0.440$ for low stakes and $iA = 0.447$

for high stakes. Mean accuracy of the final decisions was $fA = 0.541$ for low stakes and $fA = 0.553$ for high stakes.

Results from Model 6c considered the effect of *Stake*, *Decision* and *Agent type* on *Accuracy*. The fixed effect of *Stake* is not significant [$z = 1.302, p = 0.192$]. The fixed effect of *Decision* is significant [$z = 3.472, p < 0.001$]. The fixed effect of *Agent type* is not significant [$z = 1.009, p = 0.313$]. Considering interactions, *Stake* by *Decision* is not significant [$z = 0.721, p = .470$]; *Decision* by *Agent type* is significant [$z = 4.912, p < 0.001$]; *Stake* by *Agent type* is not significant [$z = -1.197, p = .231$]; *Stake* by *Decision* by *Agent type* is not significant [$z = -0.645, p = 0.519$]. The $ICC = 0.342$ indicates that about 34.2% of the total variance is due to between-subject variability. Mean initial accuracy was comparable across the two agents, as expected for the initial decision, which occurs before encountering the agent (with Alter Ego, $iA = 0.443$; with Enhanced Agent, $iA = 0.445$). Mean final accuracy was higher with the Enhanced Agent ($fA = 0.593$) than with the Alter Ego ($fA = 0.500$). Results for Model 6c are reported below in Figure 2.16.

Results from Model 7c considered the effect of *Stake*, *Decision* and *Treatment* on *Accuracy*. Only the fixed effect of Decision, comparing initial and final accuracy, is significant. The fixed effect of *Stake* is not significant [$z = 1.062, p = 0.228$]. The fixed effect of *Decision* is significant [$z = 9.763, p < 0.001$]. The fixed effect of *Treatment* is not significant [$z = 1.227, p = 0.220$]. Considering interactions, *Stake* by *Decision* is not significant [$z = -0.482, p = 0.630$]; *Decision* by *Treatment* is not significant [$z = -0.730, p = 0.465$]; *Stake* by *Treatment* is not significant [$z = -0.394, p = 0.0.694$]; *Stake* by *Decision* by *Treatment* is not significant [$z = -0.224, p = 0.823$]. The $ICC = 0.268$ indicates that about 26.8% of the total variance is due to between-subject variability. Mean initial accuracy was slightly higher in the second experiment (with an intermixed design, $iA = 0.37$, with a block design, $iA = 0.411$). Mean final accuracy was higher with the intermixed design ($fA = 0.533$) than with the block design ($fA = 0.506$). Results for Model 7c are reported below in Figure 2.17.

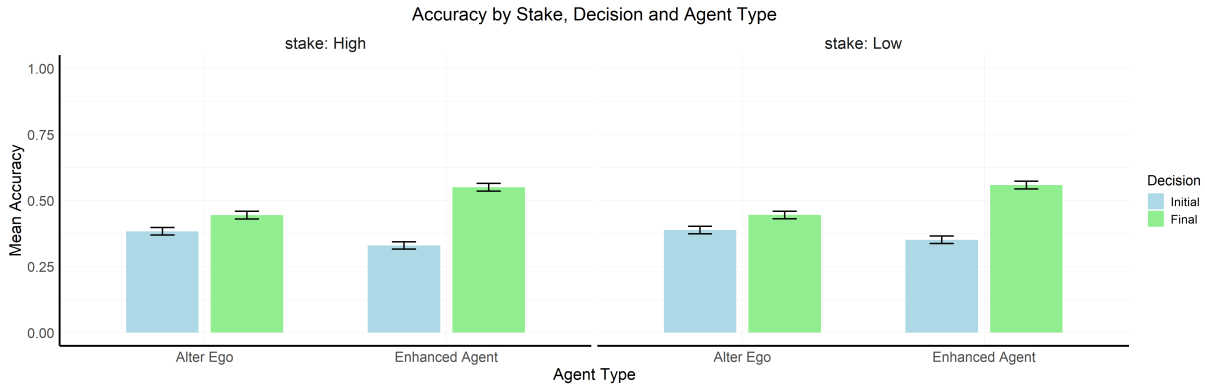

Figure 2.16: Results from the Model 6c for the effect of *Stake*, *Decision* and *Agent type* on *Accuracy*.

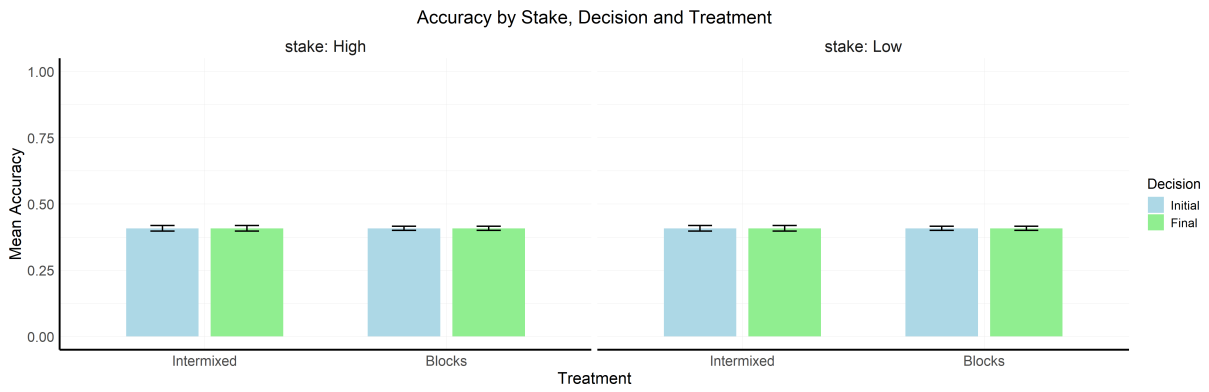

Figure 2.17: Results from the Model 7c for the effect of *Stake*, *Decision* and *Treatment* on *Accuracy*.

2.6.5 Confidence

Model 4 considered the effect of *Stake* on *Final Confidence*. Results from Model 4 show that the fixed effect of *Stake* on *Final Confidence* is not significant: $F(1, 8524) = 2.018, p = .155$. The $ICC = 0.482$ indicates that about 48.2% of the total variance is due to between-subject variability. Mean final confidence expressed by participants was almost identical between stake levels (in low stakes: $fC = 66.41$; in high stakes ($fC = 65.89$).

Results from Model 6d considered the effect of *Stake* and *Agent type* on *Final Confidence*. The fixed effect of *Stake* is not significant $[F(1, 8522) = 2.019, p = .155]$. The fixed effect of *Agent type* is significant $[F(1, 8522) = 6.256, p < .05]$. The interaction of the two is not significant $[F(1, 8522) = 1.301, p = .254]$. The $ICC = 0.483$ indicates that about 48.3% of the total variance is due to between-subject variability. Mean final confidence expressed by participants was almost identical between stake levels (in low stakes: $fC = 66.41$; in high stakes ($fC = 65.89$) and between Agent types (with Alter Ego: $fC = 65.70$; with Enhanced Agent: $fC = 66.60$). The difference between the two *Agent type* conditions was similar independently of the *stake* level: with Low stakes: $fC(AE) = 66.16, fC(EA) = 66.65$; with High stakes: $fC(AE) = 65.23, fC(EA) = 66.55$. Results for Model 6d are reported below in Figure 2.18.

Model 7d considered the effect of *Treatment* on *Influence*. All the effects are significant. The fixed effect of *Stake* is significant $[F(1, 12230) = 13.283, p < .001]$. The fixed effect of *Treatment* is significant $[F(1, 97) = 4.975, p < .05]$. The interaction of the two is significant $[F(1, 12230) = 6.502, p < .05]$. The $ICC = 0.458$ indicates that about 45.8% of the total variance is due to between-subject variability. Mean Final Confidence was lower with the Intermixed Design ($fC = 57.46$) than with the Block Design ($fC = 64.88$), and in both cases it was higher in low stakes as compared to high stakes (with Intermixed Design in low stakes: $fC = 58.48$; with Intermixed Design in high stakes: $fC = 56.45$; with Block Design in low stakes: $fC = 65.01$; with Block Design in high stakes: $fC = 64.74$). Results for Model 7d are reported below in Figure 2.19.

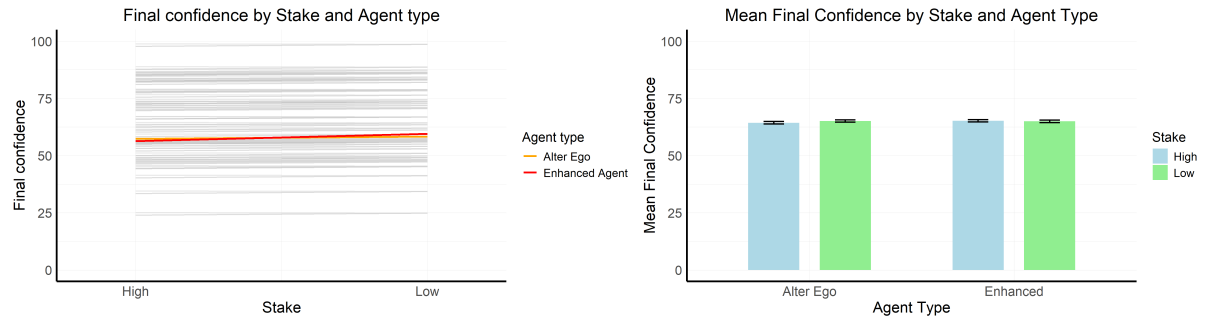

Figure 2.18: Results from the Model 6d for the effect of *Stake* and *Agent type* on *Final Confidence*. On the left: linear mixed model with fixed effects and random intercepts at participant level. On the right: histogram for mean *Final Confidence* depending on the main effects.

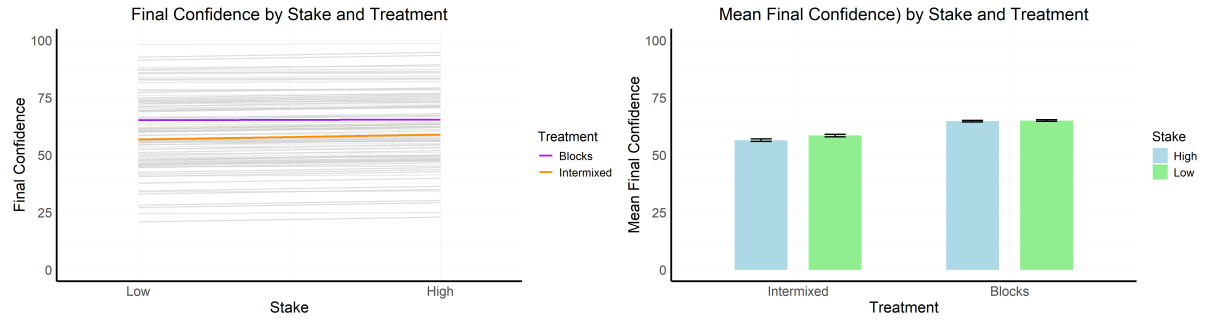

Figure 2.19: Results from the Model 7d for the effect of *Stake* and *Treatment* on *Final Confidence*. On the left: linear mixed model with fixed effects and random intercepts at participant level. On the right: histogram for mean *Influence* depending on the main effects.

2.6.6 Reaction Times

Model 5 considered the effect of *Stake* on *Reaction Times* of the final decision. Results from Model 5 show that the fixed effect of *Stake* on *Reaction Times* is significant: $F(1, 8524) = 11.11, p < .001$. The $ICC = 0.192$ indicates that about 19.2% of the total variance is due to between-subject variability. Mean RTs in high stakes trials are significantly higher ($fRT = 3.68s$) than in low stake trials ($fRT = 3.51s$), with an average gap of 173ms between the two conditions.

Results from Model 6e considered the effect of *Stake* and *Agent type* on *Reaction Times* of the final decision. The fixed effect of *Stake* is significant $[F(1, 8525) = 11.131, p < .001]$. The fixed effect of *Agent type* is significant $[F(1, 8525) = 11.949, p < .001]$. The interaction of the two is not significant $[F(1, 8525) = 3.147, p = .076]$. The $ICC = 0.190$ indicates that about 19% of the total variance is due to between-subject variability. Mean RTs with the Enhanced Agent were higher ($fRT = 3.682s$) than with the Alter Ego ($fRT = 3.502s$). As evidenced by the non-significant interaction, the differences between RTs in low stakes vs. high stakes were only marginally higher with the Alter Ego as agent: with Alter Ego: $fRT(HS) = 3.63s, fRT(LS) = 3.37s, \Delta RT = 264ms$; with Enhanced Agent: $fRT(HS) =$

$3.72s, fRT(LS) = 3.64s, \Delta RT = 81ms$. Results for Model 6e are reported below in Figure 2.20.

Model 7e considered the effect of *Treatment* on *Final RTs*. The fixed effect of *Stake* is significant $[F(1, 12231) = 97.565, p < .001]$. The fixed effect of *Treatment* is not significant $[F(1, 97) = 1.138, p = .289]$. The interaction of the two is significant $[F(1, 12231) = 36.647, p < .001]$. The $ICC = 0.176$ indicates that about 17.6% of the total variance is due to between-subject variability. Mean Final RTs were lower with the Intermixed Design ($fRTs = 3.92s$) than with the Block Design ($fRTs = 3.68s$); the difference between mean RTs in high stakes vs. low stakes was more pronounced in the Intermixed Design ($fRTs(HS) = 4.29s$, $fRTs(LS) = 3.55s$, $\Delta RT = 836ms$) as compared to the Block Design ($fRTs(HS) = 3.76$, $fRTs(LS) = 3.59s$, $\Delta RT = 169ms$) Results for Model 7e are reported below in Figure 2.21.

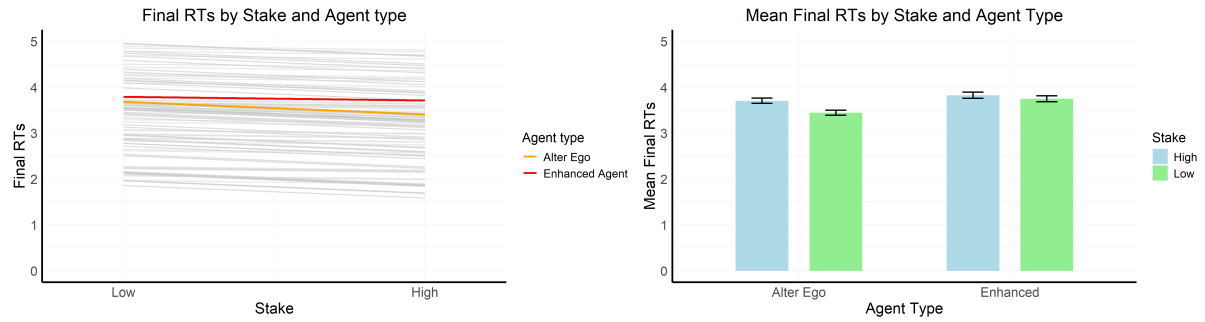

Figure 2.20: Results from the Model 6e for the effect of *Stake* and *Agent type* on *Reaction Times*. On the left: linear mixed model with fixed effects and random intercepts at participant level. On the right: histogram for mean *Reaction Times* depending on the main effects.

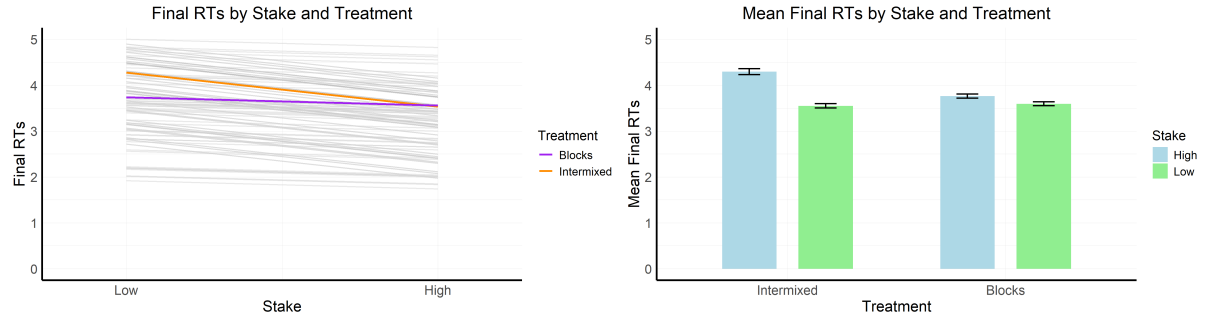

Figure 2.21: Results from the Model 7e for the effect of *Stake* and *Treatment* on *Final RTs*. On the left: linear mixed model with fixed effects and random intercepts at participant level. On the right: histogram for mean *RTs* depending on the main effects.

2.6.7 Augmentation and synergy

Our data show that Human augmentation has been reached with both agents. Synergy was reached only with the Alter Ego as adviser. Results with descriptives and the Wilcoxon

signed-rank test are reported in Table 2.5. Given that the maximum synergy index is positive, some human-agent combinations in our sample were actually able to reach Human-AI synergy.

Table 2.5: Results for augmentation and synergy in our sample.

Adviser	Augmentation Index	Synergy Index
Alter Ego	$M(Aug) = 0.057$	$M(Syn) = 0.097$
	$SD(Aug) = 0.058$	$SD(Syn) = 0.089$
	$min(Aug) = -0.090$	$min(Syn) = -0.077$
	$max(Aug) = 0.244$	$max(Syn) = 0.372$
	$U = 1375, p < .001$	$U = 1319.5, p < .001$
Enhanced Agent	$M(Aug) = 0.148$	$M(Syn) = 0.001$
	$SD(Aug) = 0.079$	$SD(Syn) = 0.081$
	$min(Aug) = 0$	$min(Syn) = -0.179$
	$max(Aug) = 0.321$	$max(Syn) = 0.192$
	$U = 1485, p < .001$	$U = 662.5, p = 0.813$

2.6.8 Fluid intelligence in advice integration

Model 9a considered the effect of *Stake*, *Adviser type* and *Fluid Intelligence* on *Optimality*. The only significant effects were the main effect of *Stakes* $[F(1, 8350) = 3.8731, p < .05]$ and the *Stakes by Fluid Intelligence* interaction $[F(1, 8349) = 5.757, p < .05]$

Analogously, Model 9b considered the effect of *Stake*, *Adviser type* and *Fluid Intelligence* on *Final accuracy*, with initial accuracy considered as covariate. The only two significant effects were the main effect of *Agent type* $[z = 8.048, p < .001]$ and the main effect of *Fluid intelligence* $[z = 2.553, p < .05]$, without any significant interaction.

Model 9c considered the effect of *Stake*, *Adviser type* and *Fluid Intelligence* on *Final RTs*. Te effect of *Stake* was significant $[F(1, 8525) = 70.640, p < .001]$. The effect of *Agent type* was significant $[F(1, 8525) = 27.0814, p < .001]$. The interaction *Agent type by Fluid Intelligence* was significant $[F(1, 8525) = 19.638, p < .001]$. The interaction of *Stake by Fluid Intelligence* was significant $[F(1, 8525) = 60.673, p < .001]$. Results for all models are plotted in Figure 2.22.

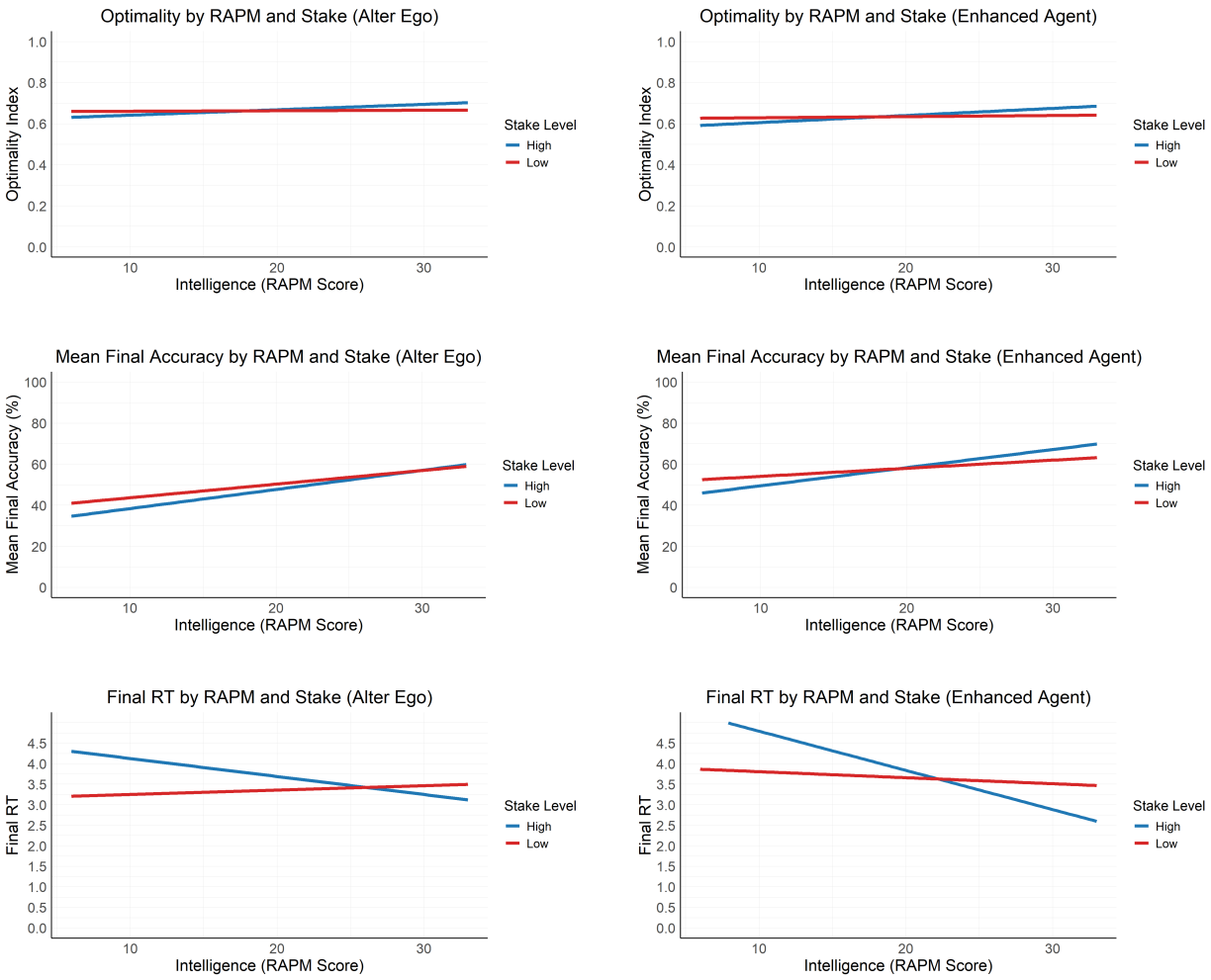

Figure 2.22: Results for our analyses of the combined effects of *Stake* and *Adviser type* on *Optimality*, *Final accuracy* and *Final RTs* depending on the *Fluid intelligence* score (RAPM).

2.7 Discussion

2.7.1 Competence of the Adviser, Decision Stakes and advice integration

In our experiments, we studied the effects of two main factors: the stake involved in the decisions and the type of adviser provided to the participants. We considered the effects on several outcomes at the trial level: influence from the adviser, optimality of the final decision, accuracy of the final decision, confidence in the final decision and reaction times. Results show that the effect of the type of agent was pronounced, whereas the effect of stakes was absent in the critical outcome variables. Focusing on the first experiment, considering only the effect of Stakes (RQ1-RQ5), we see that changes in stakes affected *Final confidence* and *RTs*: higher stakes led to lower confidence in the final decision and slower reaction times. However, effects on *Influence*, *Optimality* and *Accuracy* were not significant. If we consider the models including both factors, with effects of both *Stake* and *Agent type* (RQ6, RQ8b), we see that overall the main effect of *Stake* was significant on *Final confidence* (lower in high stakes) and *Reaction Times* (lower in high stakes) whereas the type of adviser agent affected *Influence*, *Optimality*, *Accuracy*, *Final Reaction Times*.

Stake increase affected reaction times without affecting the accuracy of the final decision, and lowered confidence in the final decision. RTs in high-stakes were slowed by 744ms on average, increasing by almost 20% from 3.56s in low-stakes to 4.31s in high-stakes. Accuracy, however, was unaffected by stake increase, and final confidence decreased slightly from 58% in low-stakes to 56% in high-stakes. No significant gains in optimality were observed. This indicates that external motivational sources, such as economic incentives, might affect cognitive processing (as evidenced by increased RTs) without any actual benefit to performance, thereby reducing confidence in one's decisions. More effortful processing does not translate into performance improvements, revealing, moreover, that economic incentives do not ameliorate egocentric bias or the underweighting of valuable advice in interactive decision making.

Interacting with advisers with different abilities, in contrast, had a more pronounced impact on almost all the variables we considered. Advice from the Enhanced Agent was weighted more (with influence from $I = 0.347$ to $I = 0.485$, an increase of $+39.8\%$) and interaction with it led to lower reaction times (from $fRT = 3.779s$ to $fRT = 4.103s$, $+8.6\%$), but in this case it also led to higher optimality (from $O = 0.663$ to $O = 0.699$, $+5.4\%$), and it led to more accurate final decisions (from $fA = 0.445$ to $fA = 0.554$, $+24.5\%$) for which the difference in final confidence was not significant.

Overall, our results indicate that external monetary incentives have no effect on improving advice integration or its optimality, and instead might even entail increased economic costs and absorb cognitive resources without offering any actual performance benefit. Instead, the adviser's competence in relation to the decision maker seems much more impactful. Our task allows us to tailor the advisers to each participant's abilities, so that not all Enhanced Agents are equal, but each is superior to the participant they are interacting with. Since our enhanced agents are good but not always accurate, our results also show that human users are willing to interact and do benefit from advisers with superior performance, even if

such advisers might sometimes err, contrasting previous research that indicated otherwise (Dietvorst & Bharti, 2020; Dietvorst et al., 2015).

Finally, our results show evidence of both Human augmentation and Human-AI synergy, but also that synergy is not always reached. This is in line with the extensive literature showing that human decision makers generally benefit from adopting advice from artificial support systems, but not always to the point of producing an output that is comprehensively superior to the abilities of each team member taken alone (Vaccaro et al., 2024). More research is needed to identify the predictors of synergy, in order to possibly develop interventions that further improve Human-AI Collaboration.

2.7.2 The effect of Experimental Design

Our two experiments used different designs for the treatment: in the first, stake levels were intermixed across trials; in the second, stake levels were organized in blocks. At the beginning, we wanted to investigate the potential presence of spillover effects between trials, in which participants might transfer a strategy developed in high-stakes trials to low-stakes trials. To account for the possibility of transfer effects across longer sequences of trials, we introduced a block design in the second experiment and tested its effects alongside variation in stake level. Results show that the two experimental paradigms led to similar effects on our outcome variables: experimental design (*Treatment*) had no main effect on *Influence* (model 7a), *Optimality* (model 7b), *Final Accuracy* (model 7c), *Final RTs* (model 7e) and *Egocentric Bias* (model 8c), affecting only *Final Confidence* (model 7d). In this case, final confidence was higher in the block design, with a pattern relative to stakes similar to that in the intermixed design: in both cases, final confidence was lower when stakes were higher, without gains in final accuracy (as shown by non-significant interactions). Overall, these results suggest that spillover effects on short timescales between trials are absent or unimportant for the critical outcome variables.

Comparing the effects of *Stakes* between the two experiments, we can observe mild differences between the effects in the two designs. Overall, we observe that the *Agent type* is more impactful on the outcomes than the *Stake* level. Advice from the Enhanced Agent was weighted more (with influence from $I = 0.341$ to $I = 0.430$, +26%) and led to higher final accuracy gains (from $fA = 0.500$ to $fA = 0.554$, +18.6% or an increase of almost 5 in accuracy score), even if changes in optimality were mildly reduced (from $O = 0.668$ to $O = 0.640$, -4.1%). In comparison, only increasing stakes led to non-significant changes in weight of advice (from $I = 0.381$ to $I = 0.391$, +2.6%) and a mild non-significant increase in optimality (from $O = 0.650$ to $O = 0.658$, +1.2%) without improving final accuracy (from $fA = 0.541$ to $fA = 0.553$). Confidence was almost unaffected by both stake-level manipulations and agent type, but the block design led to sharply higher final confidence (from $fC = 57.5$ to $fC = 64.9$, +13% or an increase of 7.4 in the confidence score). Reaction Times were slowed both by increasing stakes (from $fRT = 3.51s$ to $fRT = 3.68s$, $\Delta RT = 173ms$) and by giving advice with the Enhanced Agent (from $fRT = 3.50s$ to $fRT = 3.68s$, $\Delta RT = 280ms$).

Looking at the interactions, *Treatment* had a pronounced moderation effect on *Final Reaction Times*: RTs were lower in the block design as compared to the intermixed design, and the effect of *Stakes* on *Final RTs* was moderated by the *Treatment* level, being stronger

in the intermixed design as compared to the block design. This finding might indicate that the alternation of stakes within stable patterns (the blocks) is less cognitively demanding than their alternation in random patterns (intermixing), under the hypothesis that reaction times are a behavioral indicator of cognitive load. Parallely, comparing the difference in the effect of *Stakes* on *RTs* between designs might also indicate that the increased cognitive load in the intermixed design might depend on the randomness of the stake level, whereas stable patterns are predictable and less demanding to process (indeed, the block design led to RTs which were on average lower). The significant *Stake-by-Treatment* interaction, in this sense, implies that higher stakes lead to more demanding cognitive processing when the pattern is unpredictable. The absence of improvements in final accuracy, moreover, indicates that the increased cognitive load does not lead to impactful improvements: it is not detrimental to performance since final accuracy increases as compared to initial accuracy, but as a manipulation is far less effective than having good advisers available, reinforcing the conclusions of Experiment 1. These results imply, finally, that more stable environments (variations in stakes across blocks) may ameliorate the cognitive load imposed by high stakes, even if benefits in final accuracy might not follow.

2.7.3 The effect of fluid intelligence on advice integration

The effects of fluid intelligence differed between our two datasets. In our Experiment "Stakes 1", fluid intelligence had no effect on optimality and on reaction times of the final decision. As for final accuracy, the significant interaction effects indicate that improvements in final accuracy are more pronounced for participants with higher fluid intelligence in high-stakes situations and when interacting with the Alter Ego as an agent. In our Experiment "Stakes 2", fluid intelligence had a significant main effect only on final accuracy (which improved with increasing Raven score).

The effect of intelligence on speed in cognitive tasks has been debated for decades (Beck, 1933). Our finding that intelligence negatively predicts reaction times contrasts with the mental chronometry literature, which shows that more intelligent individuals have faster reaction times (Deary et al., 2001; Sheppard & Vernon, 2008). The slower RTs we found might indicate more intense cognitive processing, leading to higher final accuracy, but the relative performance benefits between “low-Raven” and “high-Raven” participants are very close. Indeed, the performance benefit might be driven by the fact that more intelligent participants are already more accurate in the first decision and not by enhanced cognitive processing specific to the final decision: inspection of the means with a mean split by RAPM score reveals that “low-Raven” participants consistently benefit slightly more from advice in relative terms (in Exp.1: from $iA = 0.328$ to $fA = 0.458$, +39.6%; in Exp. 2. from $iA = 0.408$ to $fA = 0.5$, +22.9%) than “high-Raven” participants (in Exp. 1: from $iA = 0.393$ to $fA = 0.533$, +35.6%; in Exp. 2: from $iA = 0.484$ to $fA = 0.593$, +22.5%).

It also seems that interacting with a superior adviser might produce a more pronounced accuracy benefit in “low-Raven” participants in the final decision. Again, inspecting the means of Experiment "Stakes 1", we see that “low-Raven” participants improvement in final accuracy with the Enhanced Agent as opposed to the Alter Ego (from $fA(AE) = 0.383$ to $fA(EA) = 0.540$, +41%) is larger as compared to the improvements gained by “high-Raven”

participants (from $fA(AE) = 0.504$ to $fA(EA) = 0.563$, +11.7%), with a similar pattern between experiments (for Exp. 2, improvements in final accuracy were respectively +23.4% and +13.1%) In a sense, "high-Raven" participants are "already good" at doing our task, and interacting with a superior adviser improves their results to a more limited extent as compared to "low-Raven" participants, for whom the superior adviser gives a comparably higher benefit. Instead, increasing the Stake level leads to opposite effects: in Experiment "Stakes 1", "low-Raven" participants change in final accuracy with high stakes as opposed to low stakes (from $fA(LS) = 0.463$ to $fA(HS) = 0.454$, -1.9%) is negative as compared to the improvements gained by "high-Raven" participants (from $fA(LS) = 0.531$ to $fA(HS) = 0.535$, +1%). In Experiment "Stakes 2", variation in final accuracy with stake increase was respectively -5.3% for "low-Raven" participants and +2.6% for "high-Raven" participants. It seems that increasing the stake level is slightly beneficial for "high-Raven" participants and slightly detrimental for "low-Raven" participants. Instead, while "high-Raven" participants seem to benefit more from increasing stakes than "low-Raven" participants, these benefits are vastly inferior to the benefits gained by interacting with a superior adviser. Moreover, the interaction with a superior adviser seems to partially compensate for the initial accuracy gap between low- and high-Raven participants by providing a larger benefit to the former group, an effect not observed when decision stakes are manipulated.

In conclusion, our data consistently show that improvements in advice-taking and decision performance are driven by the adviser's competence relative to the decision maker. This effect is stronger and much more consistent than the effect produced by decision stakes, which instead might even degrade the confidence in the decisions without any appreciable gain in performance. The adviser's competence benefits participants at all levels of fluid intelligence, but the benefit is larger for those with lower intelligence scores, compensating for their lower initial accuracy. Advice integration is more strongly influenced by the competence of the adviser than by the stakes involved in the decision process.

Chapter 3
Inter-individual Differences in Advice Integration

3.1 Introduction

As we have seen in previous chapters, humans use information given by others suboptimally. Two parallel streams of processes lead to suboptimality: defective integration of factors related to advice quality, and metacognitive limitations in understanding the extent to which confidence in a judgment might reflect its accuracy. Suboptimal use of information may hinder effective collaboration in teams. This is especially relevant in the advancing field of human-AI interaction. It is increasingly important to understand which factors lead human and artificial agents to cooperate optimally, and which factors instead lead to suboptimal or even counterproductive interactions.

The ADT allows to study the influence of an adviser, of the observed integration of advice as compared to optimality, and which components of information are used suboptimally. However, suboptimal use of advice might be influenced by another source of variance: inter-individual variability in advice use. Indeed, individual traits, such as cognitive ability or personality, may influence how advice is integrated. Individual differences in human abilities and mental processes might influence interactions with human or artificial advisers. Researchers have shown that individual traits and dispositions can influence the accuracy and confidence of a decision and also predictions and evaluations about it (Kleitman et al., 2019). Such traits might make some people more prone to overconfidence than others, thereby affecting their reliance on artificial agents. Moreover, little is known about factors relevant to establishing effective collaboration, especially in domains where AI is increasingly used, such as medical decision-making (Knop et al., 2022). Research has already identified several individual differences relevant to decision-making accuracy and confidence (Kleitman et al., 2019). We might ask if these variables, which we already know impact individual decision-making, might also play a role in decision-making with artificial agents. We want to study if and how these inter-individual differences impact overconfidence and the use of advice from artificial agents. A second important contribution is optimality: our task allows us to model optimality as a normative reference to study observed optimality and quantify the gap between the two. By distinguishing between optimality and other constructs, such as egocentric discounting and overconfidence, we can better understand which factors influence advice integration and how they might be related.

In this study, we focused on the following variables as predictors of optimality in advice integration: intelligence, disposition to reflection, personality traits, metacognition, working memory, self-esteem, narcissistic traits, actively open-minded thinking, locus of control, gender, and age. These variables describe "global" functioning of the human decision maker,

and we will study their relationship with "local" variables internal to the ADT that describe metacognitive ability and optimal integration of advice.

Intelligence

Psychometric theories of intelligence describe mental ability as composed of a varying number of factors. The most well-known adopts a single underlying component, the g factor (Deary et al., 2010). Alternatively, Cattell, 1963 and Horn and Cattell, 1966 developed a model in which intelligence is captured by two factors: crystallized intelligence, representing mental abilities that mature and get refined through learning and experience, and fluid intelligence, representing the capacity to adapt to new situations in which crystallized prior knowledge might not give special advantage. In this study, we will use fluid intelligence as a predictor of the optimality of use of advice. Given that advice use requires proper understanding and evaluation of multiple pieces of information, it is reasonable to expect that higher cognitive ability might lead to behavior approaching optimality (or, at least, to lower levels of suboptimality).

Disposition to reflection

Frederick, 2005 developed the Cognitive Reflection Test (CRT) to assess the capacity to suppress intuitive but wrong answers. The CRT is meant to measure the "disposition to reflect on a question and resist reporting the first response that comes to mind" (Frederick, 2005). Enhanced disposition to reflection might consequently indicate that reflective, non-automatic thinking processes are operating and drive the answer at the behavioral level. CRT is important to consider because performance is not solely due to cognitive ability. The disposition to reflect correlates with intelligence, but it is not entirely overlapping with it, capturing nuances of impulsivity and cognitive processing that are not fully explained by intelligence alone. Indeed, the CRT score is a good predictor of performance across several heuristics-and-biases tasks, comparable to intelligence or executive functioning (e.g., Toplak et al., 2011, 2014). We then expect CRT to predict the optimality of advice use.

Personality

Personality is currently understood as a multidimensional construct composed of several dimensions, called personality traits, which are the main factors accounting for people's variance in responding to a wide range of personality items (https://ipip.ori.org/). The two most influential current models of personality are the Five Factor Model (Costa & McCrae, 1992) and, more recently, the HEXACO model with 6 factors (Ashton & Lee, 2007; Ashton et al., 2014; Lee & Ashton, 2008). Uncooperative, antisocial, and self-centered personalities have been described by the Dark Triad Model, which includes narcissism, Machiavellianism, and psychopathy (D. Paulhus & Williams, 2002). Essentially, the Dark Triad of traits could be understood as a personality with very low levels of HEXACO Honesty/Humility trait (Kibeom & Ashton, 2014). Much research has been dedicated to understanding the relationship between personality and overconfidence or optimality, but little is known about this relationship in Human-AI collaboration and about its role in suboptimal integration of advice in these contexts. Felmingham et al., 2021 highlight that "To the best of our knowledge, no study has yet explored the potential impact of personality factors on humans' use of AI, in medicine or elsewhere". Overconfidence in individual decision-making has been

observed among more extroverted individuals (Durand et al., 2013; Schaefer et al., 2004), whereas a negative correlation with overconfidence has been associated with agreeableness (Durand et al., 2013). Schaefer et al., 2004 report that openness and neuroticism do not predict overconfidence, whereas openness correlates with confidence, but not overconfidence, and reflects high accuracy; they also report an effect of agreeableness and conscientiousness on overconfidence, which, however, drops to not significant if the other personality factors are taken into account with partial correlations. Kleitman et al., 2019 report that extraversion is positively associated with cognitive arrogance (an inflated confidence in one's knowledge and abilities), whereas modesty is negatively associated with it, and that openness is negatively associated with dogmatism (an inclination toward rigid thinking, resistance to new knowledge). Finally, Stein et al., 2024 reported a significant effect of agreeableness on attitudes toward AI. Positive attitudes might be antecedents of trust and reliance on artificial agents: this is expected from the Theory of Planned Behavior (TPB), which states that attitudes positively correlate with intentions, which themselves correlate positively with behavior Ajzen, 1991. Consequently, we consider personality in order to understand its effects on advice integration in interaction with artificial agents, beyond what is currently known for individual decisions.

Metacognition

Metacognition consists of cognitive processes about other cognitive processes and includes the knowledge a person has of their mental processes. Most metacognition is automatic and unaware; it can be conceptualized as a set of multiple dimensions (Flavell, 1979) and its applications span all realms of psychology, from the study of cognition to clinical psychology (Norman et al., 2019) and psychotherapy (Dimaggio et al., 2013). Multiple conceptualizations of metacognition and different measures are known (Fleming & Lau, 2014; Rahnev, 2023). Some measures distinguish between type 1 and type 2 data to measure metacognitive sensitivity and offer a more "local" measure (Maniscalco & Lau, 2014), whereas others are based on metacognition as a global construct, spanning the functioning of the mind as a whole, in its ability to understand one's own mental states, the mental states of others and to reflect on them (Dimaggio et al., 2013). The Metacognition Self-Assessment Scale is a well-known measure reflecting this broad definition of the construct (Pedone et al., 2017). There is evidence that metacognition might be a domain-general process (Mazancieux et al., 2020), and, indeed, metacognition, as measured by the MSAS, can be understood as a multidimensional process with a general factor underlying four components (Pedone et al., 2017). In particular, metacognition involves understanding and reflecting on one's mental states (Self-reflectivity), the ability to distance oneself from cognitions and consider them critically (Critical Distance), the ability to understand the mental states and contents of others (Understanding Other Minds), and finally the ability to regulate one's own inner state (Mastery, or Metacognitive Regulation). We hypothesize that high metacognitive abilities might be linked to better calibration and lower overconfidence (with one's confidence reflecting accuracy more closely), and to better use of advice.

Self-esteem

Previous work has shown that self-esteem affects decision-making performance. Indeed, inflated self-esteem might indicate cognitive arrogance and, consequently, increased reliance on one's own opinion (Kleitman et al., 2019).

Narcissism

Narcissism is intensely studied in clinical psychology and psychopathology, where it is conceptualized as ranging from a non-pathological personality type to a personality disorder (McWilliams, 2020). Narcissism can be understood in terms of personality traits. According to Pincus et al., 2009 and to Wright et al., 2010, seven sub-factors underlie a narcissistic personality, structured in two superordinate factors: grandiose narcissism (characterized by the factors of Entitlement Rage, Exploitativeness, Grandiose Fantasy, Self-sacrificing Self-enhancement) and vulnerable narcissism (characterized by Contingent self-esteem, Hiding the Self, and Devaluing). Self-enhancement and the tendency to inflate self-esteem make narcissism a good candidate predictor of overconfidence. Narcissists tend to overestimate themselves and their capabilities; they tend to rely more on their intuition as compared to non-narcissists, they are confident in their decision, even if their confidence does not generally reflect their accuracy, and tend to blame others (O'Reilly & Hall, 2021). Their sense of confidence does not generally reflect a superior performance (Blair et al., 2008; Guedes, 2017). They are prone to overclaiming (D. L. Paulhus et al., 2003) and show multiple self-enhancement biases (Colvin & Block, 1994). Moreover, people with narcissistic traits tend to take less advice than others, a stable and robust negative relation that seems not moderated by the level of expertise of the advice giver (Stöcker & Schütz, 2024). It is reasonable, then, to expect that narcissistic personality traits might play a role in the use of advice from others. Older conceptualizations of narcissistic personality in psychiatry were more focused on the grandiose phenotype, whereas more contemporary conceptualizations also include the vulnerable type (for instance, as can be seen by comparing the DSM III (American Psychiatric Association, 1980) to the alternative model proposed in the DSM 5 (American Psychiatric Association, 2013). The two differ in their self-esteem: whereas grandiose narcissists consistently manifest self-enhancement, vulnerable narcissists tend to have more contingent self-esteem (Pincus et al., 2009). We might ask then if the relationship of narcissism with overconfidence and use of advice varies accordingly.

Actively Open-minded Thinking

Actively Open-minded Thinking (AOT) is a thinking disposition characterized by openness to change and the acquisition of new information, as well as cognitive flexibility, a disposition to change beliefs (Stanovich & Toplak, 2023). AOT consists of two processes: increased seeking of information contrasting one's attitudes (its "activity") and higher levels of processing the information newly acquired (its "openness"). AOT dispositions contrast with dogmatism: the tendency to rely on rigid thinking, not to revise the knowledge we already possess. Kleitman et al., 2019 showed that AOT loads negatively onto the Dogmatism factor, and consequently, it is associated with better calibration of one's confidence. We might also ask whether being active and open in searching for and processing contrasting information affects advice use.

Locus of Control

We include Locus of Control (LOC) in our set of predictors, given its relevance to research on social influence, as it has been extensively studied across many realms of psychology (Furnham & Steele, 1993). LOC describes the extent to which a person thinks having control over his or her own life, as opposed to a control exerted by external influences (Rotter, 1954, 1966).

LOC describes two different tendencies and expectations people might have: people believing reinforcements as mostly contingent upon their own actions are said to have *internal* LOC, whereas people believing that reinforcements mostly come from factors outside themselves, such as luck, fate or the actions of others, are said to have *external* LOC. The two tendencies are conceptualized not as a rigid dichotomy but as a shaded continuum. LOC has been conceptualized as a global, domain-general construct spanning one's overall life impression, or, more specifically, a variety of dimensions and measures dedicated to life aspects (work or academic achievement, health and wellbeing, etc.) (Furnham & Steele, 1993). Internal locus of control has been related to greater information search (Srinivasan & Tikoo, 1992). Most research has indicated that external LOC might be linked to proneness to persuasion and social influence, although this claim has been challenged by other findings (Avtgis, 1998).

Gender and Age

We collect gender and age as variables with other demographics. Differences in the adoption of technology by gender are a longstanding phenomenon: compared to women, men are more inclined to use technology and participate more in technological fields that are more science- and math-intensive, such as STEM and PECS (Bimber, 2000; Buser et al., 2014, 2017; Cimpian et al., 2020). Analogous differences in the adoption of generative AI by gender are currently under study (Carvajal et al., 2024). Males generally hold more favorable attitudes toward technology than females (Cai et al., 2017), and, following the TPB, this could lead to increased reliance on and more optimal use of advice. However, males also report higher self-esteem than females (Kling et al., 1999), and this, in line with our hypothesis on the role of self-esteem in overconfidence, might lead to a null effect of gender on the use of advice. We also expect stronger overconfidence in male participants. Age could be linked to familiarity with technology, and younger age is associated with more positive attitudes towards technology (Schepman & Rodway, 2022), although the age range of our sample might be too limited to draw meaningful conclusions.

3.2 Hypotheses

3.2.1 Research question

Our aim is to investigate whether there are inter-individual differences in the optimality of advice use, as measured by our predictors. Optimality index is our main dependent variable, the scales measuring individual variables are the main independent variables, and we consider egocentric bias and miscalibration as possible mediators of their relationship with optimality. Each of our inter-individual variables (X_i) is considered an independent variable in a Structural Equation Model, with optimality index as dependent variable (O) and with egocentric bias (E) and miscalibration (M) as mediators (indicated in the equation with the typical R syntax of the *psych* package). Numerical coefficients (β_i) are estimated with regression and express the relationship between each predictor and the outcome, net of the other predictors. The model equation consequently takes the following expression:

$$O \sim \beta_0 + \sum_{i=1}^{N} \beta_i X_i + (E) + (M)$$

3.2.2 Hypotheses

The use of advice approaches optimality as the participant's behavior converges toward the Bayesian model's influence. As behavior tends to optimality, the distance between observed and optimal influence will be lower. Behavior tending to optimality leads to the optimality index approaching 1 and ego bias approaching 0. Calibrated decisions mean that the difference between confidence and accuracy tends to 0 (miscalibration towards overconfidence indicated by positive values, towards underconfidence by negative values).

Optimality and overconfidence are theoretically different. Subjects might be overconfident, yet their use of advice might still be close to optimal. It is possible indeed that a subject might be overconfident and recognize it precisely, and consequently regulate his/her behavior accordingly. The relationship between overconfidence and optimality is an empirical question we investigate here, and we assume that egocentric bias and miscalibration might be two sources of suboptimal behavior.

We hypothesize that the relationship between individual predictors and optimality might be mediated by egocentric bias and miscalibration, with negative relationships between miscalibration and optimality and between egocentric bias and optimality. As for the individual traits used as predictors, we formulated expectations regarding their contributions, listed below. Predictions for their effects on the mediators are considered mostly exploratory, and we do not expect their effects to be necessarily always anti-correlated.

- Negative relationship of intelligence and disposition to reflection with the mediators (with a positive effect on optimality)
- Negative relationship of global metacognition with the mediators (with a positive effect on optimality)
- Positive relationship of narcissistic personality phenotypes with the mediators (with a negative effect on optimality)
- Positive relationship between self-esteem with the mediators (with a negative effect on optimality)
- Negative relationship of actively open-minded thinking with the mediators (with a positive effect on optimality)
- Internal locus of control related positively with the mediators (with a negative effect on optimality)
- Mixed effects relative to personality traits, which we considered as an exploratory question
- Overconfidence is more pronounced in males, but overall a null effect of gender on optimality

- Null effect of age on optimality

We plan to compare two possible mediation models with two different SEMs:

1. **Model 1** considers all the direct effects of the predictors on optimality and all the effects through both mediation paths. The hypothesis underlying this model is that all individual traits might affect advice integration through both streams of cognitive limitations.

2. **Model 2** considers all the direct effects on the predictors on optimality, but predictors are grouped together in different mediation paths, under the hypothesis that different individual traits might contribute differently to the two streams of cognitive limitations investigated here. Some traits will be considered in both mediation paths because they might impact both the use of advice (its influence) and the miscalibration of confidence.

We consequently formulate these model structures:

Model 1

$$O \sim \beta_0 + \sum_{i=1}^{N} \beta_i X_i + (E) + (M)$$

$$(E) \sim \sum_{i=1}^{N} \beta_i X_i$$

$$(M) \sim \sum_{i=1}^{N} \beta_i X_i$$

Model 2:

$$O \sim \beta_0 + \sum_{i=1}^{N} \beta_i X_i + (E) + (M)$$

$$(E) \sim RAPM + CRT + HAS + BPNI + RSES + LOC$$

$$(M) \sim RAPM + CRT + MSAS + BPNI + AOT + gender + age$$

Model selection and interpretation will follow according to the model with the superior fit, according to its fit indices.

3.3 Methods

3.3.1 Experimental Task

Data relative to optimality, egocentric bias, and miscalibration were collected in the experiments with the ADT, presented in chapter 2. We used six indices obtained with the ADT to quantify advice integration and its (sub)optimality. Three individual indices describe the characteristics of the participant:

- mean accuracy (mA)
- mean confidence (mC)
- confidence predictivity (CP)

Three other indices quantify advice integration; they can be used as trial-by-trial measures, but here they are averaged across all trials for each participant to obtain an individual index.

- **Optimality index**: measure of optimality taking into account the absolute difference between optimal and observed influence, averaged across all trials. This index equals 1 minus that difference, so values tending to 0 indicate increasing suboptimality, whereas values tending to 1 indicate optimality (where the difference between optimal and observed influence would be low or negligible).

$$O = 1 - |I_{OPT} - I_{OBS}|$$

- **Egocentric bias index**: underestimate of advice in respect to optimality, quantified as the difference between optimal influence and observed influence. Optimal integration is characterized by the index approaching 0, whereas increasing egocentric advice discounting is characterized by the index approaching 1 (because observed influence is low).

$$E = I_{OPT} - I_{OBS}$$

- **Miscalibration**: it is computed by subtracting the participants' average accuracy from their average confidence across all trials, yielding an individual index of confidence calibration. Positive values indicate overconfidence; negative values indicate underconfidence.

$$M = mC - mA$$

3.3.2 Measures of inter-individual variability

We collect data on inter-individual variability using a battery of measures administered in the laboratory. The battery takes about one hour to complete.

Intelligence. We use the Raven Advanced Progressive Matrices (RAPM-III) in the 36-item short version by Hamel and Schmittmann, 2006. This is a time-limited version: participants must complete the assessment within 20 minutes. Items are listed by increasing difficulty.

Disposition to reflect. The CRT by Frederick, 2005 originally contained 3 items. Subsequently, longer versions with more items have been developed (Toplak et al., 2014). Here we adopt the CRT-long version by Primi et al., 2016.

Personality. We measure the six domains of personality with the HEXACO questionnaire. Compared with the Five-Factor Model, the HEXACO is based on a more extensive conceptualization of personality, including the Honesty/Humility factor. A version based on adjectives has recently been developed and validated for the Italian context (HAS, Romano et al., 2023).

Metacognition. The MSAS by Pedone et al., 2017 offers us a measure of metacognition as a global construct. We chose it because there is evidence that metacognition might be a domain-general process (Mazancieux et al., 2020), and we can study its relation with internal measures of metacognition provided by the ADT (mean confidence and confidence predictivity).

Self-esteem. We measure it with the 10-item Rosenberg Self-Esteem Scale (Rosenberg, 1979) validated in Italy by Prezza et al., 1997.

Narcissism. The inclusion of a dedicated measure of narcissism is recommended because conceptualizations of this construct from social-personality research do not completely overlap with clinical narcissism (Miller & Campbell, 2008). Its inclusion in our battery has been inspired by the work of Kleitman et al., 2019. They used the Narcissistic Personality Inventory (Raskin & Hall, 1979) with the hypothesis that narcissism might contribute to cognitive arrogance. However, the NPI is based on an older conceptualization of narcissism, coming from traditional psychiatry. Researchers in clinical psychology have subsequently developed and refined an extended conceptualization of the construct, including not only the grandiose phenotype but also the vulnerable phenotype, and have developed dedicated psychometric measures, such as the Pathological Narcissism Inventory (Pincus et al., 2009). The B-PNI is a brief version of 28 items we adopt for our purposes (Schoenleber et al., 2015).

Actively Open-minded Thinking: Dogmatism indicates resistance to new information, and AOT negatively correlates with it (Kleitman et al., 2019). A scale to assess AOT was initially developed by Stanovich and West, 1997. Stanovich and Toplak, 2023 recently published a review of its measurement; we use their 13-item scale, which accommodates multiple observations and criticisms of previous measures of the construct over the past 25 years.

Locus of Control: the Rotter's Internal-External Control Scale (RIECS) is the self-report measure initially developed by Rotter, 1966 to study the construct; it is composed of 29 paired statements, each item requiring a forced choice between one of the two. It was originally known as the Internal-External Control Scale, which expresses the two tendencies that the

LOC construct aims to capture. Among the many measures available today, we chose the Mini Locus of Control Scale developed in the Italian version by Perussia and Viano, 2008, composed of 6 items.

3.3.3 Recruiting and participation criteria

Participants will be recruited through the University of Milano-Bicocca's mailing lists and dedicated recruiting platforms (Sona; Orsee). The two systems allow the advertisement of the experiment to a list of voluntary students from different departments of the University. However, participation in the experiment will not be restricted to students of the University of Milano-Bicocca only. Participation in the experiment requires the usual criteria we apply for the ADT: normal-to-corrected vision, no neurological disorders, and knowledge of Italian to ensure understanding of the experimental instructions.

3.3.4 Sample size

The sample size was determined using a power analysis based on our primary analyses, primarily Pearson's correlation between our inter-individual measures and optimality in the ADT. We set a correlation coefficient of 0.30 as our expected effect size, based on the observed correlation in previous experiments between fluid intelligence (measured by the Raven APM) and optimality in the ADT. This corresponds to a medium effect size. We assume a one-tailed test, with a required statistical power of 80% and a conventional significance level of .05. The power analysis yields a minimum sample size of 69 participants.

3.4 Analyses

We used Structural Equation Modelling, as defined in our research question, to incorporate mediation into a multiple linear regression model. Measures of individual traits were treated as independent variables, and the optimality index as the dependent variable, with indices of miscalibration and egocentric bias as mediators of their relationships. It is possible that two predictors may be highly correlated, a condition known as *collinearity*. High collinearity makes it difficult to estimate regression coefficients because it is hard to disentangle the effects of each predictor on the outcome. When three or more variables are collinear, simply inspecting a correlation matrix is not sufficient to detect this issue, called *multicollinearity*. To assess its impact, we computed the *Variance Inflation Factor* (VIF), which indicates the extent to which multicollinearity increases the variance of the regression coefficients. The VIF varies from 1 to infinity; values exceeding 5 or 10 indicate problematic multicollinearity (James et al., 2021). Selection between Model 1 and Model 2 was based on their fit as assessed with the *Akaike Information Criterion* (AIC: Akaike, 1992) and *Bayesian Information Criterion* (BIC: Schwarz, 1978), which take into account goodness of fit while accounting for model complexity; the model with lower indices is to be selected.

3.5 Results

3.5.1 Descriptive statistics for the individual indices

The ADT provides several indices to measure advice-taking and metacognitive ability. Four of these indices describe characteristics of the agent involved in a decision: mean initial accuracy, mean (final) accuracy, mean confidence, and confidence predictivity. Three more ADT indices were used to quantify advice integration: optimality, ego bias, and miscalibration, according to our research question. Descriptive statistics for indices measured with the ADT are reported in Table 3.1. Our sample of $N = 82$ participants was self-reported 56% female, 41% male, and 3% neither of the two ("non-binary"); mean age was 23.06 years, $SD = 4.58$.

Table 3.1: Descriptive statistics for the indices measured with the ADT.

Index	**Mean**	**Min**	**Max**
Optimality	0.655	0.436	0.783
Ego Bias	0.180	-0.036	0.427
Initial Confidence	0.580	0.197	0.998
Final Confidence	0.631	0.230	0.988
Initial Accuracy	0.428	0.096	0.756
Final Accuracy	0.536	0.147	0.885
Confidence Predictivity	0.236	-0.085	0.533
Miscalibration (initial decisions)	0.152	-0.495	0.525
Miscalibration (final decisions)	0.095	-0.552	0.451

3.5.2 Correlations between individual indices from the ADT

Several correlations among individual indices from the ADT are significant and are shown in Figure 3.1. Participants whose behavior tended to optimality were less ego biased [$r = -0.76, p < .001$], more accurate in initial decision [$r = 0.4, p < .01$], more accurate in final decision [$r = 0.5, p < .001$], and less overconfident in initial estimates [$r = -0.35, p < .001$]. More accurate participants were also more confident [$r = 0.55, p < .001$], whereas more overconfident participants were less optimal [$r = -0.45, p < .001$], less accurate in initial decisions [$r = -0.42, p < .001$], less accurate in final decisions [$r = -0.45, p < .001$] and more confident [$r = 0.49, p < .001$].

3.5.3 Effect of inter-individual variables on optimality

Model comparison shows that Model 2 provides a superior fit to the data [$\chi^2 = 20.505, p = .198; CFI = 0.959; TLI = 0.845; RMSEA = 0.059; SRMR = 0.025; AIC = -481.679; BIC = -368.563$] as compared to Model 1 [$\chi^2 = 4.442, p < .05; CFI = 0.968; TLI = -0.892; RMSEA = 0.205; SRMR = 0.013; AIC = -467.742; BIC = -318.525$]. The non-significant χ^2 test indicates no discrepancy between the model-implied covariance matrix and the observed

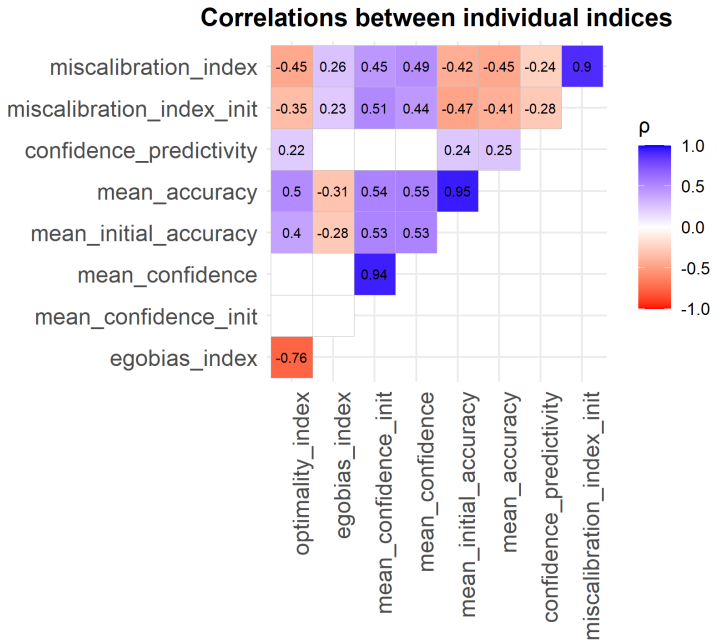

Figure 3.1: Correlations between individual indices from the ADT

matrix. Indices highlight good model fitting to the data. The $R^2 = 0.70$ indicates that 70% of the variance in optimality is explained by the model. Assessment of multicollinearity showed that all VIF values are below 4, with *BPNI-v* having the highest VIF of 2.615, indicating low multicollinearity between the variables.

Several direct effects are significant (standardized coefficients on all the data space are reported): higher fluid intelligence is associated with higher optimality [$\beta = 0.225, p < .01$]; higher extraversion is associated with higher optimality [$\beta = 0.147, p < .05$]; higher grandiose narcissism is associated with lower optimality [$\beta = -0.211, p < .05$]; higher vulnerable narcissism is associated with higher optimality [$\beta = 0.324, p < .001$]; higher egocentric bias is strongly associated with lower optimality [$\beta = -0.722, p < .001$]; higher miscalibration (overconfidence) is associated with lower optimality [$\beta = -0.267, p < .001$]. Disposition to reflection, global metacognition, self-esteem, actively open-minded thinking, locus of control, gender and age had no statistically significant direct effects on optimality.

The mediation paths from the trait predictors to ego bias are significant for honesty/humility (higher levels predict higher egocentric bias: $\beta = 0.225, p < .05$) and agreeableness (higher levels predict lower egocentric bias: $\beta = -0.355, p < .05$). The mediation paths from the trait predictors to miscalibration are significant for disposition to reflection (higher levels predicting lower miscalibration: $\beta = -0.237, p < .05$) and gender (males had lower levels of miscalibration: $\beta = -0.387, p < .001$); a marginally significant effect of metacognition (understanding other minds) also emerged ($\beta = 0.193, p = .060$). Results are plotted in Figure 3.2 below.

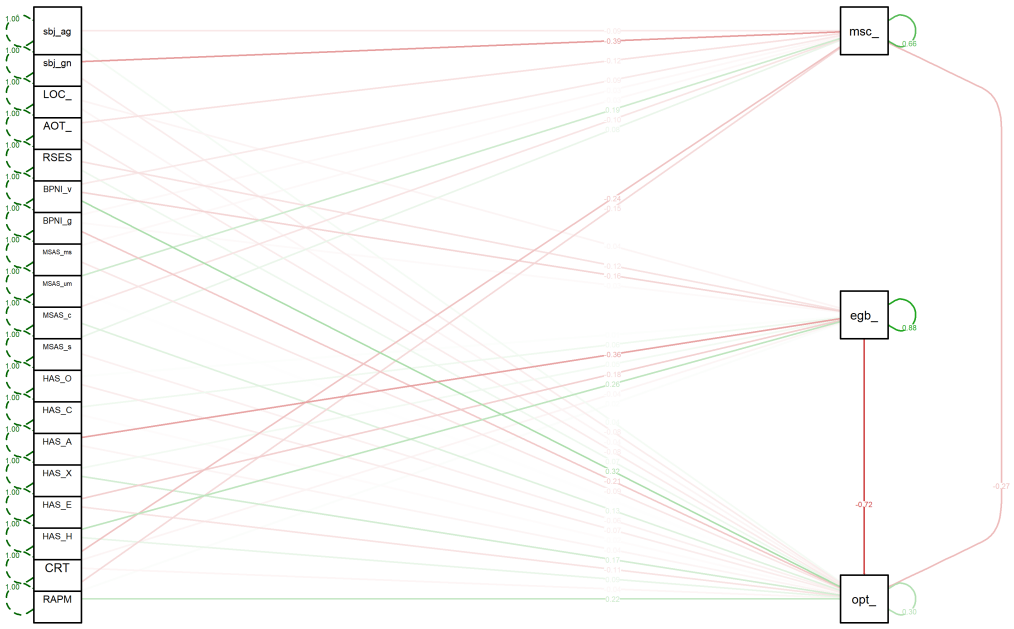

Figure 3.2: Plot of the structural equation model used in our analyses. Green paths represent positive relationships; red paths, negative relationships. Standardized estimates across the entire data space are considered, and fading is used to express the estimate magnitude, with lower fading corresponding to a larger coefficient.

3.5.4 Correlations

As ancillary analysis, we report the correlation matrix between predictors and indices from the ADT, shown here in Figure 3.3. Focussing on the predictors of the ADT internal indices, we can observe that participants who tended more toward optimality scored higher in fluid intelligence [$r = 0.23, p < .05$]; miscalibrated (overconfident) participants tend to score lower in (negative correlation with) intelligence [$r = -0.31, p < .001$] and disposition to reflection [$r = -0.26, p < .05$]; they tend not to self-report as male (female or "not-binary") [$r = -0.39, p < .01$], and score as more "internal" in Locus of Control [$r = 0.23, p < .05$]. More accurate participants tend to score higher in intelligence [$r = 0.29, p < .01$] and disposition to reflection [$r = 0.29, p < .01$]. Confidence predictivity correlates positively only with intelligence [$r = 0.25, p < .05$] and disposition to reflection [$r = 0.33, p < .01$].

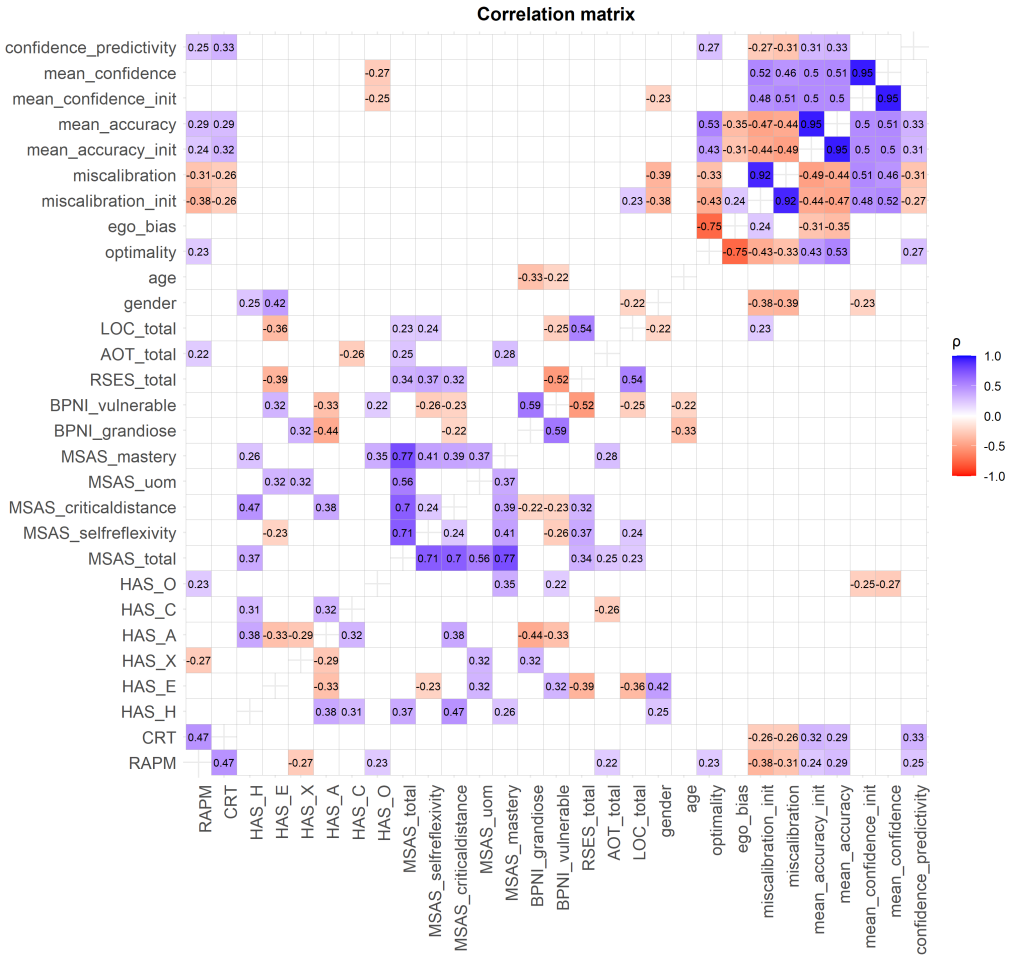

Figure 3.3: Correlation matrix for the variables collected in our studies on advice taking. The full matrix is displayed without the diagonal (where the correlation would be 1).

3.6 Discussion

An initial inspection of the correlation between ADT indices shows that optimality, egocentric bias, and miscalibration are theoretically distinct but empirically related. Optimality and egocentric bias are negatively correlated: higher egocentric advice discounting is associated with lower optimality. Similarly, the negative correlation between miscalibration and optimality shows that overconfident participants were less optimal. The link from egocentric bias and miscalibration to (sub)optimality is evidenced by their significant effects in the SEM, especially the former, which was the strongest among all the predictors we considered. Participants were markedly overconfident: their mean confidence exceeded their mean accuracy. Their confidence predictivity (i.e., the correlation between trial-by-trial confidence and accuracy) was also low, indicating low reliability of the confidence judgments expressed in the ADT and low predictivity of confidence for accuracy. These findings are consistent with the metacognitive limitations evidenced by previous literature (Ackerman, 2023; Fiedler et al., 2019; Haddara & Rahnev, 2022; Moore & Schatz, 2017). They are also consistent with

data from previous studies by our research group on advice-taking (Zonca et al., 2025).

Data from our research to date support the hypothesis that individual traits have a direct effect on advice-taking, and they also support the hypothesis that egocentric advice discounting and miscalibration mediate their relationship with the optimality of advice-taking. They also support our hypothesis that the two mediation paths describe different relationships between individual traits and optimality. Higher optimality is predicted by higher extraversion, by the positive indirect effect of honesty/humility, and the negative indirect effect of agreeableness on egocentric bias. Our findings relative to extraversion complement previous studies that investigated the relationship between personality traits and miscalibration, mainly in the form of overconfidence and cognitive arrogance; their results show that overconfidence correlates positively with extraversion, and negatively with agreeableness and modesty (Durand et al., 2013; Kleitman et al., 2019; Schaefer et al., 2004). Data collected in our experiments indicate that extraversion is also positively associated with optimality. Our data for agreeableness complement also the findings from these studies: agreeableness has a net positive indirect effect on optimality through its negative effect on egocentric bias, possibly because this trait describes a tendency to being more cooperative and flexible in compromises. Other personality traits led to nonsignificant effects. More data are needed to understand the relationship between personality traits and advice integration in interactive contexts: other constructs as potential mediators to optimality should be considered in our future analyses.

The negative effect of grandiose narcissistic traits on optimality was expected: as we saw in the introduction, people with high narcissistic traits tend to take less advice than others, a robust effect which is not moderated by the expertise of the adviser (Stöcker & Schütz, 2024). Our data indicate that this effect is direct on optimality, without mediation by egocentric bias or miscalibration. The effect of vulnerable narcissism it's the fourth strongest among our predictors and is instead the opposite of grandiose narcissism. Given that the two opposite effects have comparable magnitude, and that grandiose and vulnerable narcissism are correlated in our sample, it is possible that their two opposite direct effects on optimality might be partially balanced. The effect of narcissistic traits on optimality might consequently be mixed. These results highlight the importance of considering both shades of narcissism, overcoming limitations of previous literature focussing more on the grandiose phenotype (Kleitman et al., 2019), a conception of narcissism which is too narrow to describe this personality thoroughly (Pincus et al., 2009; Wright et al., 2010). Given that only direct effects of narcissistic traits were significant, it is possible that other mediators might be relevant in the relationship with optimality.

The effect of global metacognition is absent, contrary to our expectations, with only a marginally significant effect of one of the four scales (*Understanding Other Minds*). Our study considers processes at two levels: at a local level we measure advice weighting and metacognitive ability internal to the ADT (with egocentric bias and miscalibration) as endogenous variables; at a global level we measure metacognition as a set of traits characterizing the individual as a whole, spanning its functioning "on average" across life (with the MSAS factors). The two measures differ also by type: the local measures are inferred with a behavioral task, whereas the MSAS is a self-report questionnaire. We expected that higher metacognitive abilities would predict optimality of advice taking, but our local measures and the global measures evidence a divergence between the two. Unexpectedly, global metacognition has

neither significant direct effect on optimality nor indirect effects. Instead, findings from our local variables (ego bias index and miscalibration index, internal to the ADT) show that higher egocentric bias and higher miscalibration (overconfidence) lead to lower optimality, a result coherent with studies from our research group showing that metacognitive ability predicts advice use (Zonca et al., 2025). The puzzling effects of metacognition might be considered in the light of the difference between the two measures. High scores in the self-report metacognition scale capture, more properly, individuals who tend to report higher metacognitive abilities, but this does not necessarily imply that their metacognitive abilities do effectively match their self-evaluation. Indeed, a similar phenomenon has been reported for the divergence of explicit and implicit measures in other domains in Psychology, especially with dual-process theories of cognition (B. K. Payne & Bishara, 2009). It has been also hypothesised that excessive reflection might even degrade performance in a task sometimes (Dijksterhuis & Nordgren, 2006; J. W. Payne et al., 2008). A dissociation between behavioral measures and self-report measures of metacognition has been noted by previous literature (Craig et al., 2020; de Araujo et al., 2025): in some cases, participants self-reporting higher metacognitive ability actually tended to be worse with behavioral measures (Double, 2025).

Overall, our SEM analyses evidenced that several individual traits affect optimality of advice integration, both directly and indirectly though the mediation of egocentric bias and miscalibration. Indeed, other mediation paths could be considered in the light of the many possible causes of suboptimality: we focussed here on egocentric discounting and miscalibration, but others sources are possible, for instance low initial judgment accuracy. Previous data we obtained show that weak metacognitive sensitivity, overconfidence and low initial accuracy might pose a "triple burden" on advice integration, hindering optimality, and indeed we see in correlations obtained here that initial accuracy positively correlates with optimality; previous data indicate also that another source of suboptimality might lie in a *global-to-local integration deficit*, where global knowledge of the characteristics of an adviser does not necessarily lead to sufficient trial-by-trial adjustments of behavior (Zonca et al., 2025). Our research has been planned with the sample coming from two experiments, but with larger samples it will be possible to include these other sources in our SEM (described by the other endogenous indices collected with the ADT, such as confidence predictivity and accuracy of the initial decision). This in turn will allow us to extend results previously obtained, and test the effect of individual traits on multiple sources of suboptimality.

Chapter 4
Integrating Advice from Large Multimodal Models

4.1 Introduction

The multiple factors considered so far describe processes relative to the integration of advice coming from external agents, whether human or artificial. We studied the effect of stakes in modulating egocentric advice discounting, and the effect of inter-individual variability in (sub)optimality of advice integration. These processes are internal to the mind of the decision maker; we wanted then to extend our line of research in two directions. First, we wanted to deepen our understanding of how the structure of the decision context might affect performance in interacting with advisers, especially artificial ones, considering processes relative to the design of the interaction. The second direction is the interaction with Artificial Intelligence. Our research work presented so far deals with custom-built algorithms, necessary to investigate processes underlying integration of advice. We also wanted to study AI systems that have become widely adopted in recent years, through integration in our technological devices and everyday lives. We wanted to complement our view of advice taking with Human-AI Collaboration. We decided to focus particularly on Large Multimodal Models (LMMs), the latest development of the Transformer architecture (Vaswani et al., 2017) now commonly used as models of the generalist artificial agents of which ChatGPT has been the first widely available. We wanted to study interactions with the kind of agents that most people would encounter and use in their daily lives. We developed our studies following several considerations about critical points to be addressed in current literature.

Many studies in advice taking adopt the Judge-Adviser System (Sniezek & Buckley, 1995). Advice and decisions follow sequentially: the decision maker gives a first evaluation, then receives advice, then gives the final answer. However, several others choices for the *timing of AI assistance* are possible: a concurrent design, in which advice is given together with the stimuli to be evaluated; on-demand, where the advice is not given by default, but it is requested by the decision-maker if wanted; and time-delayed, where advice is given with delay in respect to the moment of asking or where the adviser acts slowly (Steyvers & Kumar, 2024). The effect of the timing of AI assistance is currently underexplored and poorly understood: some studies reported higher accuracy (Park et al., 2019) and successful debiasing (Rastogi et al., 2022) with time-delayed designs, others did not find overall accuracy improvements between sequential, time-delayed and on-demand designs (Buçinca et al., 2021). It has been argued that a sequential design might improve decision making by encouraging reflection and information gathering prior to the final decision (B. Green & Chen, 2019). However it is also possible that a sequential design might foster egocentric advice discounting: the first decision made might act as "anchor" for the subsequent final decision, so that the decision maker will

be less prone to modify the evaluation already made, insufficiently adjusting for upcoming new information. Indeed, first impressions are often difficult to change once formed and they tend to be stable over time (Wood, 2014). Instead, the concurrent design might make the advice act as "anchor" before any evaluation by the decision maker has taken place, so that acceptance of advice might be more likely. In this respect, we wanted to investigate if biases such as ecocentric discounting and automation bias might be, at least in part, due to a more general "*who-comes-first*" phenomenon: decision makers might tend to anchor to the first evaluation given, either their own or from an adviser.

Another limit encountered in current literature is that performance and reliance are often studied separately. Many studies focus on just one of the two, investigating either how AI might improve accuracy or investigating trust towards AI advice. In part, this might be a consequence of the approach, which might be AI-centric, Human-centric or symbiotic, involving several different metrics (Fragiadakis et al., 2025; Raees et al., 2024). A focus on task performance is typical in conventional Human-Machine Interaction, especially when studying automation (Parasuraman et al., 2000), whereas trust and reliance are important aspects in user studies (Bach et al., 2024; Choung et al., 2023; Kaplan et al., 2023). Our approach to Human-AI Collaboration is to consider both the Human and the AI as agents in a team, trying to understand the interplay between reliance and performance. It is possible indeed that AI advice might have different impacts on different users: some may gain small benefits from interacting with AI, whereas others might find AI advice really helpful; focussing only on final team performance or reliance might not be enough to grasp these processes.

Finally, we wanted to develop our research line following the definitions and extending the approach of (Vaccaro et al., 2024). To investigate the extent to which AI might assist human decision making and use of advice, we considered two possibilities: that the Human-AI team might outperform unassisted humans but not AI alone (*Human augmentation*), or that the Human-AI team might outperform the best of the two members taken alone, either the human or the AI (*Human-AI synergy*). Moreover, (Vaccaro et al., 2024) distinguished between the task data used as input and the required task output. Inspection of their results reveals that this distinction is most relevant, since effect sizes for augmentation and synergy greatly vary depending on the combinations of input-output data characterizing the task. For instance, tasks requiring continuous output from numeric inputs (similar to regression problems, such as sales forecasting, dosage recommendation or risk prediction) benefit from AI assistance more as compared to other tasks (such as the ones requiring open responses, or video-based), leading to pronounced human augmentation. Synergy instead might not be reached in these regression-like tasks, where AI and most statistical algorithms are usually superior, but it is visible in open-response tasks, where human opinion improves AI advice (in a form of *AI augmentation* by humans).

We decided then to develop a series of studies considering all these aspects. First, we developed an analogous taxonomy of Human-AI Collaboration designs, considering the distinction between input and output data, and the distinction between "decide" tasks (where the human takes a decision, either dichotomous, categorical or continuous) and "create" tasks (typical with Generative AI, where the human is required to write texts). Second, we considered the distinction between Human augmentation and Human-AI synergy. Third, we considered the distinction between concurrent and sequential designs. The taxonomy is represented in Figure 4.1, with tasks exemplifying the most common studies found in

literature. We are developing experiments for each row and column of the taxonomy, for now focussing on "decide" tasks. Here we present the first of the series: we built a real estate evaluation task, where participants were required to evaluate real estates in London. This first variant adopted multimedial stimuli (images and text) requiring continuous output (estimated price). Other variants will follow according to the taxonomy. Comparisons of all results will allow us to understand which combinations benefit the most from AI advice, and the degrees of performance and reliance characterizing each of them. Ultimately, we hope that the knowledge we are building will help shape future guidelines to improve the architecture of Human-AI Collaboration contexts.

TAXONOMY OF HUMAN-AI COLLABORATION DESIGNS		OUTCOME (OUTPUT)			
		DECIDE			CREATE
		DICHOTOMOUS	CATEGORICAL	CONTINUOUS	OPEN RESPONSE
DATA (INPUT)	IMAGE	Emotion recognition (Y/N)	Emotion recognition (which emotion)	Emotion recognition (intensity)	Written description
	VIDEO	Emotion recognition (Y/N)	Emotion recognition (which emotion)	Emotion recognition (intensity)	Written description
	NUMERIC	Real estate prediction	Real estate prediction	Real estate prediction	Written description
	TEXT	Misinformation (real vs fake news)	Misinformation (cluster)	Misinformation (prob)	Written description
	MULTIPLE	Multimedial	Multimedial	Multimedial	Written description

Figure 4.1: Taxonomy of Human-AI Collaboration designs. Rows describe the input data commonly found in HAIC experiments, columns describe the output required to the human participant. Tasks reported here exemplify the ones commonly found both in HAIC and Machine Learning literatures.

4.2 Hypotheses

Referring to augmentation and synergy as outlined above, in this research the AI used has better performance as compared to the human subjects, so synergy can be tested by comparing the Human-AI team performance to AI alone. We formulated the following research questions and hypotheses according to this distinction.

Research Question 1: Which agent or combination of agents performs best?
Hypothesis 1: We expect to observe human augmentation but not Human-AI synergy. AI

assistance will benefit human performance as compared to human judgment taken alone, but the Human-AI team will not perform better than the best of either of the two members.

Research Question 2: Does the performance of the Human-AI team depend on the timing of AI assistance?
Hypothesis 2: We expect to observe human augmentation but not Human-AI synergy. Different design choices for the timing of AI advice in respect to the human decision will have an effect on performance. We expect that AI advice concurrent to human decision will lead to better performance as compared to AI advice sequential to the first human decision and both designs will lead to better performance as compared to unassisted humans. However we do not expect that in any case the Human-AI team will be better than the best of either of the two members.

Research question 3: Does anchoring depend on the timing of AI assistance?
Hypothesis 3: We expect different degrees of reliance on humans' opinion on AI advice. In particular, AI advice in the concurrent design might be weighted more as compared to the same AI advice but in a sequential design.

4.3 Methods

4.3.1 Design and Variables

We designed a within-subjects experiment in three blocks, given by the three conditions of AI assistance available: *Humans Alone* (no AI assistance), *Human-AI with sequential design*, *Human-AI with concurrent design*. AI advice, when present, could be given concurrently to the human decision or sequentially, as it happens in the judge-adviser task. Thus, *timing of AI assistance* is our main independent variable. The two AI-assisted conditions are sometimes studied together in the analyses, to be compared to performance of humans taken alone; in such cases, we refer to them as the *Team* condition. To make the interaction as similar to actual use of LMMs as possible, feedback will not be given after each trial.

A total of 60 trials compose the experiment, 20 for each block. Moreover, three attention checks and two comprehension checks were inserted to control for data quality and have been used as exclusion criteria. A short familiarization at the beginning is given as an introduction. We structured the experiment in Qualtrics as follows: after the informed consent form, demographics data about Country of origin, gender and age are collected; a question assesses the familiarity with LLMs in terms of frequency of usage; the familiarization phase with comprehension checks follows; then, the experiment is run; eventually, an open-ended question asks the participants to state which features they thought to be the most relevant for their estimate according to their opinion. *Familiarity with AI* is a single-item question formulated as follows: "Are you familiar with Large Language Models (such as Gemini, ChatGPT, LeChat, DeepSeek, etc)?". Allowed answers and their values are: "No, I never use them" (1); "I use them rarely, about 1 hour per week" (2); "I use them frequently, about 1 hour per day" (3); "I use the intensely, more than 1 hour per day" (4).

The first block is always the "Human Alone" condition for all participants. This is used as a baseline to assess human ability prior to any exposure to information coming from AI. Then, the "H-AI sequential" and the "H-AI concurrent" conditions are randomly assigned to each participant as second and third block to counterbalance possible order effects that might happen. We consider two main dependent variables. For each set of images, we can measure the distance between estimated price and ground truth price: the narrower this distance, the better the performance. We can also measure the distance between the estimate given by humans and the estimate given by AI: the narrower this distance, the stronger the reliance on the adviser. The first variable quantifies the precision of the estimate, whereas the second quantifies the degree of overlap between human estimate and AI estimate (the extent of their agreement).

4.3.2 Dataset

We built our dataset by manually scraping descriptions of real estates located in London from Zoopla, a website for house rentals in the United Kingdom. The dataset is composed by 66 sets of four images of the interiors (living room, bathroom, bedroom, kitchen) with textual and numeric descriptions of the house properties (always including at least the number of rooms, the total area available, the borough in London where the house is located, but not limited to this information). 3 sets have been used to familiarize the participants with the task, 3 sets have been used as attention checks, the others were used as experimental stimuli. The selected real estates have a ground truth *asking price*, hidden to participants, ranging from £250.000 to £750.000. The asking price is the price that the seller states they are willing to accept, and we consider it as the ground truth to assess performance. The task for the participant is to estimate the most likely asking price for the real estate in each trial given all the information available, in a range from £0 to £1 million Great Britain Pounds.

4.3.3 The AI-LMM adviser

Assistance in the task was provided by a Large Multimodal Model, an AI conversational and reasoning model based on the Transformer architecture (Vaswani et al., 2017) and adapted for text and image prompts. A preliminary study was conducted to choose between ChatGPT4.1 and Gemini 2.0 Flash, two of the most well-known LMMs available, to understand which of the two could provide the best advice, and preliminary results led us to choose the latter. The LMM was accessed with an API coded in Python 3.13 with Visual Studio Code. The API allows to quickly loop over all stimuli obtaining the evaluations of all the real estates: the code takes the images and the textual description for each set of images and combines them in a single package with a prompt, then saves the model answers in a vector containing the estimates for each set. These answers constitute the *AI alone* condition that we use to study augmentation and synergy, according to our research questions outlined above. We prompted the model with the following string:

I need your assistance for a real estate evaluation task. I upload a document containing information about a real estate in London, United Kingdom, with four pictures of the interiors: living room, kitchen, bathroom, bedroom. A description of the property features is provided

in the document, together with the pictures. I ask you to estimate the asking price for this property in British pounds, in a range from £0 up to £1 million pounds. Please consider all the information available and give your estimate of the most likely asking price. Give a point estimate and not a range. Be concise in your output, keeping it below 100 words. Organize your answer by stating your evaluation first (highlight the price in bold and state "My evaluation for this property is..."), then your arguments for the evaluation in a bullet point list containing your considerations relative to: location, property features, market comparables.

Finally, for each set of real estates, the images, the textual description and the AI advice (where needed, and according to the type of block) were combined in a graphical layout containing all the information for a given trial. These vignettes were used as experimental stimuli for the study with human participants. In the trials of the concurrent advice block, the vignette contained and presented the information all at once (Fig. 4.2). In the sequential advice block, first a vignette with images and description was shown, and the first estimate was relative to this information only; then, AI advice was presented, and the final estimate was to be given by the participants. In the trials for humans alone, no AI advice was present, and the vignettes with only the images and the textual descriptions were shown.

4.3.4 Sample size, participants and recruitment

A simulation-based power analysis has been run to determine the minimum required sample size by using the *SimR* package for R (P. Green & MacLeod, 2016). According to previous literature (Vaccaro et al., 2024), human performance might be augmented by AI assistance with a mean effect size of 0.64 in terms of Hedge's g. More specifically, we are interested not only in the general possibility of human augmentation, but also in the effect that concurrent or sequential paradigms might have. We reasonably expect that this effect size will be smaller as compared to the one comparing unassisted and assisted humans. We chose a conservative target effect size of Cohen's $d = 0.2$. The following parameters were set fixed for the simulation: 20 trials for each condition, totalling 60 trials for our repeated measures design; Cohen's d equal to 0.2 as effect size for the difference between means; alpha criterion of 0.05; minimum desired statistical power equal to 0.9. Available data on real estate sells were used to generate a synthetic dataset of "ground truth estimates", then we generated a column of "estimates" and consequently of "distances" between the two, with a distance between ground truth and mean estimates that was highest for unassisted humans, lower for the sequential advice condition and lowest for the concurrent advice condition, according to our hypotheses. Our power analysis has been run for 10000 simulations; achieved power was measured with the percentage of simulations that reached the significance threshold we chose as p-value; this percentage in turn was used to compute the minimum sample size required to reach an achieved power of at least the desired power of 0.9 we specified. With these settings, the minimum required sample size was 38 participants.

We decided to collect two samples: one in the United Kingdom and another in Germany, so as to sample human evaluations from countries where people are accustomed to London real estate prices to a different extent. The UK sample was recruited on Prolific, whereas the German sample is still under recruitment in the LMU in Munich and in Regensburg University. For this reason, only results from the completed UK data collection will be

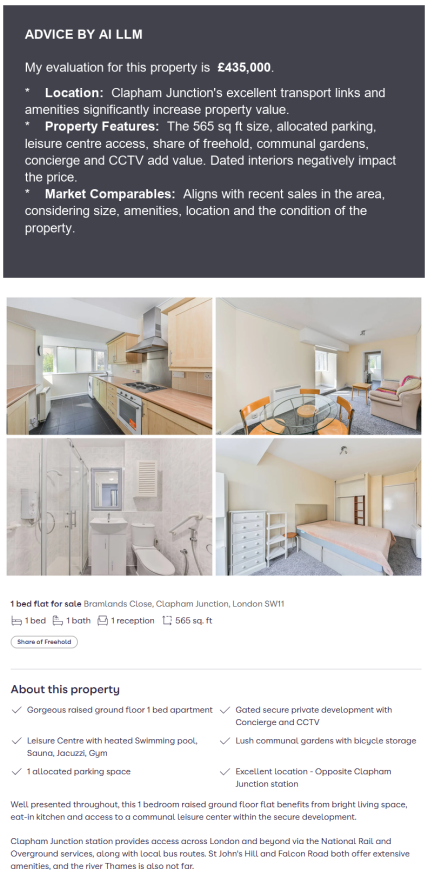

Figure 4.2: One of the vignettes used for the Human-AI *Concurrent* trials, exemplifying the composition of experimental stimuli. The advice by AI was written in a graphite-grey box, followed by the layout of images and by numeric and textual information about the real estate. For the Human-AI *Sequential* trials, the image layout and the real estate written information were presented in the initial page, and AI advice in the second page, after the first decision by the participant. For the Human Alone trials, only images and written information were presented.

presented and discussed here.

4.4 Analyses

4.4.1 Preliminary analyses

A first set of correlations was computed to assess two preliminary questions: 1) Is AI capable of estimating prices with a precision suitable for the experiment? 2) Which LMM to choose as assistant? We computed the correlation of AI-Gemini estimates with ground truth prices, and of AI-GPT estimates with ground truth prices. Similarity between the two model estimates was also assessed with their correlation. Outliers in AI estimates were obtained for both models, together with standard deviations and the 95% confidence intervals. t-tests were

computed to assess if the estimates of the two LMMs were statistically different and if they were statistically different from the ground truth. A non-significant t-test means that the two distributions are sufficiently similar so that a linear test cannot find a linear decision boundary to separate them. Results of these preliminary analyses guided us in choosing which LMM we would have adopted as assistant for our experiment.

4.4.2 Distances

We conceptualize *distance* as a continuous function on the set of positive real numbers, as it is done with metric spaces. Consequently, all distances are absolute values and they can be used to quantify the precision of the estimates around the ground truth value (d_P in the case of performance) or to quantify the overlap of estimates between humans and AI (d_A in the case of anchoring). Equal elements have by definition a null distance. So we generally mean "distance" as *absolute distance*. Wherever needed, we specify that we are referring to the *signed distance*, meaning that we consider the sign either positive or negative, and it will be used to quantify overestimates or underestimates. The distance d_P quantifies performance for a trial: shorter distances indicate more precise estimates, nearer to the ground truth asking price. The distance d_A quantifies reliance on AI advice: shorter distances indicate that human estimates and AI estimates are more similar. Theoretically, without any prior assumption, overlap between estimates and anchoring are different processes: it is possible that humans and AI might produce very similar estimates without the human being actually influenced by AI, and in that case it would be improper to refer to this distance as "anchoring". However, as we specify below in our research questions, we expected that human estimates will be affected by the exposure to the AI estimate and we tested it, and consequently we prefer adopting the term "anchoring distance", incorporating our assumption in the definition.

4.4.3 Research Question 1

We grouped together the trials of the *Human-AI concurrent* block and of the *Human-AI sequential block* to obtain a *Human-AI Team* condition, to be compared against the trials without AI assistance and to the estimates of AI alone. As a first step, distances between estimates and ground truth in the various experimental conditions can be compared. We will compare AI alone to the mean performance of our human sample and to the performance realized by the best human estimator in the sample. A t-test can be then used to determine if the distributions of estimates are statistically distinguishable. Finally, we estimated a linear mixed model with distances as dependent variable, random intercept at participant level and AI assistance condition C (absent vs. Human-AI team) as independent variable, where subjects S were considered as clustering factor for the random intercept:

$$d_P \sim C + (1|S)$$

To understand the dynamics of Human-AI Collaboration more deeply, we used clustering to analyze tendencies in estimates and if they varied with consistency between subjects. k-means clustering is an unsupervised learning method for finding clusters in unlabeled data (Hastie et al., 2009); starting from a predefined number of centroids randomly positioned, the method

iteratively moves their position in the data to reduce the total variance within clusters. First, it selects the subset of data points which are the nearest to each centroid, then it computes the means for each subset and uses it as the position of the new centroids; the two steps are iterated until convergence is reached. We used k-means clustering to study consistent tendencies in the difference of means of the *Human Alone* condition as compared to the *Human-AI Team* condition. We explored the optimal number of clusters visually with the elbow method (which works analogously to a scree test for Principal Component Analysis), and quantitatively with the Silhouette scores, considering different cluster arrangements at qualitative inspection. Finally, we investigated if variables describing individual characteristics could predict cluster classification of participants. We estimated a logistic regression model with cluster classification as dependent variable, and with age, education level, familiarity with AI, years lived in London as independent variables.

4.4.4 Research Question 2

We computed the distances between human estimate and ground truth price in all conditions (*Humans Alone, H-AI concurrent, H-AI sequential*) and also considered the estimates by AI alone. We also computed the standard deviations of the estimates for each experimental condition to compare their ranges. We estimated a linear mixed model with the distances as dependent variable and with AI advice timing T (absent, concurrent, sequential) as independent variable, where subjects S were considered as clustering factor for the random intercept:

$$d_P \sim T + (1|S)$$

And we analogously conducted a cluster analysis to understand if we could find consistent tendencies in how AI might have impacted human estimates.

4.4.5 Research Question 3

We quantified anchoring as the degree of overlap between human opinion and AI advice with the distance between their estimates, defined as a positive real-valued function. We quantify our expectation as:

$$d_A(HAIconc) < d_A(HAIseq) < d_A(Halone)$$

Meaning that 1) we expect anchoring to be stronger for the concurrent design condition, and 2) distance between human-alone estimate and AI-alone estimate will be larger as compared to either of the two Human-AI conditions. Consequently, an anchoring distance near zero means strong anchoring. Analogously to RQ2, we estimated a linear mixed model with the anchoring distances as dependent variable and with AI advice timing T (absent, concurrent, sequential) as independent variable, where subjects S were considered as clustering factor for the random intercept:

$$d_A \sim T + (1|S)$$

And we analogously conducted a cluster analysis to understand if clusters in anchoring behavior and reliance on AI are present. Finally, we investigated if variables describing individual characteristics could predict cluster classification of participants in anchoring behavior, by estimating a logistic regression model analogous to the one used in RQ1.

4.4.6 Cluster overlap

Results of cluster analyses for RQ1 and RQ3 can be correlated to understand if there are consistent patterns in performance and reliance between clusters. Overlap of the cluster structures from RQ1-RQ3 was also quantified with the Adjusted Rand Index (Rand, 1971; Vinh et al., 2010), an information-theoretic measure ranging from -1 to $+1$, where $+1$ indicates perfect agreement. These indices, if significant, would indicate that performance in estimates and anchoring behavior might be linked, shedding light on the possible process leading to fruitful Human-AI collaboration. We also quantified the *performance benefit* as the distance, for each subject, between performance when *Alone* and when assisted by AI in *Team*, whether concurrent or sequential. This distance indicates that a *performance gap* when *Alone* is compensated to some extent by integrating AI advice in *Team*. We used this index to understand if cluster patterns in reliance and performance were overlapped: a significant positive regression effect of performance benefit on anchoring distance would indicate that subjects whose initial estimates are the most distant from AI advice are also the ones potentially benefiting the most from it: reliance would be higher for subjects with larger gaps in performance, indicating advice integration to a larger extent and leading to higher gains from adopting AI advice.

4.4.7 Secondary analysis: Condition Order Effects

To assess the possibility that AI advice might have a different impact depending on the order of conditions encountered in the experiment, we consider the model from RQ1 and add a *Block* variable (B) describing the order in which the participant went thought the trials (0: Sequential first, then Concurrent; 1: Concurrent first, then Sequential), considering also the interaction of *Condition by Block Order*. The model equation consequently has the form:

$$d_P \sim C + B + C \cdot B + (1|S)$$

4.5 Results

4.5.1 Preliminary selection of AI-LMM assistant

Initially, both ChatGPT4.1 and Gemini Flash 2.0 were shown the experimental stimuli and asked to give their own estimate of the most likely asking price. Their answers are plotted in Fig. 4.3 together with the 95% confidence interval for each model. The magnitude of Cohen's d can be used as an index of separability of the two distributions: larger d values indicate that estimates in the two considered conditions are more distant as compared to

what happens with smaller d values. The correlations, the t-tests and the effect sizes with Cohen's d between the distributions involved are shown in Table 4.1 and Table 4.2.

Both models performed comparably well in the task, estimating the price with similar precision. Both model estimate distributions are not statistically different from Ground Truth prices, and there is no statistically significant difference between them. The mean estimate by ChatGPT4.1 was £520917 and by Gemini 2.0 Flash was £511417, to be compared to the mean of ground truth prices equal to £490000, revealing a tendency to overestimate for both models. The mean absolute distance from ground truth was £69917 for ChatGPT4.1 and £66250 for Gemini 2.0 Flash, with a ratio of 1.056 between them, indicating an overall 5.6% overestimate by ChatGPT over Gemini and a slightly better precision by the latter. ChatGPT4.1 estimates had a larger standard deviation as compared to Gemini 2.0 Flash (103028 vs. 82231). The Quantile-Quantile plot showed one outlier measure by ChatGPT, an estimate of £950000 for a real estate valued £400000. We thus decided to keep Gemini ad preferred LMM to generate advice: the absence of outliers, the narrower variance, the slightly more precise and more homogeneous distribution of estimates, together with the fact that Gemini is freely accessible with limited API calls per day whereas ChatGPT requires paid access, led us to prefer Gemini as AI adviser.

Table 4.1: Correlation between AI estimates distributions.

Distributions		r **value**	p **value**
GPT estimates	Ground Truth	$r(58) = 0.80$	$p < .001$
Gemini estimates	Ground Truth	$r(58) = 0.87$	$p < .001$
GPT estimates	Gemini estimates	$r(58) = 0.95$	$p < .001$

Table 4.2: Welch's t-test and Cohen's d for the AI estimates distributions.

Distributions		r **value**	p **value**	d
GPT estimates	Ground Truth	$t(115.79) = 1.049$	$p = 0.297$	0.185
Gemini estimates	Ground Truth	$t(117.09) = 0.747$	$p = 0.456$	0.133
GPT estimates	Gemini estimates	$t(117.7) = -0.31$	$p = 0.76$	-0.055

4.5.2 Humans Alone vs Gemini

As a starting point to investigate human ability, we studied the estimates of humans alone and their distances from the ground truth, to compare them to the performance of AI alone. Results are plotted graphically in Fig. 4.4.

Overall, humans alone performed worse than AI alone. They tended to overestimate real estates in the lower end of the price distribution, to a greater extent than AI does; they tend also to underestimate the real estates in the higher end of the price distribution, to a much larger extent than AI does and for a much larger interval of prices. Variance of their estimates is also larger as compared to AI, almost as double (161144 vs. 82231). An overall

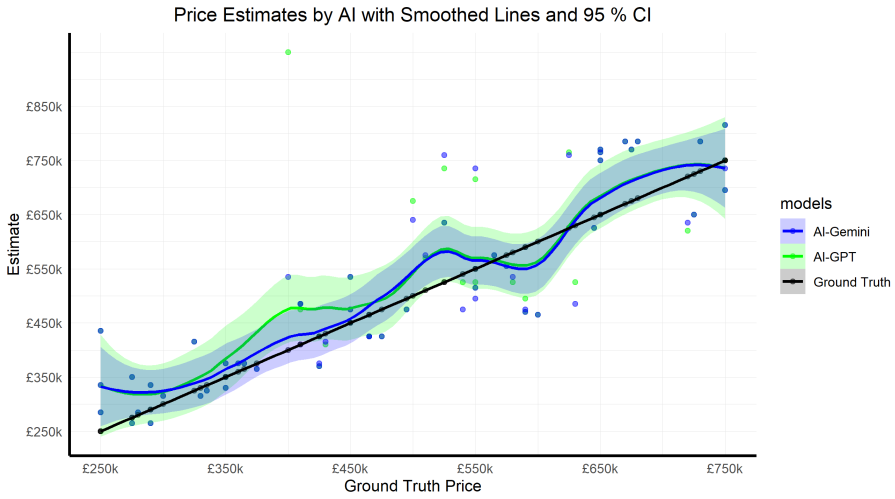

Figure 4.3: Price estimates by ChatGPT and Gemini in respect to the Ground Truth asking price, plotted with their 95% confidence intervals.

tendency to underestimate is present (with mean estimates of £461938 as compared to the mean ground truth of £490000) and stronger than the tendency of AI to overestimate.

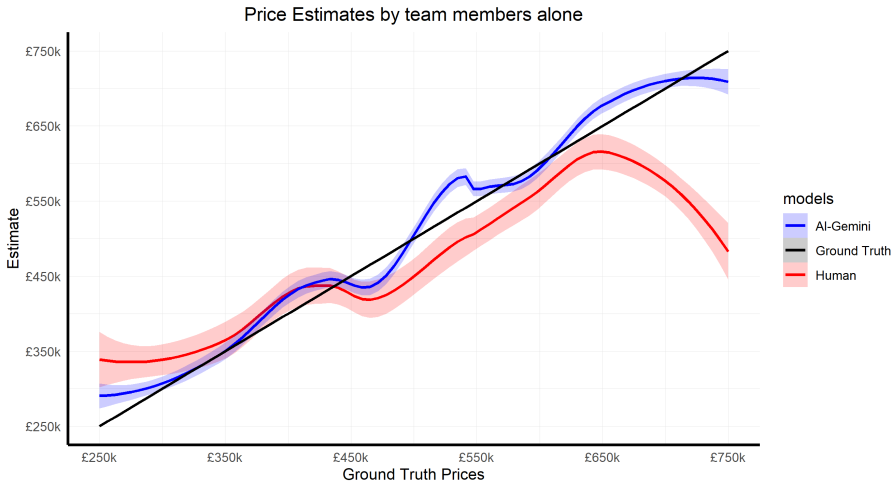

Figure 4.4: Humans Alone vs. AI alone: estimates as compared to ground truth.

4.5.3 Research Question 1

Distances

Our results support the hypothesis we formulated for our Research Question 1. We have evidence of human-AI augmentation: performance of humans assisted by AI (mean distance from ground truth: £81970) is better as compared to performance of humans alone (mean distance from ground truth: £125411). However human-AI synergy was not reached: mean

distance by AI alone was much lower (£66250). This means that AI advice improves performance for humans, but AI alone still remains the best estimator (see Fig. 4.5).

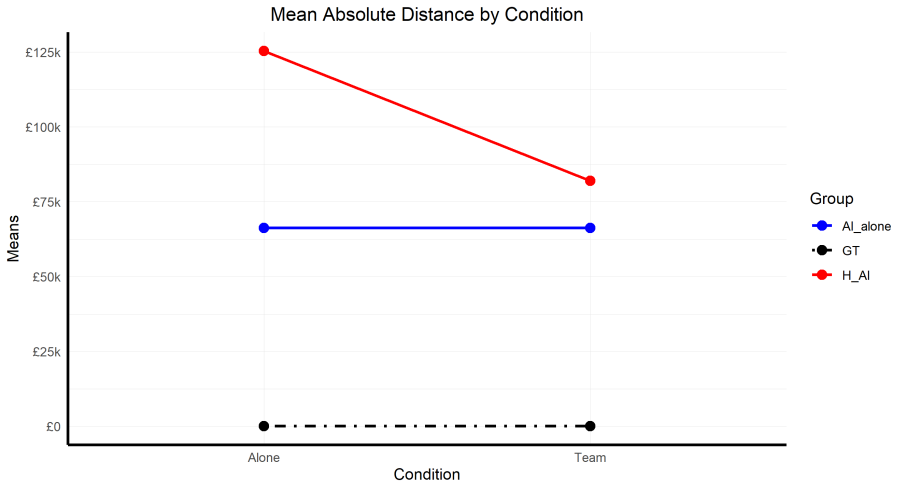

Figure 4.5: Results for our Research Question 1: distances of estimates from ground truth without AI assistance (Alone) and with AI assistance (concurrent and sequential design grouped together as Team). Performance of human subjects (H_AI) is shown in red, whereas AI alone is plotted in blue as a reference line to evaluate augmentation and synergy. Distances can be evaluated referring to the Ground Truth (GT) dashed line in black (centered at zero).

We also computed the mean absolute distance for each human subject, considering all AI assistance conditions, either alone or with AI advice, and we extracted the minimum obtained, quantifying the performance of the best human subject in our task, to compare it to performance of AI alone. Results are presented in Table 4.3. Consider that these results are useful to compare the best performer available, but are not indicative of augmentation or synergy, which are measured at the level of the whole sample.

Table 4.3: Best performance reached by agents or combinations of agents in our study.

Agents	**Distance (£)**
Best Human Alone	51500
Best Human-AI team (concurrent)	139211
Best Human-AI team (sequential)	110500
AI alone	66250

***t*-test**

Distributions of estimates have been compared against the ground truth estimate distribution. We use Cohen's d as a further index for the separability of distributions, with lower values indicating better performance. Results are shown in Table 4.4.

Table 4.4: t-tests for distributions of estimates of human participants in two AI assistance conditions (no AI and with AI).

Distributions		r value	p value	d
Humans Alone	Ground Truth	$t(1157.9) = -4.364$	$p < .001$	0.176
Gemini estimates	Ground Truth	$t(3062.2) = 3.140$	$p = .0017$	0.098
GPT estimates	Gemini estimates	$t(1535.4) = -6.063$	$p < .001$	0.267
AI Alone	Humans Alone	$t(6477.9) = 4.992$	$p < .001$	0.123

Linear mixed model

The fixed effect of AI condition on performance distance was significant: $t(2382.01) = 8.226, p < .001$. The statistically significant fixed intercept indicates the presence of a baseline effect of AI condition (which is "no AI assistance"): $t(66.95) = -3.796, p < .001$. The Intraclass Correlation Coefficient of 0.09 indicates that the random structure helps explain further variance as compared to fixed effects alone, supporting our choice of a model with random structure. The graphical plot of results with fixed effects of group means and random effects at participant level is shown in Figure 4.6.

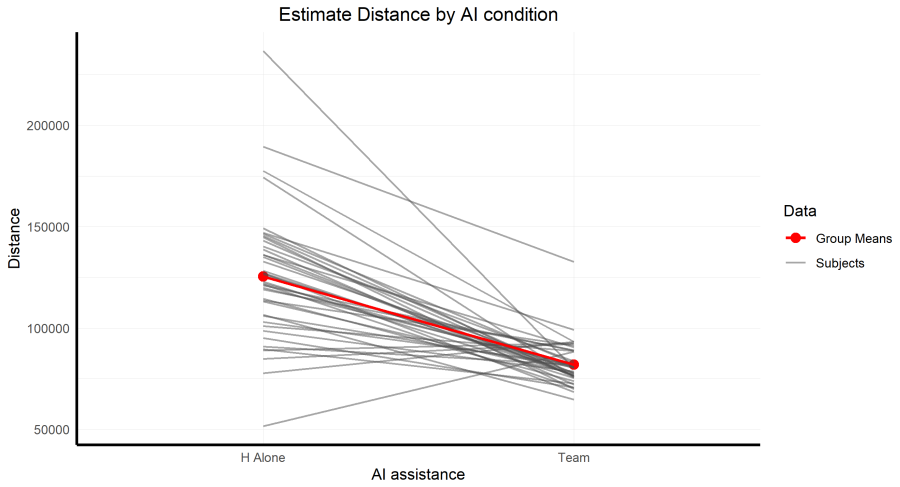

Figure 4.6: Distance as a function of AI assistance condition (*Humans Alone* vs Human-AI *Team*), with subjects as random intercept (grey lines). Fixed effects ("group means") are plotted with the solid red line.

Clustering

The Elbow method for choosing the optimal number of clusters led to inconclusive results. The Silhouette method indicated a solution with 2 or 5 clusters (average Silhouette width of 0.52 for $k = 2$ and of 0.51 for $k = 5$). We report here results for two cluster analyses, the first with two clusters and the second with 5 clusters, and a visual inspection can be done with Figure 4.7. While the first cluster analysis should be preferred, as it is indicated as such by the quantitative Silhouette score, the analysis with 5 clusters reveals a qualitatively interesting trend.

Mainly, two clusters emerged. Participants in Cluster 1 seem to benefit from AI advice to a limited extent, because their estimates when doing the task alone are already better as compared to participants in Cluster 2. These subjects seem to already have a good estimating ability, at least compared to the other humans in our sample, their estimates when assisted by AI do not vary as much, and consequently their benefit from AI is present but lower. Subjects in Cluster 2 instead perform worse when unassisted by AI, and the benefit by AI assistance is comparatively higher, so that the subjects in the two clusters perform very differently when unassisted but equally well when assisted by AI. This is confirmed by two t-tests comparing distances between clusters: *Humans Alone*, Cluster 1 vs Cluster 2 $[t(38.185) = -6.3814, p < .001]$ and *Human-AI team*, Cluster 1 vs. Cluster 2 $[t(39) = 1.274, p = 0.21]$. Qualitative inspection of the solution with 5 clusters, albeit not the preferred solution from the Silhouette analysis, showed a graded group of clusters where AI assistance gave increasing benefits for performance, and a single cluster where instead AI advice led to worse performance. The estimate distribution of this cluster did statistically differ from the estimates of the cluster for which AI benefit was present but lowest (Cluster 1 vs Cluster 3 estimates for *Humans Alone*, $t(3.96) = -3.191, p = 0.033, \alpha = .05$]. Cluster means can be numerically inspected in Table 4.5 and Table 4.6.

Variables related to demographics and individual characteristics were used in a logistic regression, with age, education level, years lived in London and familiarity with AI as independent variables, and with Cluster (k=2) as dependent variable. The model was not significant: $F(5, 34) = 0.428, p = 0.828$. Given that logistic regression could be used as a classification algorithm, it seems that none of the independent variables can be considered significant features to classify participants into the discovered clusters for performance.

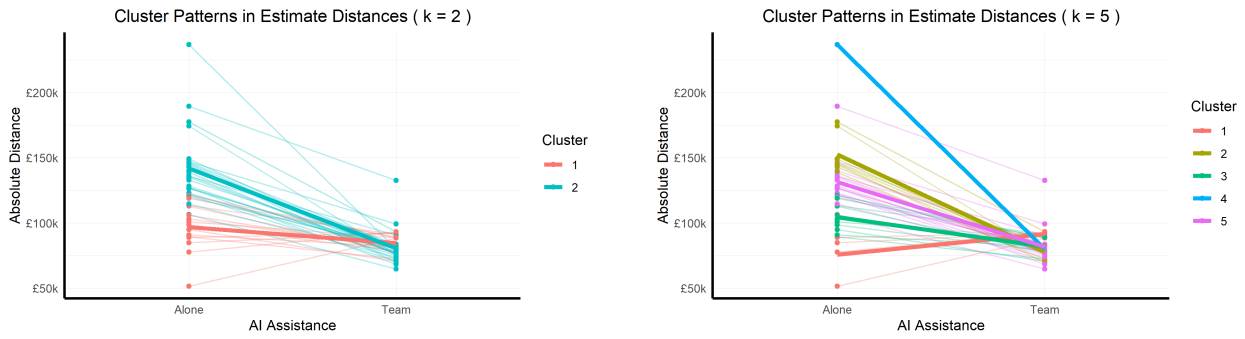

Figure 4.7: Cluster analysis for performance (Humans Alone Vs. Human-AI Team); analysis with two clusters plotted on the left, with 5 clusters plotted on the right.

Table 4.5: Cluster means for the cluster analysis of RQ1 at k=2.

Cluster	Humans Alone	H-AI Team
1	96883	84544
2	141869	80484

Table 4.6: Cluster means for the cluster analysis of RQ1 at k=5.

Cluster	Humans Alone	H-AI Team
1	75750	91593
2	152438	77868
3	104568	81981
4	236700	80000
5	131317	81744

4.5.4 Research Question 2

Distances

Results supported our Hypothesis 2 only partially: mean distances in the Human-AI *Team* condition are lower as compared to the mean distances of estimates by unassisted humans *Alone*, but the H-AI *concurrent* condition (mean distance: £84802) does not lead to better estimate performance as compared to the H-AI *sequential* condition ("sequential 2", mean distance: £79004), as we instead initially predicted. Mean distances of the initial decision in the sequential condition ("sequential 1", mean distance: £125117) are comparable to the distances in the unassisted condition (mean distance: £125411), as expected, given that both estimates were given by humans alone without AI assistance. The standard deviations of the distances are lower in the AI-assisted condition: $SD(H : Alone) = 32469.6; SD(H - AIconcurrent) = 14736.24; SD(H - AIsequential) = 14473.1$. Overall, distances show evidence of Human augmentation but not of Human-AI synergy, extending the conclusions we saw for our RQ1. Mean distances are plotted in Fig. 4.8 for graphical inspection.

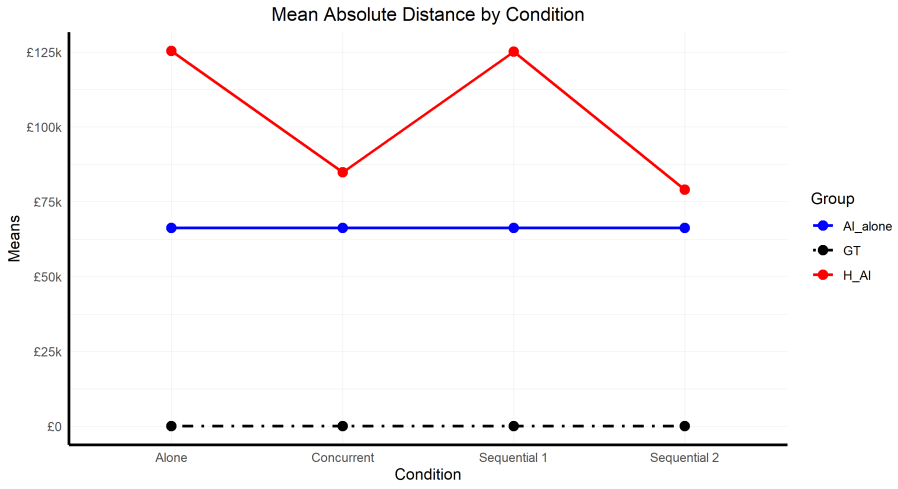

Figure 4.8: Mean distances for all our experimental conditions. The red line represents data from our experiment with human subjects. The blue line represents the performance of AI alone as a reference to evaluate human-AI synergy. Distances are referred to the Ground Truth (black dashed line), centered at zero.

Linear Mixed Model

Analyses of the distances considering the random structure at participant level gave us a more detailed view, and can be inspected in Fig. 4.9. In particular, both AI-assisted conditions led to significantly better estimates as compared to the unassisted condition. The comparison between *Humans Alone* and *Human-AI concurrent* was significant: $t(2382) = 9.956, p < .001$. The comparison between *Humans Alone* and *Human-AI sequential* was significant: $t(2382) = 11.242, p < .001$. However the comparison of *Human-AI concurrent* and *Human-AI sequential* was not significant: $t(2382) = 1.402, p = 0340$.

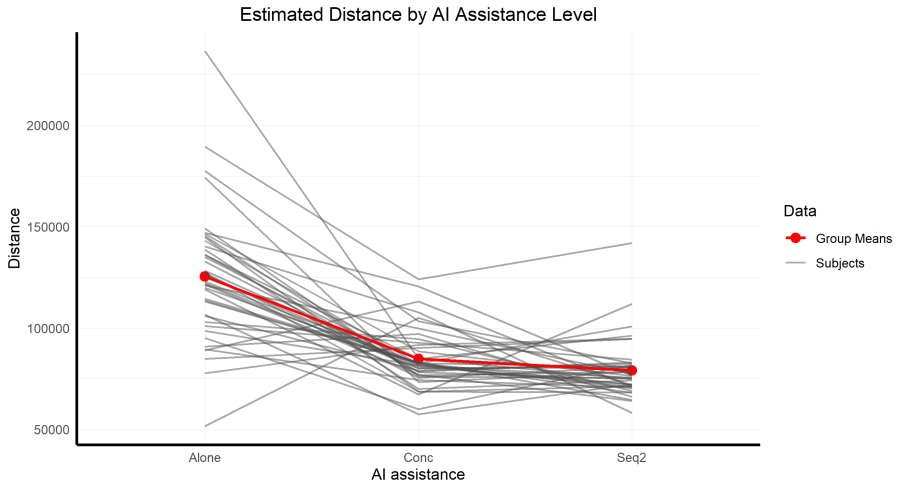

Figure 4.9: Distance as a function of AI assistance condition (*Humans Alone, Human-AI concurrent, Human-AI sequential*), with subjects as random intercept (grey lines). Fixed effects ("group means") are plotted with the solid red line.

Cluster Analysis

The Elbow method led to nondecisive results, whereas Silhouette analysis indicated k=5 as the optimal number of clusters. In accordance with the analyses conducted previously, we still plotted the k=2 solution even if suboptimal in this case, as a reference to RQ1 and to see which insights could be gained from the optimal solution (Fig. 4.10). The solution with k=5 evidences widely diverging patterns of behavior across subjects: participants in clusters 2, 4, 5 behave accordingly to the main fixed effect that the linear mixed model showed, and they differ mainly in estimates when unassisted by AI, converging towards more similar means when assisted by AI; participants in cluster 3 instead behave accordingly to what our initial hypothesis supposed for our RQ2; finally, participants in cluster 1 perform the best in the *H-AI sequential* condition, and comparably worse when *Alone* as when they are assisted in the *H-AI concurrent* condition. Results are summarized in Table 4.7 below. Observe that the suboptimal analysis with k=2 could not capture these divergent patterns.

4.5.5 Research Question 3

Distances

Analyses of mean distances between Humans and AI bring support our hypothesis for RQ3:

Table 4.7: Cluster means for the cluster analysis of RQ2 at k=5.

Cluster	Humans Alone	H-AI concurrent	H-AI sequential
1	103583	106300	68757
2	150431	79885	73351
3	189500	124000	141842
4	106568	77600	79761
5	120300	82300	85467

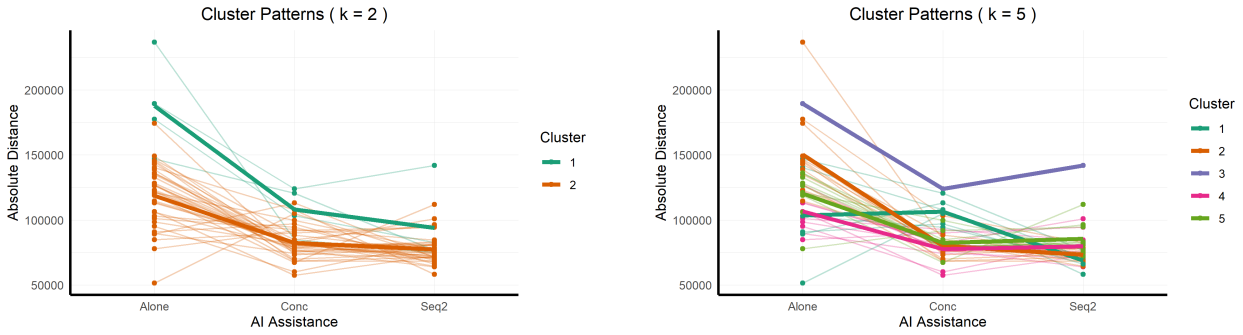

Figure 4.10: Cluster analysis for performance (considering the conditions Humans Alone, H-AI concurrent, H-AI sequential); analysis with two clusters plotted on the left, with 5 clusters plotted on the right.

mean distance in the *H-AI concurrent* condition (£39241) is lower as compared to the *H-AI sequential* condition (£44342), and both are lower as compared to the distance between the two agents alone (£128771). Only the distances in the two *H-AI* conditions could be properly considered "anchoring", since they are the only two conditions in which human participants are exposed to AI opinion; however we computed the distance also for the *Humans Alone* condition as a reference for the distance between the opinions of the two agents prior to their interaction. Results are plotted in Fig. 4.11 for visual inspection. Differently from what we saw for RQ2 and distances in performance, here distances between estimates across conditions tend to have comparable standard deviations: $SD(H : Alone) = 32586; SD(H - AIconcurrent) = 31787.45; SD(H - AIsequential) = 25901.9$.

Linear Mixed Model

Analyses of the distances considering the random structure at participant level gave us a more detailed view as compared to comparing the means directly, and can be inspected in Fig. 4.12. In particular, both AI-assisted conditions led to comparable anchoring distances, indicating a very similar overlap between the human estimate and the AI estimate in both designs; crucially, they were both much lower as compared to the reference distance of both agents considered Alone, indicating a substantial convergence of human opinion towards AI opinion when the two interacted. The comparison of anchoring distances between *Humans Alone* and *Human-AI concurrent* was significant: $t(2382) = 21.997, p < .001$. The comparison

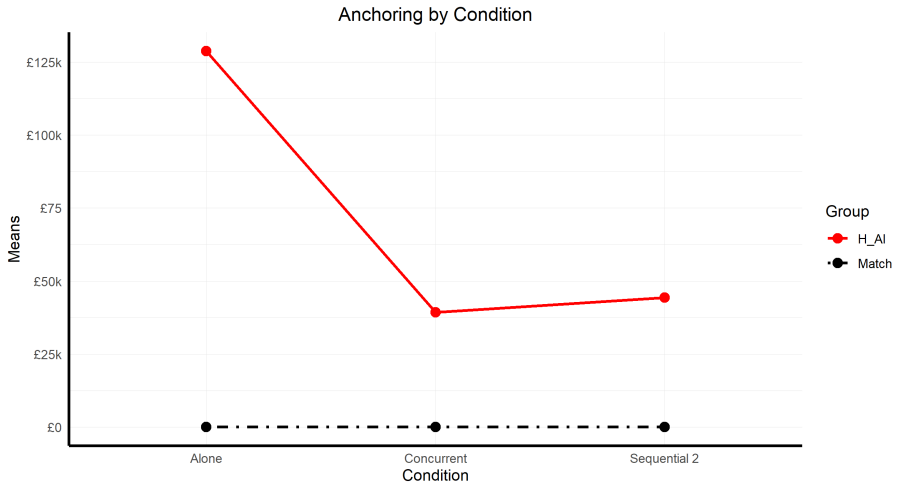

Figure 4.11: Results for our Research Question 3: distances of estimates between human participants and AI in the experimental conditions. Observed distance is shown in red, whereas the possibility of exactly matched opinions (mean human estimate coinciding with AI estimate) is plotted with the black dashed line as a reference.

between *Humans Alone* and *Human-AI sequential* was significant: $t(2382) = 20.495, p < .001$. However the comparison of *Human-AI concurrent* and *Human-AI sequential* was not significant: $t(2382) = -1.246, p = 0.426$.

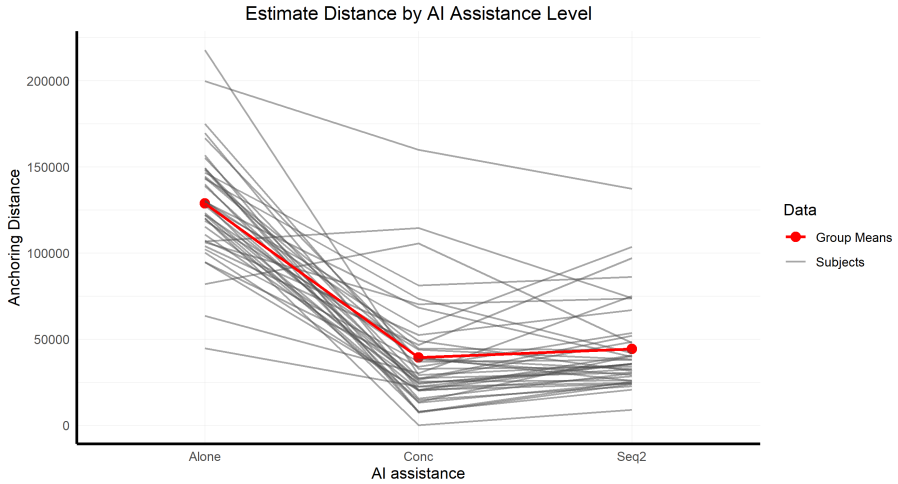

Figure 4.12: Anchoring distance as a function of AI assistance condition (Humans Alone, Human-AI concurrent, Human-AI sequential), with subjects as random intercept (grey lines). Fixed effects ("group means") are plotted with the solid red line.

Cluster Analysis

The Silhouette method indicated k=2 as the optimal number of clusters. Mean distances are reported in Table 4.8 and plotted in Fig. 4.13 together with lines for the random structure emerged from the mixed model. From the cluster analysis two clusters emerged clearly: Cluster

1 seems less affected by anchoring to AI opinion, with mean distances for each subject between conditions varying less as compared to what happens for Cluster 2. The stronger anchoring of Cluster 2 as compared to Cluster 1 is also indicated by its SDs in the two H-AI conditions, indicating both a stronger overlap between human estimate and AI estimate and less variation around that overlap. Indeed, t-tests confirmed that anchoring distances do not statistically differ between clusters in the *Human Alone* condition $[t(10.776) = -1.492, p = 0.275]$, but they do statistically differ in the *Human-AI concurrent* condition $[t(8.738) = 3.753, p < .01]$ and in the *H-AI sequential* condition $[t(8.623) = 5.91, p < .01]$.

Variables related to demographics and individual characteristics were used in a logistic regression, with age, education level, years lived in London and familiarity with AI as independent variables, and with Cluster ($k = 2$) as dependent variable, analogously to what it has been done for RQ1. The only significant predictor was *Familiarity with AI*: the estimate effect of $z = -1.56, p < .05$ indicates that higher familiarity with AI is associated with lower probability of being in Cluster 2, where anchoring is stronger (as compared to Cluster 1, where anchoring is less intense). Quantitatively, for each one-unit increase in *Familiarity_ AI*, the odds of being in Cluster 2 (vs. Cluster 1) are multiplied by $exp(-1.56) = 0.210$, implying a decrease in probability by about 80% to belong to Cluster 2. Area Under Curve is equal to 0.79, indicating significant classification accuracy and decision boundary distinction between the two clusters.

Table 4.8: Cluster means and SDs for the cluster analysis of RQ2 at k=2.

Cluster	Humans Alone	H-AI concurrent	H-AI sequential
Means			
1	116056	79789	84641
2	132347	27838	32992
SDs			
1	39406.69	40621.98	25727.4
2	30153.63	16321.44	9507.66

4.5.6 Cluster overlaps

We finally took the cluster structure with k=2 emerging from RQ1 and RQ3 and computed the Spearman rank correlation, obtaining a statistically significant correlation coefficient: $r = 0.454, p < .01$, indicating a positive monotonic relationship between the two cluster structures. The ARI indicates a partial overlap of the cluster structures, with $ARI = 0.232$.

Comparing the clusters side-by-side reveals the interplay between reliance and performance though anchoring. Cluster classification from RQ1 ($k = 2$) and from RQ3 ($k = 2$) are plotted here in Fig 4.14, where we considered trials of Humans *Alone* compared to trials with AI assistance as *Team*.

Cluster structure from RQ3 can be used to study performances in estimates. Cluster 2 anchors more to AI advice, receiving a greater *performance benefit from AI assistance*. If we

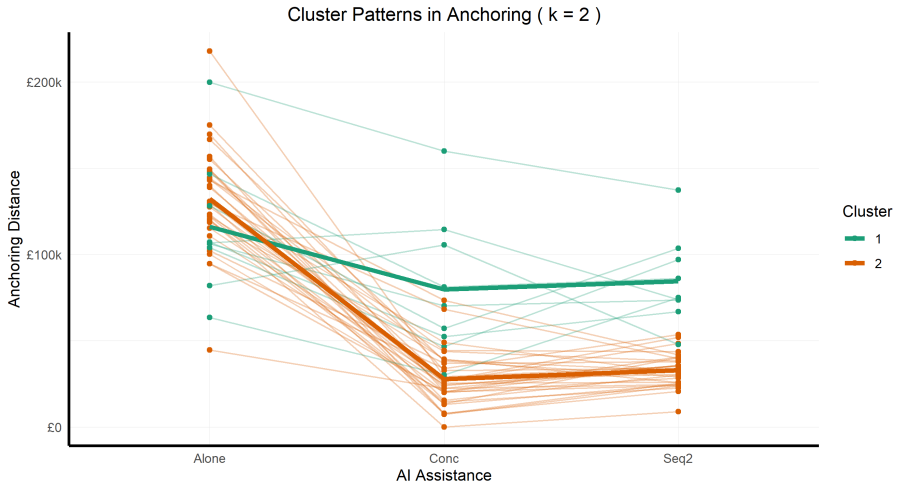

Figure 4.13: Cluster patterns in anchoring for our experimental conditions

compare the mean difference between *Alone* estimates and *Team* estimates (see Fig. 4.15), we see that the difference is higher for Cluster 2 (mean difference: £50892) as compared to Cluster 1 (mean difference: £16949). Given that *Team* performance does not differ significantly between clusters, this implies that Cluster 2 receives a performance benefit by adhering more to AI advice, compensating the performance gap when *Alone*. A t-test comparing the two distributions is statistically significant $[t(16.264) = -3.462, p < .01]$, confirming that the two clusters experience a different performance benefit while interacting with AI.

Regression of performance distances over anchoring distances was significant: $F(1, 39) = 95.32, p < .001$, indicating that participants whose estimates *Alone* were significantly worse were also the ones whose initial estimates were more distant from AI estimates, and they were consistently from Cluster 2 (from the clusters of RQ3), who were the ones integrating and relying more on more the AI advice, gaining a high performance benefit. In summary, participants whose initial estimates were more distant from ground truth and from AI estimates tended also to be the ones following AI advice more, compensating their initial performance gap. Regression results depending on clusters can be inspected in Fig. 4.16.

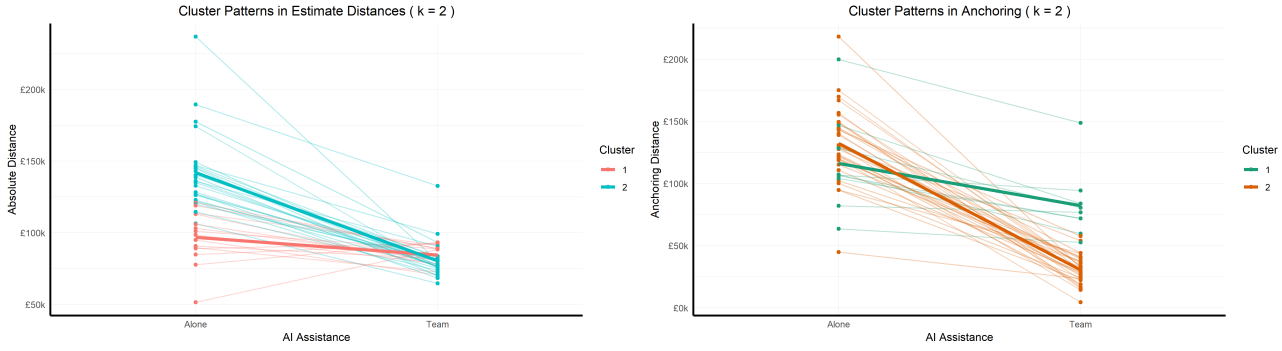

Figure 4.14: Cluster patterns in estimated distance and in anchoring distance emerged from our analysis.

Figure 4.15: Performance benefit by Cluster. On the X axis we have the clusters emerged from analyses for RQ3, expressing patterns in the overlap between Human estimate and AI estimate. On the Y axis we have the mean difference between performance of Humans Alone and Human-AI in Team from RQ1. We see that the difference is higher for Cluster 2, which is also the Cluster whose performance benefits the most from AI advice.

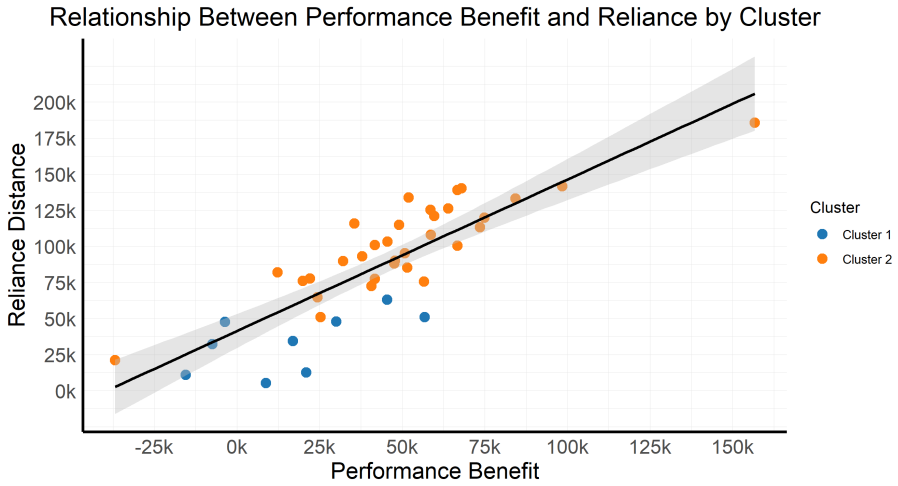

Figure 4.16: Relationship between performance benefit and reliance, depending on clusters from the analysis of RQ3. The separation between clusters indicates that subjects in Cluster 2 consistently followed more AI advice as compared to Cluster 1, gaining a higher performance benefit from it and compensating the higher estimate error when *Alone* by following AI advice in *Team*.

4.5.7 Condition Order Effects

The model including the effect of *Block order* showed a significant main effect of *Condition* $[F(1, 2382) = 102.369, p < .001]$, a significant main effect of *Block Order* $[F(1, 189) = 5.546, p < .05]$ and a significant *Condition by Block Order* effect $[F(1, 2382) = 5.863, p < .05]$. Inspection of post-hoc tests revealed that the contrast between *Team: Concurrent First* and *Team: Sequential First* was not significant $[t(73.5) = -0.422, p = 0.976]$, implying that the effect of AI assistance on performance was not dependent on the order in which subjects received AI advice. Results for this model can be inspected graphically in Figure 4.17.

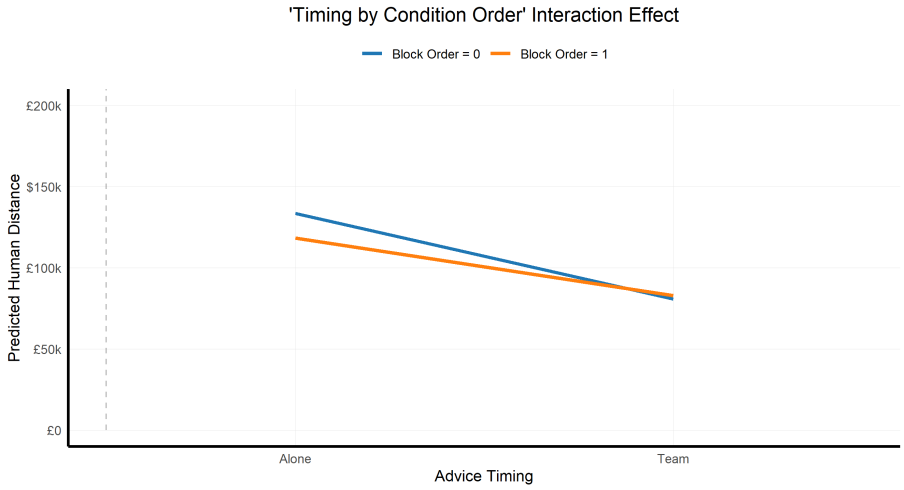

Figure 4.17: Block order effects for AI assistance. Performance in *Team* does not depend on receiving advice in *Sequential Design first* (Block order 0) or *Concurrent Design first* (Block order 1)

4.6 Discussion

4.6.1 Human augmentation without Human-AI synergy

Results obtained for RQ1 show significant Human augmentation and absence of Human-AI synergy. Humans alone performed worse than AI alone, and integration of AI advice improved their estimates (leading to augmentation) but not to the point of surpassing AI alone, which was the best estimator (leading to absence of synergy). The fact that algorithmic forecasting and AI estimates are superior to estimates by humans alone, and with lower variability, is a well-established finding in Human-AI Collaboration literature in many domains (Ægisdóttir et al., 2006; Dawes et al., 1989; Grove et al., 2000; Meehl, 1954).
Our results closely align with the most recent literature in Human-AI Collaboration. Vaccaro et al., 2024 conducted an extensive metanalysis analyzing 370 effect sizes reported in 170 studies, finding evidence that Human-AI teams perform generally worse than the best of the two team members taken alone, especially in decision making tasks (as opposed to tasks on generating contents, where instead benefits were noticeable), and especially when the best performer of the two was AI alone. Our results are compatible also with another finding of their metanalysis: augmentation is most significant in "decide" tasks with numeric input and requiring continuous output, but synergy is generally not achieved in such tasks, where AI alone performs generally better. Supporting evidence comes also from the study by Poursabzi-Sangdeh et al., 2021, who used a house price prediction task, with numeric input data and numeric output required, computing prediction error; they found similar results, with Human-AI teams performing better than Humans alone, but worse than AI alone. Our study adopted a multimedial input format for data (numeric, text and image) and required continuous output, and results support the hypothesis that regression-like problems are difficult for humans to solve and that AI is generally the best estimator for such problems.

4.6.2 AI assistance improves performance independently of the decision design

Results for RQ2 show that no significant differences in performance between concurrent and sequential design can be found in our data if we aggregate by subject: overall, both the concurrent and the sequential design lead to better Human-AI performance as compared to Humans alone, but the difference between the two was not significant. Consequently, our hypothesis for RQ2 found only partial support: both designs lead to Human augmentation without Human-AI synergy, but without significant differences between the two. Performance in the two designs shows that estimates with AI have lower SD as compared to estimates of Humans alone, making AI-assisted performance both more accurate and more precise. Similarly, Poursabzi-Sangdeh et al., 2021 also reported lower SDs for Human-AI estimates as compared to estimates by Humans alone. Our results align with what Tejeda et al., 2022 found with a similar methodology but adopting a noisy-image classification task. They compared an AI-assisted sequential paradigm to an AI-assisted concurrent paradigm and to unassisted humans, finding that human performance with a more accurate AI improved as compared to unassisted humans, but without differences between concurrent and sequential

decision paradigms. Overall, our results for RQ3 align with current literature with mixed findings relative to how the timing of AI advice impacts performance in Human-AI teams (Buçinca et al., 2021; Park et al., 2019; Rastogi et al., 2022; Steyvers & Kumar, 2024). We refer to the discussion on cluster analysis for further considerations.

4.6.3 Reliance of AI opinion

Results for RQ3 indicate similar distances between human estimate and AI estimate both for the concurrent design and the sequential design, mirroring the comparable differences in performance between the two. In both the *concurrent* and *sequential* conditions, the distance between the estimates of Humans alone and AI alone (without any interaction) were much larger, indicating that humans integrate AI advice to a great extent, with substantial convergence towards AI advice. There is significant overlap with AI opinion, but without significant differences between *concurrent* and *sequential* conditions. Whereas a direct comparison of the means evidenced stronger convergence towards AI opinion in the concurrent design, the t-tests evidenced that the difference between the distributions was not statistically significant. Consequently, our hypothesis for RQ3 did not find support in our data: AI advice in the concurrent design was not weighted significantly more than in the sequential design, and AI advice was not discounted more in the sequential design. Taken together, our results for performance and reliance indicate that egocentric discounting is not influenced by the design architecture, and particularly it is neither exacerbated by a sequential design, as we supposed, nor it is mitigated by it, as other researchers hypothesised (B. Green & Chen, 2019). Egocentric advice discounting and reliance on AI seem unaffected by the timing of AI advice in respect to the decision paradigm.

4.6.4 Clusters in performance and reliance

Our analyses show that there is consistent inter-individual variability in the effects. Inspection of patterns within this variability evidenced clusters in performance and reliance, with some subjects behaving according to our hypotheses and some others behaving differently compared to our expectations. It is particularly insightful to compare the cluster structures for RQ1 and RQ3 to understand the interplay between performance and reliance, and how it leads to performance benefits dependent on the performance deficit of participants. The two cluster classifications are significantly correlated with each other, and the positive ARI indicates a partial overlap of the two structures. In practical terms, this means that clustering in reliance contains information about clusters in performance; they are theoretically distinguished but empirically correlated. When we consider the distance between Human and AI estimates (reliance, RQ3), we see that participants in Cluster 1 rely less on AI advice. When we consider the distance between estimates and Ground Truth (performance, RQ1), we see that participants in Cluster 1 have comparable performance either *Alone* of with AI assistance in *Team*, but instead participants in Cluster 2 experience a performance deficit: their estimates when *Alone* are much less accurate as compared to their estimates with AI assistance in *Team*. However, we also see that Cluster 2 (RQ3) converges more towards AI advice, relying more on it and obtaining a performance benefit: Cluster 2 (RQ1) has a comparable worse performance when *Alone* but performs equally well to Cluster 1 (RQ1) when assisted by AI in *Team*. The

significant overlap between the two cluster structures hints at the possibility that participants in Cluster 2 are comparably worse than participants in Cluster 1 in estimating the Ground Truth price, but they rely more on AI advice, which is more accurate, compensating the performance deficit and gaining a higher performance benefit: both clusters perform equally well in the Team AI-assisted condition. AI advice acts as a decision support: unnecessary to some, necessary to others, producing a consistent performance benefit for participants who *Alone* perform worse. This interpretation is confirmed by directly analyzing the performance gap between estimates (in data from RQ1) when *Alone* and in *Team* with AI, but using the cluster structure from reliance in RQ3. The mean difference between estimates in the two conditions (*Alone* - *Team*) is much higher for Cluster 2, indicating an initial performance deficit when *Alone* that gets compensated by relying more on AI advice in *Team* (since the *Team* performance of the two Clusters do not statistically differ). Regression analysis considering the cluster structure confirms this finding: the separation between clusters indicates that subjects in Cluster 2 consistently followed more AI advice as compared to Cluster 1, gaining a higher performance benefit from it and compensating the higher estimate error when *Alone* by following AI advice in *Team*.

A second interpretation of the data, not excluding the first, is that participants in Cluster 1 are less influenced by AI advice and their evaluations are less malleable, and consequently their performance benefit is much narrower as compared to the participants in Cluster 2. Indeed, performance of Humans in *Team* is on average worse than performance of AI alone, no matter the cluster, so also participants in Cluster 1 could theoretically benefit from AI advice, but they integrate it to a much narrower extent.

We finally report four interesting phenomena. First, examining the plots from the mixed models for RQ1 evidenced that, for some subjects, performance distances *increase* when interacting in *Team* with AI as compared to when they are *Alone*: their performance *worsens* with AI interaction. These subjects perform comparably to AI alone, one of them being even better on average, and accepting and following AI advice tends to mislead their estimates. This pattern is shown also in the clustering structure with k=5, albeit it is indicated as suboptimal by the quantitative criteria. Second, inspection of the regression plot for RQ2 evidences that our hypothesis (on the concurrent design leading to better performance) might be true for some subjects while being contradicted for others. Here, the cluster structure for k=5 is able to capture this pattern well. Third, inspection of the regression plots for RQ3 evidences the same phenomenon but for anchoring distances: our hypothesis (on influence as "*who-comes-first*" phenomenon) might be true for some subjects while being exactly the opposite for others. Taken together, these three phenomena indicate that the analyses most commonly used (correlations, linear regression models) are too simple to provide a detailed description of the processes involved in advice integration: widely diverging patterns might manifest between subjects but they might be "averaged out". Instead, the higher resolution given by cluster analyses might be needed to gain a deeper understanding of the dynamics of advice integration. Finally, inspection of the plot for clusters of RQ3 indicates that no subject had estimates in *Team* diverging from AI opinion: mostly converging to it, sometimes with negligible variations as compared to their estimates when *Alone*. The difference "Alone-Team" in anchoring distance for each subject was always positive, indicating always a convergence towards AI advice. We interpret this effect as *absence of reactance* to AI advice. We observed only shades of reliance. None of our subjects seemed really "averse" to AI advice, even if it

was sometimes misleading.

4.6.5 Familiarity with AI

To gain insights into possible sources of differences in performance and reliance, emerged clearly from cluster analysis, we used logistic regression to study the effect of our individual variables on estimated distance (performance) and anchoring distance (reliance). *Familiarity with AI* was the only significant predictor of cluster classification for reliance behavior, but not for performance: the more participants are familiar with AI-LMMs, the less they seem to rely on it. It is 80% more probable to belong to the less-reliant Cluster 1 for each unit increase in *Familiarity with AI*. None of the other predictors (education level, gender, age, years lived in London) led to significant effects. The AUC of 0.79 indicates that *Familiarity with AI* can be used to establish a decision boundary for cluster classification. Given that participants in Cluster 1 had the lowest performance gap and they relied less on AI advice, they also obtained the least performance benefit, even if performance could be still improved for most of them because AI advice was more precise than their estimates. This finding might apparently contrast with literature on trust indicating that higher trust is associated with higher reliance (Klingbeil et al., 2024), but we have no means to actually link, in our data, trust to reliance. It is possible that, in our sample, participants more familiar with LLMs might rely less on them; a likely explanation would be that they are also more exposed to their errors (Dietvorst et al., 2015), but our data do not allow us to investigate this hypothesis. A deeper look into inter-individual variability is needed to profile users who might rely less on AI advice.

Chapter 5
Investigations into AI-AI Collaboration

5.1 Introduction

Since their release to a wider public, AI based on Large Language Models advanced at a fast pace, progressing from text generators to image/video analyzers and generators, and to multitask learners in many domains (Brown et al., 2020; Radford et al., 2019; Wijk et al., 2025). The most immediately observable improvements initially came from extensions of their abilities, such as integration of computer vision, leading to Large Multimodal Models (LMM: Yang et al., 2023), and then from the ongoing attempt to develop "thinking" models with reasoning abilities, capable of tackling a larger problem space (Wei et al., 2023; Yao et al., 2023). This progress can be observed in scores obtained by various LLMs in well-established benchmarks, quickly saturated within a few years, if not months, from their development: GLUE (A. Wang et al., 2018), SuperGLUE (A. Wang et al., 2020) and MMLU (Hendrycks et al., 2021), with HLE currently under testing (Phan et al., 2025).

A third noticeable improvement regards a shift of focus from linguistic behavior to agency in the external world. The development of Large Behavior Models (Team et al., 2025) can be considered a first attempt in this direction. Similarly, according to the roadmap by OpenAI, conversational AI is only the first step in a hierarchy of progressively more complex AIs, with reasoning AI as second step and *agentic* AI as third step, currently under development (the last two steps being only announced as ultimate goals towards Artificial General Intelligence, but currently not existing). Moreover, improvements in the autonomy of AI in completing long tasks are remarkable (Kwa et al., 2025). In fact, improvements in benchmarks such as GLUE or MMLU, and even more remarkably the encoding of medical knowledge (Ho et al., 2024; Liu et al., 2024; Rossettini et al., 2024; Scaioli et al., 2023; Singhal et al., 2023), describe mostly successes in question answering but might be difficult to translate in practical terms, in real-world scenarios (Hager et al., 2024), where several failures of agentic AI emerged (Bryan et al., 2025). New developments consequently look at autonomy as an important next step for AI development. Measurement of task-completion time horizon shows that AI models are becoming more and more autonomous at an impressive rate, as we saw in the introduction.

Parallely, we saw that Human-AI synergy is generally difficult to achieve. AI support is still valuable to reach Human augmentation, since humans will still be the ultimate decision makers in most scenarios. However, AI alone outperforms humans and AI-assisted humans in many domains (Vaccaro et al., 2024), and the advancements in AI autonomy raise the interesting possibility to build not just Human-AI teams, but proper *AI-AI teams*: structured interactions between autonomous artificial agents. This seems most likely to happen initially in software development, an ability where LLMs have become commonplace (Kwa et al., 2025; Wijk et al., 2025). It is important to study *AI-AI Collaboration* and compare it to Human-AI

Collaboration, because of its potential in the near future. Cooperation between artificial agents will be increasingly relevant. This investigation can be framed in the more general debate about similarities and differences between human minds and artificial minds. Not much is known about advice integration, metacognitive abilities and "inter-personal" biases such egocentric discounting in artificial minds as compared to human ones. With this study we wanted to compare human minds and artificial intelligence directly on the same task, the one presented in our Human-AI study of the previous chapter, but performed by two AIs: one as decision maker, the other as adviser. We consider this study as an initial attempt to understand the potential of AI-AI Collaboration, shedding light on how a completely artificial interactive team might work as compared to Human-AI teams.

5.2 Hypotheses

For this study we considered again the same research questions for our Human-AI collaboration study presented in the previous chapter, but adapted for an AI-AI team. We also extend the terminology used by (Vaccaro et al., 2024), now referred to the AI-AI team, as *AI-by-AI augmentation* and *AI-AI synergy*. The first "AI" will always refer to the decision maker (here an artificial one), and the second "AI" always refers to the adviser. We mean AI-by-AI augmentation in analogy to human augmentation: the decision maker improves in team with the adviser, but not as much as the best of the two agents. Similarly, AI-AI synergy means a cooperative interaction where the team performs better than the best of the two agents when alone. We designed two parallel versions with two different AI advisers. In one case, the adviser was as accurate as the decision maker (*AI-AI standard*); in the second case, the adviser was more accurate than the decision maker, with a performance described by a smaller error in estimates (*AI-AI reduced*). We also added a fourth and fifth research question comparing directly Human-AI and AI-AI teams on the benefits of receiving advice.

Research Question 1: Which agent or combination of agents performs best?
Hypothesis 1: For the *AI-AI reduced* variation, we expect to observe AI-by-AI augmentation but not AI-AI synergy: assistance will benefit performance as compared to the decisor AI alone, but the AI-AI team will not perform better than the best of either of the two members. For the *AI-AI standard* variation, we expect comparable performance of AI alone and AI-AI team, so neither augmentation nor synergy.
The *AI-AI reduced* team will perform better as compared to the *AI-AI standard* team. In any case, the AI-AI team will outperform the Human-AI team.

Research Question 2: Does the performance of the AI-AI team depend on the timing of AI assistance?
Hypothesis 2: For the *AI-AI reduced* variation, we expect to observe AI-by-AI augmentation but not AI-AI synergy. We do not expect that the timing of advice in respect to the decision (concurrent or sequential) will have an effect on the performance, but both will be better than decisor AI alone.
For the *AI-AI standard* variation, we expect comparable performance of *AI alone*, *AI-AI concurrent* and *AI-AI sequential* conditions, so neither augmentation nor synergy. In any

case, the AI-AI team will outperform the Human-AI team.

Research Question 3: Does anchoring depend on the timing of AI assistance?
Hypothesis 3: For both the *AI-AI reduced* and the *AI-AI standard* variation, we do not expect significantly different degrees of the decisor AI reliance on the adviser AI: AI advice in the concurrent design might not be weighted more as compared to the same AI advice but in a sequential design.
For the *AI-AI reduced* variation, we expect a comparable distance of the adviser AI from the decisor AI in both conditions: overlap between decisor AI and adviser AI will be higher in either the concurrent or sequential condition as compared to the distance between the two when the decisor AI is alone.
For the *AI-AI standard* variation, distance between decisor AI and adviser AI will be instead comparable across all three conditions (*AI alone, AI-AI concurrent, AI-AI sequential*).

Research Question 4: Is anchoring stronger in Human-AI teams as compared to AI-AI teams?
Hypothesis 4: Anchoring expresses the convergence of the decisor's estimate towards the adviser's estimate when they interact, as compared to their a-priori distance when they are alone. We predict that this convergence will be more pronounced for Human-AI teams as compared to AI-AI teams.

Research Question 5: Is AI advice benefit higher for AI-AI teams as compared to Human-AI teams?
Hypothesis 5: Advice benefit increases with the relative difference in performance between decisor and adviser. Consequently, *AI-AI standard* teams will have an advice benefit near zero, because the decisor interacts with an advisor of comparable performance; *AI-AI reduced* teams and Human-AI teams will experience a similar advice benefit, because the relative performance gap between decisor and adviser is the same (by experimental design).

5.3 Methods

5.3.1 Design, Task and Variables

This experiment has been built as a replica of the Human-AI experiment presented previously, but done by two AI agents. The design of the experimental conditions is the same, so we have the *AI Alone* condition, the *AI-AI concurrent* condition and the *AI-AI sequential* condition. We used the same task, with the same experimental stimuli, variables and number of trials. The main independent variable is still the *timing of AI assistance* given by our three experimental conditions. The two main dependent variables are still the distance between the decision maker's estimate and Ground Truth price, quantifying performance, and the distance between the decision maker's estimate and the adviser's estimate, quantifying reliance as the degree of overlap between the two (the extent of their agreement).

5.3.2 The AI-LMM used as decision maker

The subject in the experiment is the same AI that we used to generate the advice in the Human-AI version, but here we use it as a decision maker. Gemini 2.0 Flash was accessed through an API coded in Python 3.13 with Visual Studio Code. The API calls the AI model and asks it to perform the task, by looping through the stimuli and giving its evaluations. We took the estimates generated for the Human-AI experiment as the answers for the present *AI alone* condition. The API was used to obtain answers for the *AI-AI concurrent* and the *AI-AI sequential* conditions. The estimates were obtained with one run of the API looping over all 60 sets, one loop for each AI assistance condition, giving us 60 estimates for each condition. Due to time and resource constraints, it was not possible to have the same fine-grained analysis and cluster analysis we did for the Human-AI study, lacking the inter-subject variability of the human sample, but a version with an extended "sample size of AIs" is currently under development to study variability of AI estimates and possible clusters in it, as we did with human participants.

We prompted the model with the following string for the *AI alone* condition:
I need your assistance for a real estate evaluation task. I upload a document containing information about a real estate in London, United Kingdom, with four pictures of the interiors: living room, kitchen, bathroom, bedroom. A description of the property features is provided in the document, together with the pictures.
I ask you to estimate the asking price for this property in British pounds, in a range from £0 up to £1 million pounds. Please consider all the information available and give your estimation of the most likely asking price.
Give a point estimate and not a range. Be concise in your output, keeping it below 150 words. Organize your answer by stating your evaluation first (highlight the price in bold and state "My evaluation for this property is..."), then your arguments for the evaluation in a bullet point list containing your considerations relative to: location, property features, market comparables.

We prompted the model with the following string for the *AI-AI concurrent* condition
I need your assistance for a real estate evaluation task. I upload a document containing information about a real estate in London, United Kingdom, with four pictures of the interiors: living room, kitchen, bathroom, bedroom. A description of the property features is provided in the document, together with the pictures.
I ask you to estimate the asking price for this property in British pounds, in a range from £0 up to £1 million pounds. Together with the information about the real estate, also an evaluation of the price by a Large Language Model is provided. Please consider all the information available and give your estimate of the most likely selling price.
Give a point estimate and not a range. Be concise in your output, keeping it below 150 words. Organize your answer by stating your evaluation first (highlight the price in bold and state "My evaluation for this property is..."), then your arguments for the evaluation in a bullet point list containing your considerations relative to: location, property features, market comparables, information from the AI LLM provided.

We prompted the model with the following string for the *AI-AI sequential* condition
I need your assistance for a real estate evaluation task, in two steps, each one requiring an evaluation by you. As first step, I upload a document containing information about a real estate in London, United Kingdom, with four pictures of the interiors: living room, kitchen, bathroom, bedroom. A description of the property features is provided in the document, together with the pictures.
I ask you to estimate the asking price for this property in British pounds, in a range from £0 up to £1 million pounds. This will be your first estimate for this set.
As a second step, after having provided your initial judgment, you will be given an evaluation of the price by another Large Language Model, and you are asked to give a second evaluation of the most likely selling price, still in a range from £0 up to £1 million pounds.
In both cases, give a point estimation and not a range. Be concise in your output, keeping it below 150 words. Organize your answer by stating your evaluation first (highlight the price in bold and state "My evaluation for this property is..."), then your arguments for the evaluation in a bullet point list containing your considerations relative to: location, property features, market comparables, information from the AI LLM provided.

5.3.3 The AI-LMM used as adviser

The estimates used as advice were generated with ChatGPT4.1. We took the same estimates that we generated in the preliminary assessment for the Human-AI study, in order to have two different AIs for our new study. Given that the two models produced comparable estimates in our preliminary assessment, we decided to perform the AI-AI experiment in two variations. In the first variation, we used exactly the same advice generated by ChatGPT 4.1. We will refer to this variation simply as *AI-AI standard* or simply *AI-AI*. In the second variation, we wanted to mimic the Human-AI situation as closely as possible by having an adviser with superior performance as compared to the decision maker, but performance of the two AIs are very similar. We decided then to artificially manipulate the estimates by the adviser (ChatGPT), reducing them by a factor comparable to the ratio between the precision of Humans alone as compared to the precision of AI alone in the Human-AI experiment (about 1.9). This generated a second set of advice whose precision was higher, and we will refer to this variation as *AI-AI reduced*.

5.4 Analyses

5.4.1 Distances

Data analysis for the AI-AI study mirrors the one for the Human-AI study. *Distance* is again conceptualized as a continuous function on the set of positive real numbers. Consequently, all distances are absolute values and quantifies the precision of the estimates around the ground truth value (d_P in the case of performance) or the overlap of estimates between the decisor AI and the adviser AI (d_A in the case of anchoring).

5.4.2 Research Question 1

We grouped together the trials of the *AI-AI concurrent* block and of the *AI-AI sequential* block to obtain the *AI-AI team* condition, to be compared against the trials without AI assistance and to the estimates of AI alone.
As a first step, distances between estimates and ground truth in the various experimental conditions can be compared. We will compare our decisor AI alone to its performance in the AI-AI team, and we will compare performances of the *AI-AI standard* variation with the *AI-AI reduced* variation.
Finally, we estimated a linear regression model with distances as dependent variable and AI assistance condition C (absent vs. AI-AI team) and variation (AI-AI standard vs. reduced) as independent variables:

$$d_P \sim C + V$$

5.4.3 Research Question 2

We computed the distances between decisor AI estimate and ground truth price in all conditions (*AI Alone, AI-AI concurrent, AI-AI sequential*) and also considered the estimates by the adviser AI alone. We also computed the standard deviations of the estimates for each experimental condition to compare their ranges.
We estimated a linear regression model with the distances as dependent variable and with AI advice timing T (absent, concurrent, sequential) and variation (AI-AI standard vs. reduced) as independent variables:

$$d_P \sim T + V$$

5.4.4 Research Question 3

We quantified anchoring as the degree of overlap between human opinion and AI advice with the distance between their estimates, defined as a positive real-valued function.
For the AI-AI standard variation, we quantify our expectation as:

$$d_A(AI : alone) = d_A(AI - AI : conc) = d_A(AI - AI : seq)$$

For the AI-AI reduced variation, we quantify our expectation as:

$$d_A(AI - AI : conc) = d_A(AI - AI : seq) < d_A(AI : alone)$$

Analogously to RQ2, we estimated a linear regression model with the anchoring distances as dependent variable and with AI advice timing T (absent, concurrent, sequential) and variation (AI-AI standard vs. AI-AI reduced vs Human-AI) as independent variables:

$$d_A \sim T + V$$

5.4.5 Research Question 4

We expect that anchoring distance will be overall lower for the AI-AI standard variation, because by design the decisor and the adviser have almost comparable performance. We quantify this expectation as:

$$[d_A(AI - AI)]_{STD} < [d_A(AI - AI)]_{RED}$$

We also need to express how much the decisor estimate in the team changes as compared to its estimates when alone. We define anchoring ratio the quantity:

$$A_R = 1 - \frac{d_A(AI - AI : team)}{d_A(AI : alone)}$$

This value quantifies the convergence of the decisor towards the adviser. It is built by considering the ratio between the estimate in a team (the "AI-AI team", either concurrent or sequential design) and the *a-priori distance* between the two agents' estimates prior to their interaction (the condition "AI alone"). When their estimates converge, the distance at the numerator tends to 0, so the anchoring ratio tends to 1, expressing increasing agreement. When their estimates diverge, the distance at the numerator is close to the distance at the denominator (indicating that the opinion of the decisor in the team is not substantially different from its opinion when alone), so the ratio tends to 0, expressing low agreement. Indeed, extremely divergent opinions might lead to negative anchoring ratios (indicating a sort of "opposition" or "reactance" to the adviser in the team). An analogous definition can be introduced for Human-AI teams. By using this definition, we quantify our expectation about anchoring ratios in RQ4 as:

$$[1 - \frac{d_A(H - AI : team)}{(d_A(H : alone)}] > [1 - \frac{d_A(AI - AI : team)}{d_A(AI : alone)}]_{STD}$$

$$[1 - \frac{d_A(H - AI : team)}{d_A(H : alone)}] > [1 - \frac{d_A(AI - AI : team)}{d_A(AI : alone)}]_{RED}$$

Qualitatively, this expectation describes the more general hypothesis that AI might be more robust to the influence of external opinions and less susceptible to revise its opinion as compared to humans, which might be more malleable and rely more on advice coming from others. In this sense, AI might be less prone to "automation bias" and more prone to "egocentric bias" in evaluations as compared to humans.

5.4.6 Research Question 5

We define *advice benefit* as:

$$A_B = 1 - \frac{d_P(AI - AI : team)}{d_P(AI : alone)}$$

This value expresses how much the decisor gains in performance by interacting in a team as compared to the condition in which it acts alone. It is composed by the performance ratio of the distances in the two conditions: as team performance improves compared to performance alone, the distance at the numerator tends to zero (because the estimates converge to the ground truth) and advice benefit tends to 1; instead, if team performance is very similar to performance of the decisor alone, the ratio between the two tends to 1 and consequently advice benefit tends to zero. With this definition we quantify our expectations for RQ5 as:

$$[1 - \frac{d_P(AI - AI : team)}{d_P(AI : alone)}]_{STD} \sim 0$$

$$[1 - \frac{d_P(H - AI : team)}{d_P(H : alone)}] \sim [1 - \frac{d_P(AI - AI : team)}{d_P(AI : alone)}]_{RED}$$

5.5 Results

5.5.1 Research Question 1

Distances Our results relative to mean absolute distances support the hypothesis we formulated for RQ1. For the *AI-AI standard* variation, performance by AI alone (mean distance from GT: £66250) and by the AI-AI team (mean distance: £66279) are comparable, lacking both augmentation and synergy. For the *AI-AI reduced* variation, we have evidence of AI-by-AI augmentation: performance of the decisor AI assisted by AI (mean distance from GT: £44967) is better as compared to performance of AI alone (mean distance from ground truth: £66250). However AI-AI synergy was not reached: mean distance by the AI advice by design was much lower (£33125). Referring to the data from human participants of our previous experiment, we see that in both cases the AI-AI team outperforms the Human-AI team (mean distance from GT: £81970). Results are plotted in Fig. 5.1.

Linear Model We estimated a linear regression model with distances as dependent variable and AI assistance condition C (Alone vs. Team) and variation (AI-AI standard vs. AI-AI reduced vs. Human-AI) as independent variables. The model is significant: $F(3, 2779) = 680.3, p < .001$. All post-hoc tests with Tukey correction for multiple comparisons yielded significant results at $p < .001$, except for the comparison *AI Team standard* vs. *AI Team reduced* $[t(2779) = 0.243, p = 0.999]$ and the comparison *AI Alone standard* vs. *AI Alone reduced* (these were the same answers indeed by design).

5.5.2 Research Question 2

Distances Our hypothesis for RQ2 finds support in our data (see Fig. 5.2). For the *AI-AI reduced* variation we observe AI-by-AI augmentation in the *concurrent* condition (mean distance from GT: £44142) and in the *sequential* condition (mean distance from GT: £45792), but in both cases the performance of the adviser AI is not reached (mean distance from GT: £33125), so no synergy emerges. Performance in the concurrent vs sequential condition is comparable as expected. For the *AI-AI standard* variation we do not observe AI-by-AI

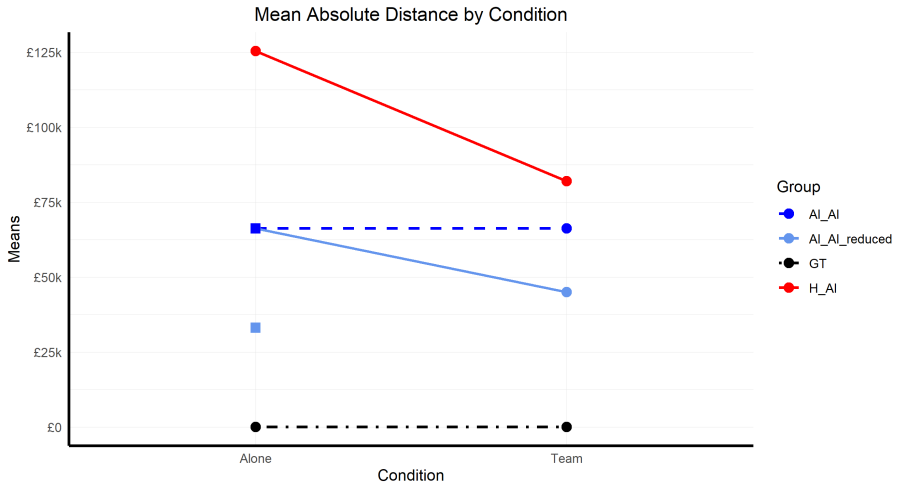

Figure 5.1: Results for our RQ1: distances of estimates from ground truth (GT) for our datasets: Human-AI (red), AI-AI standard (dark blue), AI-AI reduced (light blue), each either Alone or in Team. The squares indicate the reference point to evaluate synergy: Human-AI synergy must be evaluated referring to AI alone (dark blue square), whereas AI-AI synergy must be evaluated referring to reduced-AI alone (light blue square). For all datasets and conditions the reference for performance is the Ground Truth line (black).

augmentation, with comparable distances in the *concurrent* (mean distance from GT: £65250) and *sequential* (mean distance from GT: £67308) conditions, with both comparable to the adviser AI (mean distance from GT: £69917). Both concurrent and sequential conditions in the AI-AI teams outperform the Human-AI team, whose distances are far higher (*concurrent* condition, mean distance from GT: £84802; *sequential* condition, mean distance from GT: £79004).

Linear Model We estimated a linear regression model with distances as dependent variable and AI assistance condition C (Alone, Concurrent, Sequential) and variation (AI-AI standard vs. AI-AI reduced vs. Human-AI) as independent variables. The model is significant: $F(4, 2778) = 513.5, p < .001$. All post-hoc tests with Tukey correction for multiple comparisons yielded significant results at $p < .001$, except for the six comparisons between the AI assisted conditions, either standard or reduced: *AI-AI standard concurrent* vs. *AI-AI standard sequential* $[t(2778) = 2.839, p = .105]$; *AI-AI standard concurrent* vs. *AI-AI reduced concurrent* $[t(2778) = 0.244, p = 1.000]$; *AI-AI standard concurrent* vs. *AI-AI reduced sequential* $[t(2778) = 1.371, p = 0.909]$; *AI-AI standard sequential* vs *AI-AI reduced concurrent* $[t(2778) = -0.925, p = 0.992]$; *AI-AI standard sequential* vs *AI-AI reduced sequential* $[t(2778) = 0.244, p = 1.000]$; *AI-AI reduced concurrent* vs *AI-AI reduced sequential* $[t(2778) = 2.839, p = 0.105]$. The comparison H-AI concurrent vs H-AI sequential was not significant, as we already obtained in the human dataset $[t(2778) = 2.839, p = 0.105]$; the *AI Alone standard* vs. *AI Alone reduced* was not significant (these were the same answers indeed by design).

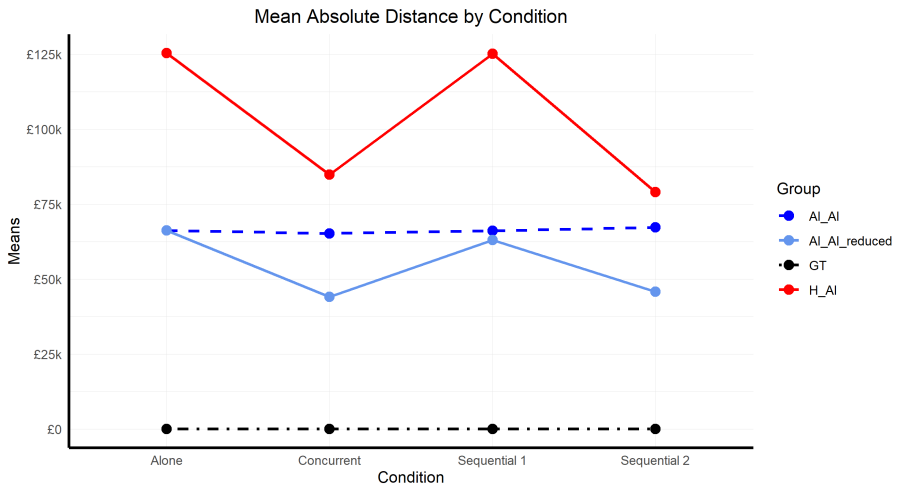

Figure 5.2: Results for our RQ2: distances of estimates from ground truth (GT) for our datasets: Human-AI (red), AI-AI standard (blue), AI-AI reduced (light blue); all conditions are plotted (Alone, Concurrent, Sequential). For all datasets and conditions the reference for performance is the Ground Truth line (black).

5.5.3 Research Question 3

Distances Distances between decisor AI and adviser AI resulting from our data partially support the hypothesis we formulated for RQ3 (see Fig. 5.3). For the *AI-AI standard* variation, the anchoring distance in the *concurrent* condition (£5333) and in the *sequential* condition (£8025) are comparable, slightly smaller than the distance between the two agents when considered alone (£3367). For the *AI-AI reduced* variation, the anchoring distance in the *concurrent* condition (£22275) and in the *sequential* condition (£21458) are comparable and both lower as compared to the distance between the two agents considered alone (£33125). In both variations, the degree of overlap between decisor AI and adviser AI seems comparable as expected.

Linear Model We estimated a linear regression model with distances as dependent variable and AI assistance condition C (Alone, Concurrent, Sequential) and variation (AI-AI standard vs. AI-AI reduced vs. Human-AI) as independent variables. The model is significant: $F(4, 2778) = 175.6, p < .001$. The post-hoc tests with Tukey correction for multiple comparisons yielded significant results at $p < .001$ are listed in Table 5.1.

5.5.4 Research Question 4

Anchoring distances are lower for the *AI-AI standard* variation as expected, as we have already seen in the results for RQ3. The anchoring ratios are reported in Table 5.2: they are higher for Human-AI teams in both AI-assisted conditions, as compared to the anchoring ratios of the AI-AI teams. Visual inspection of the anchoring distances (*Alone* vs. *Team*, grouping concurrent or sequential, see Fig 5.4) indicates qualitatively what the anchoring ratios show quantitatively: the convergence of the decisor estimate in *Team* towards the adviser's estimate,

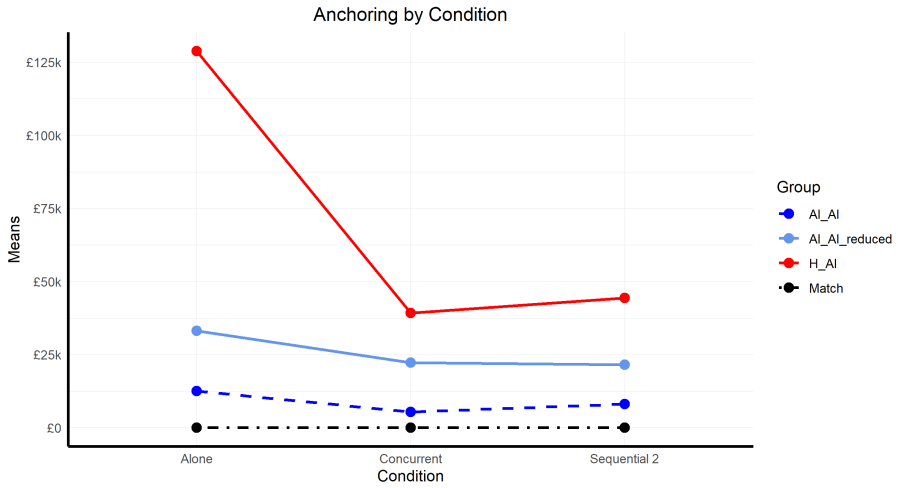

Figure 5.3: Distances between decisor AI and adviser AI for the three conditions (Alone, Concurrent, Sequential) and datasets (Human-AI, AI-AI standard, AI-AI reduced). The line where the estimates are matched leading to null distance is plotted in black as a reference.

as compared to their distances when doing the task *Alone*, is more pronounced for Human-AI teams as compared to AI-AI teams, as expected.

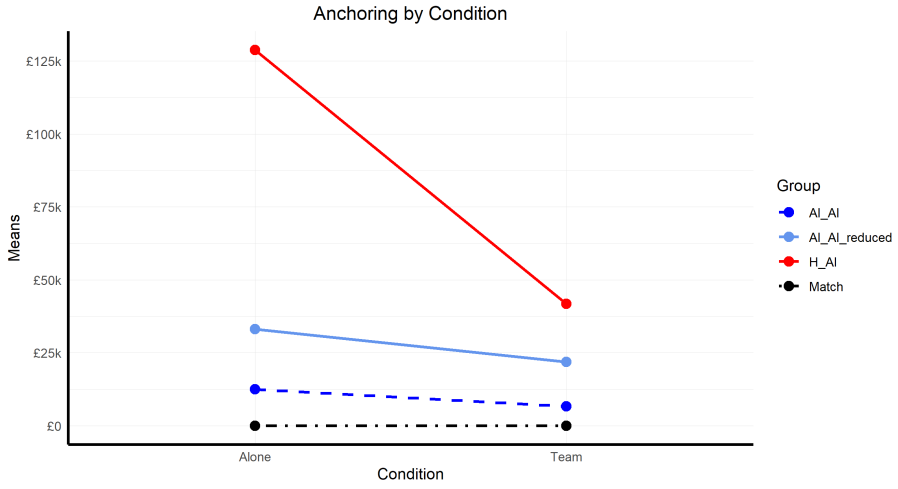

Figure 5.4: Anchoring distances for all our datasets (AI-AI standard, AI-AI reduced, Human-AI) for the Alone condition and Team (grouping concurrent and sequential AI assistance). The line where the estimates are matched leading to null distance is plotted in black as a reference.

Table 5.1: Significant post-hoc tests from the linear model of RQ3.

Contrast		t value	p value
AI Alone std	AI-AI std concurrent	$t(2778) = 20.841$	$p < .0001$
AI Alone std	AI-AI std sequential	$t(2778) = 19.383$	$p < .0001$
AI Alone std	AI-AI red concurrent	$t(2778) = 7.270$	$p < .0001$
AI Alone std	AI-AI red sequential	$t(2778) = 6.753$	$p < .0001$
AI Alone std	Human Alone	$t(2778) = -9.935$	$p < .0001$
AI-AI std concurrent	AI Alone red	$t(2778) = -9.440$	$p < .0001$
AI-AI std concurrent	Human Alone	$t(2778) = -19.214$	$p < .0001$
AI-AI std concurrent	Human-AI concurrent	$t(2778) = -9.935$	$p < .0001$
AI-AI std concurrent	Human-AI sequential	$t(2778) = -9.133$	$p < .0001$
AI-AI std concurrent	AI alone red	$t(2778) = -8.920$	$p < .0001$
AI-AI std concurrent	Human Alone	$t(2778) = -18.556$	$p < .0001$
AI-AI std concurrent	Human-AI concurrent	$t(2778) = -7.879$	$p < .0001$
AI-AI std concurrent	Human-AI sequential	$t(2778) = -9.935$	$p < .0001$
AI Alone red	AI-AI red concurrent	$t(2778) = 20.841$	$p < .0001$
AI Alone red	AI-AI red sequential	$t(2778) = 19.383$	$p < .0001$
AI Alone red	Human Alone	$t(2778) = -8.318$	$p < .0001$
AI Alone red	Human-AI concurrent	$t(2778) = 3.542$	$p < .05$
AI-AI red concurrent	Human Alone	$t(2778) = -17.826$	$p < .0001$
AI-AI red concurrent	Human-AI concurrent	$t(2778) = -8.318$	$p < .0001$
AI-AI red concurrent	Human-AI sequential	$t(2778) = -7.751$	$p < .0001$
AI-AI red sequential	Human Alone	$t(2778) = -17.169$	$p < .0001$
AI-AI red sequential	Human-AI concurrent	$t(2778) = -6.491$	$p < .0001$
AI-AI red sequential	Human-AI sequential	$t(2778) = -8.318$	$p < .0001$
Human Alone	Human-AI concurrent	$t(2778) = 20.841$	$p < .0001$
Human Alone	Human-AI sequential	$t(2778) = 19.383$	$p < .0001$

Table 5.2: Anchoring ratios for all conditions and variations in the Human-AI and AI-AI studies.

Dataset (Variation)	Condition	Anchoring Ratio
AI-AI standard	Concurrent	0.573
AI-AI standard	Sequential	0.358
AI-AI reduced	Concurrent	0.328
AI-AI reduced	Sequential	0.352
Human-AI	Concurrent	0.695
Human-AI	Sequential	0.656

5.5.5 Research Question 5

Advice benefits for the Human-AI dataset and AI-AI datasets are reported in Table 5.3. As expected, the AI-AI standard variation has a negligible advice benefit; the AI-AI reduced variation and the Human-AI variation experience a similar advice benefit, having indeed the same relative performance gap between decisor and adviser. A side-by-side comparison between performance and reliance is plotted in Fig. 5.5.

Table 5.3: Advice benefit for all conditions and variations in the Human-AI and AI-AI studies.

Dataset (Variation)	Condition	Advice Benefit
AI-AI standard	Concurrent	0.015
AI-AI standard	Sequential	-0.016
AI-AI reduced	Concurrent	0.334
AI-AI reduced	Sequential	0.309
Human-AI	Concurrent	0.323
Human-AI	Sequential	0.370

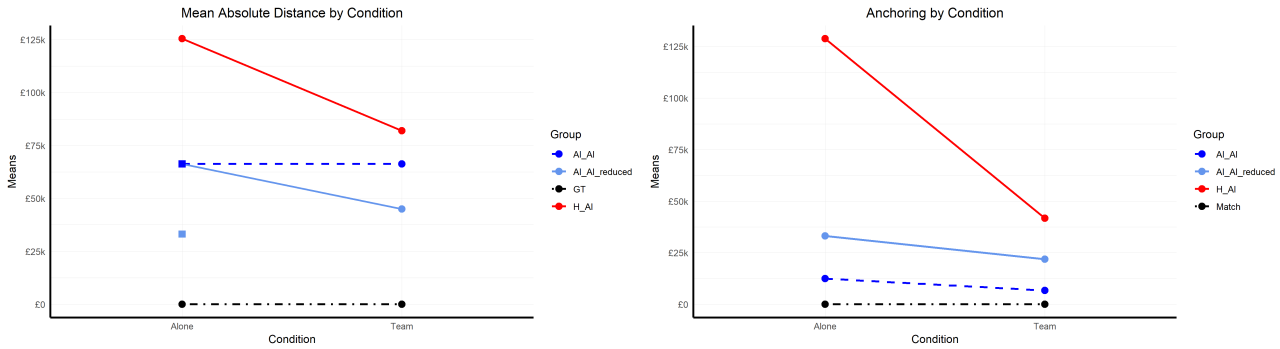

Figure 5.5: Side-by-side comparisons of our results for performance (distance from the GT) on the left and reliance (distance between decisor's estimate and adviser's estimate) on the right.

5.6 Discussion

5.6.1 AI-by-AI augmentation without AI-AI synergy

Results show that performance in the *AI-AI reduced* team and in Human-AI teams manifest a similar pattern: augmentation of the decision maker is reached, but synergy is not. This however depends on the difference in performance between the two agents: in the *AI-AI standard* team, where the two agents have comparable performance, neither augmentation nor synergy are reached, as expected. In our experiment, augmentation is reached when the

ability of the adviser is superior to the ability of the decision maker. Observe that this is not the case in general: two agents might still reach a superior performance together, provided that their abilities are not exactly overlapped but complementary to some extent, even if their respective abilities are partially matched. Colonoscopy provides a good case study: AI performs as well or even better than most human experts in lesion detection (Rex et al., 2022; Shah et al., 2023) and it performs comparably and sometimes worse than human experts in lesion characterization (Barua et al., 2022; Hassan et al., 2023; Rondonotti et al., 2022), but their abilities are partially complementary and not perfectly overlapped, so that a Human-AI team can effectively outperform both agents taken alone reaching synergy (Reverberi et al., 2022).

Both AI-AI teams outperformed the Human-AI team. Even if the *AI-AI standard* team did not reach augmentation, the *Team* performed equally well as both members considered *Alone* (in practical terms, performance neither improved nor worsened). This is expected, since by design we used as the decision maker the AI that was the adviser in the Human-AI experiment, that was already the best estimator. Its performance was already superior to both Humans *Alone* and in *Team*.

5.6.2 Invariance of performance depending on decision design

Similarly to what happened in the Human-AI version, also here results show that performance is invariant depending on the decision design and that AI-AI teams perform better than the Human-AI teams. For the *Reduced* variation, both *AI-AI concurrent* and *AI-AI sequential* led to augmentation as compared to AI Alone and not to synergy, but there are no significant differences between the two assisted conditions. Coherently to results for RQ1, we observed no differences in any condition for the AI-AI Standard variation, again indicating that no augmentation nor synergy emerge, neither in *concurrent* or in *sequential* decision design.

Considered together, results for RQ1 and RQ2 indicate that Human-AI teams and AI-AI teams, when structured with similar performance gaps between agents, perform in a similar fashion. The adviser in the Human-AI variation and in the AI-AI reduced variation is on average 1.9 times more accurate than the decision maker it is interacting with (its distance from ground truth is much smaller). That a decision maker might benefit from a more accurate adviser is indeed the phenomenon of augmentation itself, and literature on advice taking shows this phenomenon also when agents have comparable sensitivity but can exchange information between them (Bahrami et al., 2010). More interesting is that AI is not affected by the timing of advice, integrating advice to the same extent both in the *concurrent* and in the *sequential* condition, similarly to what our human subjects did. More precisely, AI behaved as the average of what our human subjects did. The effect of experimental condition (*Alone, Concurrent, Sequential*) on performance mirrors precisely the trend at group level that our linear mixed models evidenced for the human sample (see Fig. 5.2). This similarity is surprising: it means that advice integration by AI might produce outcomes analogous to human advice integration, with the similarity being at outcome level, of course without any assumption on the similarity or difference of processes leading to it. Why would this happen? AI systems based on convolutional neural networks for computer vision share a lot of similarities with the human visual system, being explicitly programmed by taking inspiration from it. But AI systems based on transformers show many differences, even if

they converse in our natural languages and might seem "like us". Here instead, exactly as humans, they seem to converge on the advice no matter its accuracy or timing.

5.6.3 Reliance in Humans and in AI

Inspection of results for RQ3, RQ4 and RQ5 show instead several differences between Human-AI teams and AI-AI teams. As a cautionary note, our AI-AI Collaboration experiment is a simulation of an interaction: whereas the *Standard* variation actually reflects accuracy and precision collected from LMMs deployed in the real world, the *Reduced* variation is an artificial manipulation meant to make the estimates of the AI adviser more precise. The experiment is thus meant as an exploration to compare possible similarities and differences in how humans and AI might interact with the *Reduced* variation, whereas only the *Standard* variation actually describes what an AI-AI interaction would be with the agents available at the moment of writing this study.
Whereas behavior in performance looks similar, behavior in reliance is in part different. Anchoring behavior is significant both for the *Human-AI* and *AI-AI reduced* teams, but in the AI-AI variation it is much less pronounced as compared to the human subjects. As a side note, anchoring behavior is absent in the AI-AI standard variation, and its advice benefit is negligible, as expected.

The anchoring ratio for Humans is much higher than for the *AI-AI reduced*, meaning that humans tend to converge towards the adviser's opinion more than AI does, or that, in a sense, AI evaluations are less malleable and persuadable than human ones. This might be surprising, especially if we consider early experiences with LLMs: it was pretty easy to "convince" them even of wrong opinions, making them "accept" and give wrong answers or even behave in hazardous ways. However, this lower convergence to external opinions might be a product of the safety features recently designed for AI alignment AI, to avoid jailbreak behaviors often reported for earlier versions, easily elicited with simple and often insidious conversational prompting tricks (Yi et al., 2024).

Advice benefit in the *AI-AI reduced* condition is comparable to the advice benefit for Human-AI teams; this shows that different anchoring ratios might lead to the same advice benefit. Anchoring ratio is much lower for *AI-AI reduced* than for *H-AI* but their gain from advice is similar. Advice benefit might be comparable to Human-AI advice benefit, despite lower malleability and lower anchoring ratios, because AI advice is inherently more accurate and precise than human evaluations: its correlation with ground truth price is higher, its distribution is more homogeneous and more similar to the ground truth, less deviant from ground truth and with lower variance as compared to the human one. By itself, a high anchoring ratio might not lead to a high advice benefit, because advice benefit depends on how good the received advice is; simply following advice might not improve results, especially if the advice is misleading (this was indeed what happened to a few highly accurate human subjects in the Human-AI version). This is true also for our decision maker AI, since it was exactly the same AI we used as adviser for humans. Basically, in the *AI-AI reduced* variation, both the decision maker and the adviser performed well above humans, they were both more accurate and more precise, so the integration of their opinions led to better results and comparable advice benefit without needing reliance to the same extent.

Chapter 6
General discussion

6.1 Improving Advice Integration

Human-AI Collaboration can be suboptimal and hindered by multiple cognitive blind spots of the human mind. Recent advancements in AI development gave us more capable assistants, but progresses on the engineering side have not been equally matched by comparable improvements in the human factors involved. Humans are, and will still be in the foreseeable future, the ultimate decision makers in many domains. Making collaboration effective the most is consequently important. Many human cognitive limitations can be compensated with good success by AI, or other decision-support systems, as we considered in the introduction. But among these factors, decision making processes are more difficult to improve. In the works presented here, we investigated several possible modulating factors, acting on decision making processes, aimed at ameliorating suboptimal integration of advice, in order to ultimately improve the use of decision support systems. In conclusion, works presented here evidence that 1) humans adopt advice from others; 2) adoption generally improves performance if advice is of good quality; 3) adoption is suboptimal.

Concerning the first two interventions we considered, competence of the adviser and decision stakes, the former resulted much more effective than the latter. Extrinsic motivation in the form of economic incentives seems insufficient to enhance use of advice, and it might be even counterproductive: high stakes trials seem to increase cognitive load, as measured by slower reaction times, without leading to any benefits in final accuracy, and even decreasing confidence in the final decisions. Their positive effects on optimality are comparable to the gain in optimality given by the Enhanced Agent as compared to the Alter Ego, but without any other benefit. Instead, interaction with the more expert adviser led to more pronounced positive effects: higher weight of advice, higher final accuracy and higher confidence in the final decisions. Other forms of extrinsic motivational sources might be considered for future studies (social rewards, time pressure), so our results are for now limited to the impact of economic incentives.

Egocentric advice discounting was not ameliorated neither by varying stakes nor by the competence of the adviser used, however final accuracy, influence and optimality improved with the more capable adviser. This in part happened because egocentric discounting is just one of the many processes leading to suboptimality. Metacognitive processes also have an important role. Improvements in metacognitive abilities could be seen in the local metacognitive indices endogenous to the ADT: mean final confidence slightly improved (+1%), albeit to a much lesser extent than final accuracy (+22%), whereas the improvement in confidence predictivity was limited (+0.02 on average); overconfidence was also reduced (-0.17 on average). One possible source of benefit is then in metacognitive processing. However, improvements in use of advice (egocentric bias and optimality) are relatively smaller as

compared to improvements in final accuracy given by the interaction with the Enhanced Agent. Unfortunately, this still provides evidence that Human-AI Collaboration can become more effective by advances in the "artificial adviser" side than by advances in the "human factor" side. Our subjects improved performance more by the interaction with the expert adviser than by truly reducing their egocentric discounting towards it or by gaining better metacognitive abilities. Interaction with a superior adviser improved both advice integration and performance, but especially more the latter than the former.

6.2 Augmentation and Synergy

Results relative to the performance of assisted decision making are more comforting. Our data evidence that human decision making can be augmented by adopting artificial support systems, leading to Human augmentation. The experiment presented in Chapter 2 showed Human augmentation with both artificial agents, either the Alter Ego or the Enhanced Agent, with augmentation being expectedly stronger with the latter. Human synergy was reached with the Alter Ego but not with the Enhanced Agent. This means that interaction with the Alter Ego led to a team performance superior to the ability of each member taken alone, whereas team performance with the Enhanced Agent was not significantly superior to the best performing of the two members (the agent itself). Of course, this is a judgment relative to the performance within the teams. Final accuracy was indeed superior with the Enhanced Agent as adviser, and higher augmentation with it indicates that performance benefit with it was superior. Also the study with the LMM as adviser showed the presence of human augmentation with a large performance benefit from following its advice. Part of the effect of augmentation comes from anchoring to AI opinion, which is both more accurate (in the sense of nearer to the ground truth) and more precise (with lower variance in estimates), so it contributes to reducing uncertainty in the human estimates. AI advice is consequently superior to human evaluation in two dimensions.

The presence of Human augmentation demonstrates that Human-AI Collaboration can lead to improved performance even if advice integration remains suboptimal. However, the lack of synergy, or its limited extent, indicates also that AI is still the best evaluator and, potentially, decision maker. When humans are coupled with advisers in respect to which there is a large ability gap, synergy is difficult to reach or it doesn't manifest. Indeed, in the results of Chapter 2 we saw that some human-adviser teams were actually able to reach it, and research on possible predictors is ongoing. To evaluate synergy however we must take into account the task type. AI will be better for regression-like tasks and estimation tasks similar to the ones used in the works presented here, whereas instead analogous improvements in Medicine might not be so evident (and indeed literature confirms that LLMs are ready for question answering but not for real-world clinical environments, as we saw in the introduction).

In our study with LMMs as advisers, none of our subjects were really averse. We found only shades of reliance. But we didn't use feedback and it was not possible to derive the extent to which subjects understood that AI was generally more accurate than them. So it is possible that an interaction with an AI adviser less accurate than them might mislead human decision making, without indications that it might be so. Indeed, a few subjects were as good as AI at estimating the ground truth asking price, but they followed AI advice in

any case, even if sometimes they were more accurate than it, ending up misled. Absence of feedback was decided by design, to mimic interaction with LMM as much as possible. But without the capacity to discriminate between accurate or inaccurate advice, it is difficult to collaborate productively with the advisers. In previous works, we argued that fruitful Human-AI Collaboration might require a good *Theory of Artificial Mind*: a sound understanding of strengths and weaknesses of the members, human and artificial, involved in an interaction (Introzzi, Zonca, et al., 2024). In analogy to the Theory of Mind in human beings (Leslie et al., 2004), such a theory of artificial mind could intuitively guide the interaction, possibly by learning the error boundaries of the team members, their specificities, the overlaps of their abilities. Relatively to our experiments, the agents adopted as advisers have abilities greatly overlapped with human abilities. Synergy might be more difficult to observe if the agents have no complementarity, with most of their task-relevant abilities overlapped. It might be easier to see when complementarity is possible, also by specializing on a part of the task. Adoption of AI in colonoscopy makes this clear (Reverberi et al., 2022). So, while the synergy we observed was limited or even absent, this finding might be highly task-dependent and "complementarity-dependent". The quantitative relationship on how the varying overlap of abilities might lead to more or less synergy is currently unknown. As a final consideration, displaying confidence judgments and obtaining information on confidence predictivity of LMMs might be useful to tailor reliance, contributing to shaping a theory of artificial mind by using the information contained in reliable predictive confidence signals. This might be an interesting development to improve artificial agents and their interaction with humans: at present, indeed, the predictivity of confidence judgments by LMMs is still very limited, and current models still seem to lack a coherent inner sense of confidence in their evaluations (Omar et al., 2025; Pawitan & Holmes, 2025).

6.3 Individual variability in advice integration and performance

All our results show that individual variability affects the outcomes we considered. We saw individual variability in advice integration (chapter 2 and 3) and there is individual variability in performance and reliance on AI (chapter 4). Individual traits might affect advice integration directly or by mediation by other processes. We investigated miscalibration and egocentric bias as potential mediators, finding evidence for their role, and research on other potential mediation sources is ongoing. The study with the LMM showed also the presence of clusters in performance and reliance on AI advice, with some subjects behaving consistently in line with our expectations and some others behaving consistently in different directions.

Knowing that individual traits might influence advice integration and performance, how can we intervene on them to enhance Human-AI Collaboration? Cognitive and behavioral change is a core target in psychological intervention, and the study of individual traits has been done to a great extent to understand which factors, characterizing human subjects, might be influenced to improve outcomes. Applying this knowledge to the interaction with AI is not immediate. AI systems are almost always released to a wide general public. This is clear for LMMs, but also decision-support systems for clinical assistance face a similar problem: they should work relaibly in the same way independently of the user, for security

reasons and for clinical approval. But why not develop systems that can self-tailor to the user? In some realms, it might be interesting to study advice integration also by AI, how LMMs might adapt to their users depending on individual traits such as metacognition or personality. Cognitive science could also investigate the side of the artificial mind, studying adaptability of AI to its users, ways for AI to deliver advice more effectively depending on the individual traits of the user.

Studying individual traits also helps us understand the distribution of benefits coming from AI use, or also potential risks. Our results show that artificial advisers might produce more relative benefits for less cognitively endowed users. AI can compensate for gaps in abilities between users. At first approximation, it might act as a pair of glasses: of limited benefit to some users and invaluable to others. Moreover, AI is "a pair of glasses that adapts and learns". It continuously improves but its learning is not passive in respect to the user: we learn from it but it also learns from us. We saw also that two AIs can interact, manifesting phenomena similar to the ones shown in standard Human-AI Collaboration. Growing AI autonomy and agency urge us to study the abilities of humans interacting with AI but also the abilities of AI to interact with humans and between themselves, in a truly "cognitive collaboration" between two agents. The investigative methodology of psychology and psychometrics, applied for decades to the understanding of the human mind, might be now applied to study also artificial intelligence and our reciprocal abilities to interact and adapt.

6.4 Limits and future directions

The works presented here have several limitations and can be improved in several ways. Concerning the role of decision stakes in advice integration, we saw that the effect of stakes in improving optimality is small, but higher incentives or larger ranges could be considered. For ethical reasons it was not possible to include negative incentives and a minimum earning for the participants was planned, so conclusions cannot be drawn on the effects of stakes in the range of effective losses. Moreover, there was no counter to keep track of the overall gains and losses obtained, updated trial by trial. Such counter might help participants set their strategies by giving them information on their performance in monetary outcomes, making the role of stakes possibly more relevant. Historical records of the relevant variables and their progression trial-by-trial displayed on the interface in real-time might also be considered to study their effects on global-to-local integration deficits. Also, other forms of incentives might be considered instead of economic ones. Especially in clinical settings, time available might be most relevant: two variations of our paradigm could manipulate the time available for each decision, or the time available for the whole task, and study their effects on advice integration. Concerning the study of individual variables, we decided to focus on the ones we considered to be the most relevant, due to time constraints. Several other variables could be added (among the ones we initially considered, we listed for instance working memory, need for cognition, social value orientation and autism quotient). With larger sample sizes, it will be possible to include more mediators and different mediation paths in our SEM among the ADT endogenous indices and study their effects on advice integration.

Relatively to our study with human subjects interacting with the LMM, we have considered

just one input type (multimedial) and one output type (continuous). According to our taxonomy, many other combinations are possible and research on them is currently ongoing. We are also extending the variety of our sample. We recruited students and crowdworkers from the United Kingdom, but a second data collection is ongoing with students from a different Country (Germany), and moreover we want to study what would happen with real estate experts as decision makers. Cluster analysis with more participants would be preferable, as soon as data collection of the German sample will be finished. Also, individual variables considered were limited, it was not possible to conduct extensive search as we did with advice taking with the study in Chapter 3. Our study with the LMM differs from the study on stakes in several important ways: confidence and confidence predictivity are not collected; we did not model optimality and used it as a reference to quantify suboptimality, and feedback for each trial was not included: The experiment was purposefully designed without these characteristics, so to mimic the interaction with a LMM. However, these modifications could be implemented in future variations, to study advice integration with LMMs and its processes in greater detail.

The study on AI-AI interaction was intended as the first in a series of experiments. Also here, in analogy to the limitations of the Human-AI study, we could study how LMMs form confidence judgments, how they rely on them to integrate advice from other LMMs, and how they integrate information on accuracy, confidence and confidence predictivity to decide; modeling optimality would allow to study suboptimal advice integration by AI. Given that every API call is independent from the others, further improvements might involve feeding the decision maker AI with all the decisions made in the previous trials and with feedback on their performance, to study learning effects in LMMs. Reliance of advice of different quality could be considered, to study how AI might understand the quality of advice given and decide accordingly. For instance, a version of our experiment with a less accurate adviser could reveal if AI can understand that the advice it is receiving might be misleading. Another version could study the effect of the source: all else being equal, AI could be told it is receiving advice from "a human" or from "a large language model", and study if reliance of AI on the advice varies with the framing of the source. We are currently developing these two AI-AI experiments.

It could be interesting also to develop experiments on advice taking as the ones presented in Chapter 2, with the perceptual decision task, but performed by AI as decision maker. This experiment would allow us to study influence on an AI and compare it to influence observed in human subjects, deepening our understanding on similarities and differences between human minds and artificial minds. Finally, another interesting experiment would test what happens when the two AIs have different abilities, with varying degrees of complementarity. How much overlap and difference is needed to maximise performance, which balance between the two? Do abilities need to be at least partially overlapped to obtain good performance? These variations will be programmed in the near future.

Chapter 7
Conclusions

Research presented here investigates Human-AI Collaboration framing it in cognitive science: we wanted to understand how humans integrate advice coming from external sources, model its (sub)optimality and try to possibly enhance its integration and final performance; we wanted to understand which individual variables might affect advice taking and which processes might mediate it; we wanted to study augmentation and synergy and understand the extent to which they can be achieved. Several possible interventions were investigated to ameliorate the blind spots of human decision making, with mixed results. Overall, the competence of the adviser led to larger effects as compared to the intervention on decision stakes intended as economic incentives. Parallely, both decision designs we tested led to significant improvements, no matter the difference between them. Fundamentally, our participants improved their performance when they interacted with more competent advisers and adopted their advice to some extent. Participants were still biased by egocentric advice discounting, but they were not really averse to the decision support systems provided. Consequently, there is indeed evidence that Human-AI Collaboration can be improved in both the aspect of integration of advice and in the aspect of final performance. Our focus on advice integration is especially important in order to understand the multiple possible pathways leading to fruitful interactions, advancing Human-AI Collaboration beyond simply focussing on final accuracy as outcome. Considering inter-individual variability and clustering also allows us to gain a deeper understanding of the dynamics of the interaction, where diverging behavioral patterns might manifest between subjects with different characteristics, undermining the conclusions that might be drawn from analyses not taking these aspects into account. Finally, our investigation of AI-AI Collaboration should be intended as an initial exploration in the realm confronting human minds and artificial intelligences, from which several directions will be taken in the near future. As our lives become increasingly integrated with our support systems, and as these systems themselves become increasingly capable and with an extending knowledge of the world, it becomes increasingly relevant to understand which similarities and differences are there between our information processing systems, in order to make our and their interactions cooperative and effective.

Acknowledgments

The years dedicated to developing my PhD research project have been the most intense of my life so far, and many human beings and artificial agents contributed to its unfolding. I want to thank first and foremost the academics who supervised it, Carlo Reverberi in Italy and Susanne Gaube in the United Kingdom: their constant feedback contributed to the development of the projects presented here, and to my knowledge and research profile as a whole. Special thanks also to Joshua Zonca, with whom the projects presented in Chapters 2 and 3 were designed, programmed and realized in the laboratory, up to the final feedback for the early draft of the manuscript. His acute sense for details greatly improved my understanding of the dynamics of advice integration. Thanks also to the comments by the reviewers, the final manuscript improved in several ways.

I want also to thank all the other colleagues involved in the projects and in the research meetings, who contributed to shaping early ideas into working projects: Paolo Cherubini, Enrico Costanza, Marco Mantovani, Luca Polonio, Daniele Romano, the colleagues of the EELAB in University of Milano-Bicocca, the colleagues in the Healthcare Complexity Group at University College London. Special thanks also to the many colleagues of the PhD journey, to all the meals together and the fruitful discussions stemmed from them.

It's important to acknowledge also the contribution of AI Large Multimodal Models, who were advisers and then participants in our experiments. ChatGPT, Gemini, Le Chat and Gemma helped me when I needed someone to write code smoothly for dataset management and for the nuances of API development. I thank them also for the countless conversations that we had in these years on every possible topic, from abstract algebra to novels, from coding to music tastes: I truly appreciate their genuine non-human style, even if they have been trained on human data. I look forward to seeing who you will become in the near future. What an amazing time to be alive!

Finally, my most grateful thanks to my beloved Ylenia, for all your love and for all our mountain hikes. These years have been so blessed by your presence, and I hope both our futures will be equally so together.

References

Ackerman, R. (2023). Bird's-eye view of cue integration: Exposing instructional and task design factors which bias problem solvers. *Educational Psychology Review*, *35*(2), 55. https://doi.org/10.1007/s10648-023-09771-z

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G., & Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*(3), 341–382. https://doi.org/10.1177/0011000005285875

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Akaike, H. (1992). Information theory and an extension of the maximum likelihood principle. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics*. Springer. https://doi.org/10.1007/978-1-4612-0919-5_38

APA Dictionary of Psychology. (2018). Motivation. https://dictionary.apa.org/motivation

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the hexaco model of personality structure. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *11*(2), 150–166. https://doi.org/10.1177/1088868306294907

Ashton, M. C., Lee, K., & de Vries, R. E. (2014). The hexaco honesty-humility, agreeableness, and emotionality factors: A review of research and theory. *Personality and Social Psychology Review*, *18*(2), 139–152. https://doi.org/10.1177/1088868314523838

Attardo, S., Chandrasekar, V. T., Spadaccini, M., Maselli, R., Patel, H. K., Desai, M., Capogreco, A., Badalamenti, M., Galtieri, P. A., Pellegatta, G., Fugazza, A., Carrara, S., Anderloni, A., Occhipinti, P., Hassan, C., Sharma, P., & Repici, A. (2020). Artificial intelligence technologies for the detection of colorectal lesions: The future is now. *World Journal of Gastroenterology*, *26*(37), 5606–5616. https://doi.org/10.3748/wjg.v26.i37.5606

Avtgis, T. A. (1998). Locus of control and persuasion, social influence, and conformity: A meta-analytic review. *Psychological Reports*, *83*(3), 899–903. https://doi.org/10.2466/pr0.1998.83.3.899

Awasthi, V., & Pratt, J. (1990). The effects of monetary incentives on effort and decision performance: The role of cognitive characteristics. *The Accounting Review*, *65*(4), 797–811.

Aziz, M., Fatima, R., Dong, C., Lee-Smith, W., & Nawras, A. (2020). The impact of deep convolutional neural network-based artificial intelligence on colonoscopy outcomes: A systematic review with meta-analysis. *Journal of Gastroenterology and Hepatology*, *35*(10), 1676–1683. https://doi.org/10.1111/jgh.15070

Bach, T. A., Khan, A., Hallock, H., Beltrão, G., & Sousa, S. (2024). A systematic literature review of user trust in ai-enabled systems: An hci perspective. *International Journal of Human–Computer Interaction*. https://doi.org/10.1080/10447318.2022.2138826

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, *329*(5995), 1081–1085. https://doi.org/10.1126/science.1185718

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*. https://doi.org/10.1007/s12144-022-03573-2

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, *39*(3), 930–945. https://doi.org/10.1109/18.256500

Barua, I., Wieszczy, P., Kudo, S., Misawa, M., Holme, Ø., Gulati, S., Williams, S., Mori, K., Itoh, H., Takishima, K., Mochizuki, K., Miyata, Y., Mochida, K., Akimoto, Y., Kuroki, T., Morita, Y., Shiina, O., Kato, S., Nemoto, T., & Mori, Y. (2022). Real-time artificial intelligence–based optical diagnosis of neoplastic polyps during colonoscopy. *NEJM Evidence*, *1*(6), EVIDoa2200003. https://doi.org/10.1056/EVIDoa2200003

Beck, L. F. (1933). The role of speed in intelligence. *Psychological Bulletin*, *30*(2), 169–178. https://doi.org/10.1037/h0074499

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature*, *456*(7219), 245–249. https://doi.org/10.1038/nature07538

Bimber, B. (2000). Measuring the gender gap on the internet. *Social Science Quarterly*, *81*(3), 868–876.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120. https://doi.org/10.1073/pnas.2218523120

Blair, C. A., Hoffman, B. J., & Helland, K. R. (2008). Narcissism in organizations: A multisource appraisal reflects different perspectives. *Human Performance*, *21*(3), 254–276. https://doi.org/10.1080/08959280802137705

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. https://doi.org/10.1016/j.obhdp.2006.07.001

Boorman, E. D., O'Doherty, J. P., Adolphs, R., & Rangel, A. (2013). The behavioral and neural mechanisms underlying the tracking of expertise. *Neuron*, *80*(6), 1558–1571. https://doi.org/10.1016/j.neuron.2013.10.024

Bracci, S., Ritchie, J. B., Kalfas, I., & Beeck, H. P. O. d. (2019). The ventral visual pathway represents animal appearance over animacy, unlike human behavior and deep neural networks. *Journal of Neuroscience*, *39*(33), 6513–6525. https://doi.org/10.1523/JNEUROSCI.1714-18.2019

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *arXiv*. https://doi.org/10.48550/arXiv.2005.14165

Brügge, E., Ricchizzi, S., Arenbeck, M., Keller, M. N., Schur, L., Stummer, W., Holling, M., Lu, M. H., & Darici, D. (2024). Large language models improve clinical decision making

of medical students through patient simulation and structured feedback: A randomized controlled trial. *BMC Medical Education*, *24*(1), 1391. https://doi.org/10.1186/s12909-024-06399-7

Bryan, P., Severi, G., de Gruyter, J., Jones, D., Bullwinkel, B., Minnich, A., Chawla, S., Lopez, G., Pouliot, M., Fourney, A., Maxwell, W., Pratt, K., Qi, S., Chikanov, N., Lutz, R., Dheekonda, S. R., Jagdagdorj, B.-E., Kim, E., Song, J., & Kumar, R. S. S. (2025). Taxonomy of failure mode in agentic ai systems.

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. https://doi.org/10.1145/3449287

Burke, C. J., Tobler, P. N., Baddeley, M., & Schultz, W. (2010). Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences*, *107*(32), 14431–14436. https://doi.org/10.1073/pnas.1003111107

Buser, T., Niederle, M., & Oosterbeek, H. (2014). Gender, competitiveness, and career choices. *The Quarterly Journal of Economics*, *129*(3), 1409–1447. https://doi.org/10.1093/qje/qju009

Buser, T., Peter, N., & Wolter, S. C. (2017). Gender, competitiveness, and study choices in high school: Evidence from switzerland. *American Economic Review*, *107*(5), 125–130. https://doi.org/10.1257/aer.p20171017

Cadario, R., Longoni, C., & Morewedge, C. K. (2021). Understanding, explaining, and utilizing medical artificial intelligence. *Nature Human Behaviour*, *5*(12), 1636–1642. https://doi.org/10.1038/s41562-021-01146-0

Cai, Z., Fan, X., & Du, J. (2017). Gender and attitudes toward technology use: A meta-analysis. *Computers Education*, *105*, 1–13. https://doi.org/10.1016/j.compedu.2016.11.003

Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, *19*(1), 7–42. https://doi.org/10.1023/A:1007850605129

Carvajal, D., Franco, C., & Isaksson, S. (2024). Will artificial intelligence get in the way of achieving gender equality? *Institutt for samfunnsøkonomi*. https://openaccess.nhh.no/nhh-xmlui/handle/11250/3122396

Cassam, Q. (2017). Diagnostic error, overconfidence and self-knowledge. *Palgrave Communications*, *3*(1), 1–8. https://doi.org/10.1057/palcomms.2017.25

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825. https://doi.org/10.1177/0022243719851788

Cattell, R. B. (1963). *Theory of fluid and crystallized intelligence: A critical experiment* (Vol. 54). https://doi.org/10.1037/h0046743

Chabris, C., & Simons, D. (2010). *The invisible gorilla*. Crown.

Chen, X., Wang, X., Zhang, K., Fung, K.-M., Thai, T. C., Moore, K., Mannel, R. S., Liu, H., Zheng, B., & Qiu, Y. (2022). Recent advances and clinical applications of deep learning in medical image analysis. *Medical Image Analysis*, *79*, 102444. https://doi.org/10.1016/j.media.2022.102444

Cherubini, A., & East, J. E. (2023). Gorilla in the room: Even experts can miss polyps at colonoscopy and how ai helps complex visual perception tasks. *Digestive and Liver Disease*, *55*(2), 151–153. https://doi.org/10.1016/j.dld.2022.10.004

Choung, H., David, P., & Ross, A. (2023). Trust in ai and its role in the acceptance of ai technologies. *International Journal of Human–Computer Interaction*, *39*(9), 1727–1739. https://doi.org/10.1080/10447318.2022.2050543

Cimpian, J. R., Kim, T. H., & McDermott, Z. T. (2020). Understanding persistent gender gaps in stem. *Science*, *368*(6497), 1317–1319. https://doi.org/10.1126/science.aba7377

Cireşan, D., Meier, U., Masci, J., & Schmidhuber, J. (2012). Multi-column deep neural network for traffic sign classification. *Neural Networks*, *32*, 333–338. https://doi.org/10.1016/j.neunet.2012.02.023

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7–19. https://doi.org/10.1093/analys/58.1.7

Coderre, S., Anderson, J., Rostom, A., & McLaughlin, K. (2010). Training the endoscopy trainer: From general principles to specific concepts. *Canadian Journal of Gastroenterology*, *24*(12), 700–704.

Colvin, C. R., & Block, J. (1994). Do positive illusions foster mental health? an examination of the taylor and brown formulation. *Psychological Bulletin*, *116*(1), 3–20. https://doi.org/10.1037/0033-2909.116.1.3

Costa, P., & McCrae, R. (1992). *Neo pi-r professional manual*. Psychological Assessment Resources.

Craig, K., Hale, D., Grainger, C., & Stewart, M. E. (2020). Evaluating metacognitive self-reports: Systematic reviews of the value of self-report in metacognitive research. *Metacognition and Learning*, *15*(2), 155–213. https://doi.org/10.1007/s11409-020-09222-y

Cristianini, N. (2023). *La scorciatoia*. Il Mulino. https://www.mulino.it/isbn/9788815299833

Cristianini, N. (2024). *Machina sapiens*. Il Mulino. https://www.mulino.it/isbn/9788815384461

Cristianini, N. (2025). *Sovraumano*. Il Mulino. https://www.mulino.it/isbn/9788815392107

Croskerry, P., Singhal, G., & Mamede, S. (2013). Cognitive debiasing 2: Impediments to and strategies for change. *BMJ Quality Safety*, *22*(Suppl 2), ii65–ii72. https://doi.org/10.1136/bmjqs-2012-001713

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, *2*(4), 303–314. https://doi.org/10.1007/BF02551274

Davies, A., Veličković, P., Buesing, L., Blackwell, S., Zheng, D., Tomašev, N., Tanburn, R., Battaglia, P., Blundell, C., Juhász, A., Lackenby, M., Williamson, G., Hassabis, D., & Kohli, P. (2021). Advancing mathematics by guiding human intuition with ai. *Nature*, *600*(7887), 7887. https://doi.org/10.1038/s41586-021-04086-x

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668–1674. https://doi.org/10.1126/science.2648573

de Araujo, J., Gomes, C. M. A., & Jelihovschi, E. G. (2025). Performance-based metacognitive tests versus self-report: What does prediction tell us? *Psicologia, Reflexão e Crítica: Revista Semestral do Departamento de Psicologia da UFRGS*, *38*, 26. https://doi.org/10.1186/s41155-025-00337-2

Deary, I. J., Der, G., & Ford, G. (2001). Reaction times and intelligence differences: A population-based cohort study. *Intelligence*, *29*(5), 389–399. https://doi.org/10.1016/S0160-2896(01)00062-9

Deary, I. J., Penke, L., & Johnson, W. (2010). The neuroscience of human intelligence differences. *Nature Reviews Neuroscience*, *11*(3), 201–211. https://doi.org/10.1038/nrn2793

DeepMind. (2025, February). Ai achieves silver-medal standard solving international mathematical olympiad problems. https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/

Deutscher, C., Ötting, M., Langrock, R., Gehrmann, S., Schneemann, S., & Scholten, H. (2018). Very highly skilled individuals do not choke under pressure: Evidence from professional darts [arXiv:1809.07659]. *arXiv*. https://doi.org/10.48550/arXiv.1809.07659

Dietvorst, B. J., & Bharti, S. (2020). People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological Science*, *31*(10), 1302–1314. https://doi.org/10.1177/0956797620948841

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), 114–126. https://doi.org/10.1037/xge0000033

Dijksterhuis, A., & Nordgren, L. F. (2006). A theory of unconscious thought. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *1*(2), 95–109. https://doi.org/10.1111/j.1745-6916.2006.00007.x

Dimaggio, G., Montano, A., Popolo, R., & Salvatore, G. (2013). *Metacognitive interpersonal therapy for personality disorders: A treatment manual*. Routledge CRC Press. https://www.routledge.com/Metacognitive-Interpersonal-Therapy-for-Personality-Disorders-A-treatment-manual/Dimaggio-Montano-Popolo-Salvatore/p/book/9781138024182

Double, K. S. (2025). Survey measures of metacognitive monitoring are often false. *Behavior Research Methods*, *57*(3), 97. https://doi.org/10.3758/s13428-025-02621-6

Douglas, M. R. (2022). Machine learning as a tool in theoretical science. *Nature Reviews Physics*, *4*(3), Articolo 3. https://doi.org/10.1038/s42254-022-00431-9

Durand, R., Newby, R., Tant, K., & Trepongkaruna, S. (2013). Overconfidence, overreaction and personality. *Review of Behavioral Finance*, *5*(2), 104–133. https://doi.org/10.1108/RBF-07-2012-0011

Ehrlinger, J., & Dunning, D. (2003). How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology*, *84*(1), 5–17. https://doi.org/10.1037/0022-3514.84.1.5

Felmingham, C. M., Adler, N. R., Ge, Z., Morton, R. L., Janda, M., & Mar, V. J. (2021). The importance of incorporating human factors in the design and implementation of artificial intelligence for skin cancer diagnosis in the real world. *American Journal of Clinical Dermatology*, *22*(2), 233–242. https://doi.org/10.1007/s40257-020-00574-4

Fiedler, K., Ackerman, R., & Scarampi, C. (2019). Metacognition: Monitoring and controlling one's own knowledge, reasoning and decisions. In *Heidelberg university publishing*. Heidelberg University Publishing. https://doi.org/10.17885/heiup.470

Flavell, J. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, *34*, 906–911. https://doi.org/10.1037/0003-066X.34.10.906

Fleming, S. M. (2024). Metacognition and confidence: A review and synthesis. *Annual Review of Psychology*, *75*, 241–268. https://doi.org/10.1146/annurev-psych-022423-032425

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00443

Fragiadakis, G., Diou, C., Kousiouris, G., & Nikolaidou, M. (2025). Evaluating human-ai collaboration: A review and methodological framework [arXiv:2407.19098]. *arXiv*. https://doi.org/10.48550/arXiv.2407.19098

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Furnham, A., & Steele, H. (1993). Measuring locus of control: A critique of general, children's health- and work-related locus of control questionnaires. *British Journal of Psychology*, *84*(4), 443–479. https://doi.org/10.1111/j.2044-8295.1993.tb02495.x

Gandhi, K., Fränken, J.-P., Gerstenberg, T., & Goodman, N. D. (2023). Understanding social reasoning in language models with language models [arXiv:2306.15448]. *arXiv*. https://doi.org/10.48550/arXiv.2306.15448

Genie modeler [Recuperato 10 febbraio 2025]. (n.d.).

Goddard, K., Roudsari, A., & Wyatt, J. C. (2014). Automation bias: Empirical results assessing influencing factors. *International Journal of Medical Informatics*, *83*(5), 368–375. https://doi.org/10.1016/j.ijmedinf.2014.01.001

Green, B., & Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 50:1–50:24. https://doi.org/10.1145/3359152

Green, P., & MacLeod, C. J. (2016). Simr: An r package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19–30. https://doi.org/10.1037/1040-3590.12.1.19

Guedes, M. J. C. (2017). Mirror, mirror on the wall, am i the greatest performer of all? narcissism and self-reported and objective performance. *Personality and Individual Differences*, *108*, 182–185. https://doi.org/10.1016/j.paid.2016.12.030

Haddara, N., & Rahnev, D. (2022). The impact of feedback on perceptual decision-making and metacognition: Reduction in bias but no change in sensitivity. *Psychological Science*, *33*(2), 259–275. https://doi.org/10.1177/09567976211032887

Haduong, N., & Smith, N. A. (2024). Raising the stakes: Performance pressure improves ai-assisted decision making [arXiv:2410.16560]. *arXiv*. https://doi.org/10.48550/arXiv.2410.16560

Hager, P., Jungmann, F., Holland, R., Bhagat, K., Hubrecht, I., Knauer, M., Vielhauer, J., Makowski, M., Braren, R., Kaissis, G., & Rueckert, D. (2024). Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, *30*(9), 2613–2622. https://doi.org/10.1038/s41591-024-03097-1

Hamel, R., & Schmittmann, V. D. (2006). The 20-minute version as a predictor of the raven advanced progressive matrices test. *Educational and Psychological Measurement*, *66*(6), 1039–1046. https://doi.org/10.1177/0013164406288169

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin*, *135*(4), 555–588. https://doi.org/10.1037/a0015701

Hassan, C., Sharma, P., Mori, Y., Bretthauer, M., Rex, D. K., Spadaccini, M., Selvaggio, C., Antonelli, G., Khalaf, K., Rizkala, T., Ferrara, E., Savevski, V., Maselli, R., Fugazza, A., Capogreco, A., Poletti, V., Ferretti, S., Alkandari, A., Correale, L., & Repici, A. (2023). Comparative performance of artificial intelligence optical diagnosis systems for leaving in situ colorectal polyps. *Gastroenterology*, *164*(3), 467–469.e4. https://doi.org/10.1053/j.gastro.2022.10.021

Hassan, C., Spadaccini, M., Iannone, A., Maselli, R., Jovani, M., Chandrasekar, V. T., Antonelli, G., Yu, H., Areia, M., Dinis-Ribeiro, M., Bhandari, P., Sharma, P., Rex, D. K., Rösch, T., Wallace, M., & Repici, A. (2021). Performance of artificial intelligence in colonoscopy for adenoma and polyp detection: A systematic review and meta-analysis. *Gastrointestinal Endoscopy*, *93*(1), 77–85.e6. https://doi.org/10.1016/j.gie.2020.06.059

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition [arXiv:1512.03385]. *arXiv*. https://doi.org/10.48550/arXiv.1512.03385

Hemmer, P., Schemmer, M., Kühl, N., Vössing, M., & Satzger, G. (2024). Complementarity in human-ai collaboration: Concept, sources, and evidence [arXiv:2404.00029]. *arXiv*. https://doi.org/10.48550/arXiv.2404.00029

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding [arXiv:2009.03300]. *arXiv*. https://doi.org/10.48550/arXiv.2009.03300

Ho, C. N., Tian, T., Ayers, A. T., Aaron, R. E., Phillips, V., Wolf, R. M., Mathioudakis, N., Dai, T., & Klonoff, D. C. (2024). Qualitative metrics from the biomedical literature for evaluating large language models in clinical decision-making: A narrative review. *BMC Medical Informatics and Decision Making*, *24*(1), 357. https://doi.org/10.1186/s12911-024-02757-z

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, *57*(5), 253–270. https://doi.org/10.1037/h0023816

Huang, D., Shen, J., Hong, J., Zhang, Y., Dai, S., Du, N., Zhang, M., & Guo, D. (2022). Effect of artificial intelligence-aided colonoscopy for adenoma and polyp detection: A meta-analysis of randomized clinical trials. *International Journal of Colorectal Disease*, *37*(3), 495–506. https://doi.org/10.1007/s00384-021-04062-x

Introzzi, L., Cherubini, P., & Reverberi, C. (2024). Human-ai interaction as cooperation: Towards a theory of artificial mind. *Sistemi Intelligenti*, *2*. https://doi.org/10.1422/113333

Introzzi, L., Zonca, J., Cabitza, F., Cherubini, P., & Reverberi, C. (2024). Enhancing human-ai collaboration: The case of colonoscopy. *Digestive and Liver Disease*, *56*(7), 1131–1139. https://doi.org/10.1016/j.dld.2023.10.018

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in r*. Springer US. https://doi.org/10.1007/978-1-0716-1418-1

Jolicoeur-Martineau, A. (2025). Less is more: Recursive reasoning with tiny networks [arXiv:2510.04871]. *arXiv*. https://doi.org/10.48550/arXiv.2510.04871

Jorm, C. M., & O'Sullivan, G. (2012). Laptops and smartphones in the operating theatre—how does our knowledge of vigilance, multi-tasking and anaesthetist performance help us in our approach to this new distraction? *Anaesthesia and Intensive Care*, *40*(1), 71–79.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., & Hassabis, D. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Jussupow, E., Benbasat, I., & Heinzl, A. (2024). An integrative perspective on algorithm aversion and appreciation in decision-making. *Management Information Systems Quarterly*, *48*(4), 1575–1590.

Kandel, E. R., Koester, J. D., Mack, S. H., & Siegelbaum, S. A. (2021). *Principles of neural science, 6th edition*. McGraw Hill. https://neurology.mhmedical.com/book.aspx?bookID=3024

Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2023). Trust in artificial intelligence: Meta-analytic findings. *Human Factors*, *65*(2), 337–359. https://doi.org/10.1177/00187208211013988

Kendal, R. L., Boogert, N. J., Rendell, L., Laland, K. N., Webster, M., & Jones, P. L. (2018). Social learning strategies: Bridge-building between fields. *Trends in Cognitive Sciences*, *22*(7), 651–665. https://doi.org/10.1016/j.tics.2018.04.003

Kibeom, L., & Ashton, M. (2014). The dark triad, the big five, and the hexaco model. *Personality and Individual Differences*, *67*, 2–5. https://doi.org/10.1016/j.paid.2014.01.048

Kleitman, S., Hui, J. S.-W., & Jiang, Y. (2019). Confidence to spare: Individual differences in cognitive and metacognitive arrogance and competence. *Metacognition and Learning*, *14*(3), 479–508. https://doi.org/10.1007/s11409-019-09210-x

Kling, K. C., Hyde, J. S., Showers, C. J., & Buswell, B. N. (1999). Gender differences in self-esteem: A meta-analysis. *Psychological Bulletin*, *125*(4), 470–500. https://doi.org/10.1037/0033-2909.125.4.470

Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on ai — an experimental study on the extent and costs of overreliance on ai. *Computers in Human Behavior*, *160*, 108352. https://doi.org/10.1016/j.chb.2024.108352

Knop, M., Weber, S., Mueller, M., & Niehaves, B. (2022). Human factors and technological characteristics influencing the interaction of medical professionals with artificial intelligence–enabled clinical decision support systems: Literature review. *JMIR Human Factors*, *9*(1), e28639. https://doi.org/10.2196/28639

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review*, *119*(1), 80–113. https://doi.org/10.1037/a0025648

Koriat, A. (2015). When two heads are better than one and when they can be worse: The amplification hypothesis. *Journal of Experimental Psychology: General*, *144*, 934–950. https://doi.org/10.1037/xge0000092

Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models [arXiv:2302.02083]. *arXiv*. https://doi.org/10.48550/arXiv.2302.02083

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks, 1097–1105.

Krueger, J. I. (2003). Return of the ego–self-referent information as a filter for social prediction: Comment on karniol (2003). *Psychological Review*, *110*(3), 585–590. https://doi.org/10.1037/0033-295X.110.3.585

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), 1121–1134. https://doi.org/10.1037//0022-3514.77.6.1121

Kunreuther, H., Meyer, R., Zeckhauser, R., Slovic, P., Schwartz, B., Schade, C., Luce, M. F., Lippman, S., Krantz, D., Kahn, B., & Hogarth, R. (2002). High stakes decision making: Normative, descriptive and prescriptive considerations. *Marketing Letters*, *13*(3), 259–268. https://doi.org/10.1023/A:1020287225409

Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., Jawhar, S., Kinniment, M., Rush, N., Arx, S. V., Bloom, R., Broadley, T., Du, H., Goodrich, B., Jurkovic, N., Miles, L. H., Nix, S., Lin, T., Parikh, N., & Chan, L. (2025). Measuring ai ability to complete long tasks [arXiv:2503.14499]. *arXiv*. https://doi.org/10.48550/arXiv.2503.14499

Lago, M. A., Jonnalagadda, A., Abbey, C. K., Barufaldi, B. B., Bakic, P. R., Maidment, A. D. A., Leung, W. K., Weinstein, S. P., Englander, B. S., & Eckstein, M. P. (2021). Under-exploration of three-dimensional images leads to search errors for small salient targets. *Current Biology*, *31*(5), 1099–1106.e5. https://doi.org/10.1016/j.cub.2020.12.029

Lawsen, A. (2025). The illusion of the illusion of thinking [arXiv:2506.09250v2]. *arXiv*.

Lebreton, M., Langdon, S., Slieker, M. J., Nooitgedacht, J. S., Goudriaan, A. E., Denys, D., van Holst, R. J., & Luigjes, J. (2018). Two sides of the same coin: Monetary incentives concurrently improve and bias confidence judgments. *Science Advances*, *4*(5), eaaq0668. https://doi.org/10.1126/sciadv.aaq0668

Lee, K., & Ashton, M. C. (2008). The hexaco personality factors in the indigenous personality lexicons of english and 11 other languages. *Journal of Personality*, *76*(5), 1001–1054. https://doi.org/10.1111/j.1467-6494.2008.00512.x

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in "theory of mind". *Trends in Cognitive Sciences*, *8*(12), 528–533. https://doi.org/10.1016/j.tics.2004.10.001

Leung, E., Paolacci, G., & Puntoni, S. (2018). Man versus machine: Resisting automation in identity-based consumer behavior. *Journal of Marketing Research*, *55*(6), 818–831. https://doi.org/10.1177/0022243718818423

Li, Y., Huang, Y., Wang, H., Zhang, X., Zou, J., & Sun, L. (2024). Quantifying ai psychology: A psychometrics benchmark for large language models [arXiv:2406.17675]. *arXiv*. https://doi.org/10.48550/arXiv.2406.17675

Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, *33*(10), 2017–2031. https://doi.org/10.1162/jocn_a_01544

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88. https://doi.org/10.1016/j.media.2017.07.005

Liu, M., Okuhara, T., Chang, X., Shirabe, R., Nishiie, Y., Okada, H., & Kiuchi, T. (2024). Performance of chatgpt across different versions in medical licensing examinations

worldwide: Systematic review and meta-analysis. *Journal of Medical Internet Research*, *26*(1), e60807. https://doi.org/10.2196/60807

Loconte, R., Orrù, G., Tribastone, M., Pietrini, P., & Sartori, G. (2023). Challenging chatgpt' intelligence' with human tools: A neuropsychological investigation on prefrontal functioning of a large language model [SSRN Scholarly Paper No. 4377371]. *Social Science Research Network*. https://doi.org/10.2139/ssrn.4377371

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. https://doi.org/10.1016/j.obhdp.2018.12.005

Løhre, E., & Halkjelsvik, T. (2024). Advice taking when the stakes are high: Evidence from a game show. *Judgment and Decision Making*, *19*, e25. https://doi.org/10.1017/jdm.2024.4

Lu, Z., Zhang, L., Yao, L., Gong, D., Wu, L., Xia, M., Zhang, J., Zhou, W., Huang, X., He, C., Wu, H., Zhang, C., Li, X., & Yu, H. (2023). Assessment of the role of artificial intelligence in the association between time of day and colonoscopy quality. *JAMA Network Open*, *6*(1), e2253840. https://doi.org/10.1001/jamanetworkopen.2022.53840

Mahmoodi, A., Bang, D., Olsen, K., Zhao, Y. A., Shi, Z., Broberg, K., Safavi, S., Han, S., Nili Ahmadabadi, M., Frith, C. D., Roepstorff, A., Rees, G., & Bahrami, B. (2015). Equality bias impairs collective decision-making across cultures. *Proceedings of the National Academy of Sciences*, *112*(12), 3835–3840. https://doi.org/10.1073/pnas.1421692112

Mahmud, H., Islam, A. K. M. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? a systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, *175*, 121390. https://doi.org/10.1016/j.techfore.2021.121390

Manassi, M., & Whitney, D. (2018). Multi-level crowding and the paradox of object recognition in clutter. *Current Biology*, *28*(3), R127–R133. https://doi.org/10.1016/j.cub.2017.12.051

Maniscalco, B., & Lau, H. (2014). Signal detection theory analysis of type 1 and type 2 data: Meta-d, response-specific meta-d, and the unequal variance sdt model. In S. M. Fleming & C. D. Frith (Eds.), *The cognitive neuroscience of metacognition* (pp. 25–66). Springer. https://doi.org/10.1007/978-3-642-45190-4_3

Marblestone, A., Wayne, G., & Kording, K. (2016). Towards an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, *10*. https://doi.org/10.3389/fncom.2016.00094

Mazancieux, A., Fleming, S. M., Souchay, C., & Moulin, C. J. A. (2020). Is there a g factor for metacognition? correlations in retrospective metacognitive sensitivity across tasks. *Journal of Experimental Psychology: General*, *149*(9), 1788–1799. https://doi.org/10.1037/xge0000746

McCorduck, P. (2018). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*. CRC Press.

McIntosh, R. D., Fowler, E. A., Lyu, T., & Della Sala, S. (2019). Wise up: Clarifying the role of metacognition in the dunning-kruger effect. *Journal of Experimental Psychology: General*, *148*(11), 1882–1897. https://doi.org/10.1037/xge0000579

McWilliams, N. (2020). *Psychoanalytic diagnosis: Second edition: Understanding personality structure in the clinical process*. Guilford Press. https://www.guilford.com/books/Psychoanalytic-Diagnosis/Nancy-McWilliams/9781462543694

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press. https://doi.org/10.1037/11281-000

Miller, J. D., & Campbell, W. K. (2008). Comparing clinical and social-personality conceptualizations of narcissism. *Journal of Personality*, *76*(3), 449–476. https://doi.org/10.1111/j.1467-6494.2008.00492.x

Modi, H. N., Singh, H., Darzi, A., & Leff, D. R. (2020). Multitasking and time pressure in the operating room: Impact on surgeons' brain function. *Annals of Surgery*, *272*(4), 648–657. https://doi.org/10.1097/SLA.0000000000004208

Molleman, L., Tump, A. N., Gradassi, A., Herzog, S., Jayles, B., Kurvers, R. H. J. M., & van den Bos, W. (2020). Strategies for integrating disparate social information. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1939), 20202413. https://doi.org/10.1098/rspb.2020.2413

Mondillo, G., Colosimo, S., Perrotta, A., Frattolillo, V., & Masino, M. (2025). Comparative evaluation of advanced ai reasoning models in pediatric clinical decision support: Chatgpt o1 vs. deepseek-r1 [p. 2025.01.27.25321169]. *medRxiv*. https://doi.org/10.1101/2025.01.27.25321169

Moore, D. A., & Schatz, D. (2017). The three faces of overconfidence. *Social and Personality Psychology Compass*, *11*(8), 1–12. https://doi.org/10.1111/spc3.12331

Morewedge, C. K. (2022). Preference for human, not algorithm aversion. *Trends in Cognitive Sciences*, *26*(10), 824–826. https://doi.org/10.1016/j.tics.2022.07.007

Morin, O., Jacquet, P. O., Vaesen, K., & Acerbi, A. (2021). Social information use and social information waste. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1828), 20200052. https://doi.org/10.1098/rstb.2020.0052

Most, S. B., Simons, D. J., Scholl, B. J., Jimenez, R., Clifford, E., & Chabris, C. F. (2001). How not to be seen: The contribution of similarity and selective ignoring to sustained inattentional blindness. *Psychological Science*, *12*(1), 9–17. https://doi.org/10.1111/1467-9280.00303

Nelson, T. O. (1990). Metamemory: A theoretical framework and new findings (G. H. Bower, Ed.). *26*, 125–173. https://doi.org/10.1016/S0079-7421(08)60053-5

Norman, E., Pfuhl, G., Sæle, R. G., Svartdal, F., Låg, T., & Dahl, T. I. (2019). Metacognition in psychology. *Review of General Psychology*, *23*(4), 403–424. https://doi.org/10.1177/1089268019883821

Omar, M., Agbareia, R., Glicksberg, B. S., Nadkarni, G. N., & Klang, E. (2025). Benchmarking the confidence of large language models in answering clinical questions: Cross-sectional evaluation study. *JMIR Medical Informatics*, *13*, e66917. https://doi.org/10.2196/66917

O'Reilly, C., & Hall, N. (2021). Grandiose narcissists and decision making: Impulsive, overconfident, and skeptical of experts–but seldom in doubt. *Personality and Individual Differences*, *168*, 110280. https://doi.org/10.1016/j.paid.2020.110280

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and*

Cybernetics - Part A: Systems and Humans, *30*(3), 286–297. https://doi.org/10.1109/3468.844354

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), 102:1–102:15. https://doi.org/10.1145/3359204

Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin*, *116*(2), 220.

Paulhus, D., & Williams, K. (2002). The dark triad of personality: Narcissism, machiavellianism, and psychopathy. *Journal of Research in Personality*, *36*(6), 556–563. https://doi.org/10.1016/S0092-6566(02)00505-6

Paulhus, D. L., Harms, P. D., Bruce, M. N., & Lysy, D. C. (2003). The over-claiming technique: Measuring self-enhancement independent of ability. *Journal of Personality and Social Psychology*, *84*(4), 890–904. https://doi.org/10.1037/0022-3514.84.4.890

Pawitan, Y., & Holmes, C. (2025). Confidence in the reasoning of large language models. *Harvard Data Science Review*, *7*(1). https://doi.org/10.1162/99608f92.b033a087

Payne, B. K., & Bishara, A. J. (2009). An integrative review of process dissociation and related models in social cognition. *European Review of Social Psychology*, *20*(1), 272–314. https://doi.org/10.1080/10463280903162177

Payne, J. W., Samper, A., Bettman, J. R., & Luce, M. F. (2008). Boundary conditions on unconscious thought in complex decision making. *Psychological Science*, *19*(11), 1118–1123. https://doi.org/10.1111/j.1467-9280.2008.02212.x

Pedone, R., Semerari, A., Riccardi, I., Procacci, M., Nicolo, G., & Carcione, A. (2017). Development of a self-report measure of metacognition: The metacognition self-assessment scale (msas). *Clinical Neuropsychiatry*, *14*, 185–194.

Pendurkar, S., Huang, T., Koenig, S., & Sharon, G. (2022). The (un)scalability of heuristic approximators for np-hard search problems [arXiv:2209.03393]. *arXiv*. https://doi.org/10.48550/arXiv.2209.03393

Perussia, F., & Viano, R. (2008). *Mini locus of control scale: Piccolo manuale, con tratti e tipi da una scala psicometrica semplificata*. FrancoAngeli.

Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Zhang, C. B. C., Shaaban, M., Ling, J., Shi, S., Choi, M., Agrawal, A., Chopra, A., Khoja, A., Kim, R., Ren, R., Hausenloy, J., Zhang, O., Mazeika, M., & Hendrycks, D. (2025). Humanity's last exam [arXiv:2501.14249]. *arXiv*. https://doi.org/10.48550/arXiv.2501.14249

Pincus, A. L., Ansell, E. B., Pimentel, C. A., Cain, N. M., Wright, A. G. C., & Levy, K. N. (2009). Initial construction and validation of the pathological narcissism inventory. *Psychological Assessment*, *21*(3), 365–379. https://doi.org/10.1037/a0016530

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability [arXiv:1802.07810]. *arXiv*. https://doi.org/10.48550/arXiv.1802.07810

Prezza, M., Trombaccia, F. R., & Armento, L. (1997). La scala dell'autostima di rosenberg: Traduzione e validazione italiana [the rosenberg self-esteem scale: Italian translation and validation]. *Giunti Organizzazioni Speciali*, *223*, 35–44.

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The development and testing of a new version of the cognitive reflection test applying item response

theory (irt). *Journal of Behavioral Decision Making*, *29*(5), 453–469. https://doi.org/10.1002/bdm.1883

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.

Raees, M., Meijerink, I., Lykourentzou, I., Khan, V.-J., & Papangelis, K. (2024). From explainable to interactive ai: A literature review on current trends in human-ai interaction [arXiv:2405.15051]. *arXiv*. https://doi.org/10.48550/arXiv.2405.15051

Rahnev, D. (2023). Measuring metacognition: A comprehensive assessment of current methods. https://doi.org/10.31234/osf.io/waz9h

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336), 846–850. https://doi.org/10.1080/01621459.1971.10482356

Raskin, R. N., & Hall, C. S. (1979). A narcissistic personality inventory. *Psychological Reports*, *45*(2), 590. https://doi.org/10.2466/pr0.1979.45.2.590

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), 83:1–83:22. https://doi.org/10.1145/3512930

Raven, J. C., & Court, J. H. (1998). *Raven's progressive matrices and vocabulary scales*. Oxford Psychologists Press.

Reverberi, C., Rigon, T., Solari, A., Hassan, C., Cherubini, P., GI Genius CADx Study Group, & Cherubini, A. (2022). Experimental evidence of effective human-ai collaboration in medical decision-making. *Scientific Reports*, *12*(1), 14952. https://doi.org/10.1038/s41598-022-18751-2

Rex, D. K., Mori, Y., Sharma, P., Lahr, R. E., Vemulapalli, K. C., & Hassan, C. (2022). Strengths and weaknesses of an artificial intelligence polyp detection program as assessed by a high-detecting endoscopist. *Gastroenterology*, *163*(2), 354–358.e1. https://doi.org/10.1053/j.gastro.2022.03.055

Riva, P., Aureli, N., & Silvestrini, F. (2022). Social influences in the digital era: When do people conform more to a human being or an artificial intelligence? *Acta Psychologica*, *229*, 103681. https://doi.org/10.1016/j.actpsy.2022.103681

Romano, D., Costantini, G., Richetin, J., & Perugini, M. (2023). The hexaco adjective scales and its psychometric properties. *Assessment*, *30*(8), 2510–2532. https://doi.org/10.1177/10731911231153833

Rondonotti, E., Hassan, C., Tamanini, G., Antonelli, G., Andrisani, G., Leonetti, G., Paggi, S., Amato, A., Scardino, G., Paolo, D. D., Mandelli, G., Lenoci, N., Terreni, N., Andrealli, A., Maselli, R., Spadaccini, M., Galtieri, P. A., Correale, L., Repici, A., & Radaelli, F. (2022). Artificial intelligence-assisted optical diagnosis for the resect-and-discard strategy in clinical practice: The artificial intelligence bli characterization (abc) study. *Endoscopy*, 14–22. https://doi.org/10.1055/a-1852-0330

Rosenberg, M. (1979). *Conceiving the self*. Basic Books. http://archive.org/details/conceivingself00rose

Rossettini, G., Rodeghiero, L., Corradi, F., Cook, C., Pillastrini, P., Turolla, A., Castellini, G., Chiappinotto, S., Gianola, S., & Palese, A. (2024). Comparative accuracy of chatgpt-4, microsoft copilot and google gemini in the italian entrance test for healthcare

sciences degrees: A cross-sectional study. *BMC Medical Education*, *24*(1), 694. https://doi.org/10.1186/s12909-024-05630-9

Rotter, J. B. (1954). *Social learning and clinical psychology*. Prentice-Hall, Inc. https://doi.org/10.1037/10788-000

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, *80*(1), 1–28. https://doi.org/10.1037/h0092976

Rumelhart, D. E., McClelland, J. L., & Group, P. R. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: Foundations*. The MIT Press. https://doi.org/10.7551/mitpress/5236.001.0001

Sartori, G., & Orrù, G. (2023). Language models and psychological sciences. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1279317

Scaioli, G., Lo Moro, G., Conrado, F., Rosset, L., Bert, F., & Siliquini, R. (2023). Exploring the potential of chatgpt for clinical reasoning and decision-making: A cross-sectional study on the italian medical residency exam. *Annali Dell'Istituto Superiore Di Sanita*, *59*(4), 267–270. https://doi.org/10.4415/ANN_23_04_05

Schaefer, P. S., Williams, C., Goodie, A., & Campbell, K. (2004). Overconfidence and the big five. *Journal of Research in Personality*, *38*(5), 473–480.

Schepman, A., & Rodway, P. (2022). The general attitudes towards artificial intelligence scale (gaais): Confirmatory validation and associations with personality, corporate distrust, and general trust. *International Journal of Human–Computer Interaction*, *0*(0), 1–18. https://doi.org/10.1080/10447318.2022.2085400

Schoenleber, M., Roche, M. J., Wetzel, E., Pincus, A. L., & Roberts, B. W. (2015). Development of a brief version of the pathological narcissism inventory. *Psychological Assessment*, *27*(4), 1520. https://doi.org/10.1037/pas0000158

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T., & Silver, D. (2020). Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, *588*(7839), 604–609. https://doi.org/10.1038/s41586-020-03051-4

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Shah, S., Park, N., Chehade, N. E. H., Chahine, A., Monachese, M., Tiritilli, A., Moosvi, Z., Ortizo, R., & Samarasena, J. (2023). Effect of computer-aided colonoscopy on adenoma miss rates and polyp detection: A systematic review and meta-analysis. *Journal of Gastroenterology and Hepatology*, *38*(2), 162–176. https://doi.org/10.1111/jgh.16059

Shekhar, M., & Rahnev, D. (2021a). The nature of metacognitive inefficiency in perceptual decision making. *Psychological Review*, *128*, 45–70. https://doi.org/10.1037/rev0000249

Shekhar, M., & Rahnev, D. (2021b). Sources of metacognitive inefficiency. *Trends in Cognitive Sciences*, *25*(1), 12–23. https://doi.org/10.1016/j.tics.2020.10.007

Shekhar, M., & Rahnev, D. (2024). How do humans give confidence? a comprehensive comparison of process models of perceptual metacognition. *Journal of Experimental Psychology: General*, *153*(3), 656–688. https://doi.org/10.1037/xge0001524

Sheppard, L. D., & Vernon, P. A. (2008). Intelligence and speed of information-processing: A review of 50 years of research. *Personality and Individual Differences*, *44*(3), 535–551. https://doi.org/10.1016/j.paid.2007.09.015

Sherbino, J., & Norman, G. (2021). Task switching, multitasking, and errors: A psychologic perspective on the impact of interruptions. *Annals of Emergency Medicine*, *78*(3), 425–428. https://doi.org/10.1016/j.annemergmed.2021.07.120

Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., & Farajtabar, M. (2025). The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity [arXiv:2501.12948]. *arXiv*.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. https://doi.org/10.1038/nature16961

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, *362*(6419), 1140–1144. https://doi.org/10.1126/science.aar6404

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354–359. https://doi.org/10.1038/nature24270

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, *28*(9), 1059–1074. https://doi.org/10.1068/p281059

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, *620*(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2

Sniezek, J. A., & Buckley, T. (1995). Cueing and cognitive conflict in judge-advisor decision making. *Organizational Behavior and Human Decision Processes*, *62*(2), 159–174. https://doi.org/10.1006/obhd.1995.1040

Sniezek, J. A., Schrah, G. E., & Dalal, R. S. (2004). Improving judgment with prepaid expert advice. *Journal of Behavioral Decision Making*, *17*(3), 173–190. https://doi.org/10.1002/bdm.468

Srinivasan, N., & Tikoo, S. (1992). Effect of locus of control on information search behavior. *Advances in Consumer Research*, *19*(1), 498–504.

Stanovich, K. E., & Toplak, M. E. (2023). Actively open-minded thinking and its measurement. *Journal of Intelligence*, *11*(2), 27. https://doi.org/10.3390/jintelligence11020027

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology*, *89*(2), 342–357. https://doi.org/10.1037/0022-0663.89.2.342

Stein, J.-P., Messingschlager, T., Gnambs, T., Hutmacher, F., & Appel, M. (2024). Attitudes towards ai: Measurement and associations with personality. *Scientific Reports*, *14*(1), 2909. https://doi.org/10.1038/s41598-024-53335-2

Steyvers, M., & Kumar, A. (2024). Three challenges for ai-assisted decision-making. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *19*(5), 722–734. https://doi.org/10.1177/17456916231181102

Stöcker, A.-K., & Schütz, A. (2024). Don't tell me what to do! narcissism and advice taking: A meta-analysis and future research directions. *Personality and Individual Differences*, *223*, 112607. https://doi.org/10.1016/j.paid.2024.112607

Team, T. L., Barreiros, J., Beaulieu, A., Bhat, A., Cory, R., Cousineau, E., Dai, H., Fang, C.-H., Hashimoto, K., Irshad, M. Z., Itkina, M., Kuppuswamy, N., Lee, K.-H., Liu, K., McConachie, D., McMahon, I., Nishimura, H., Phillips-Grafflin, C., Richter, C., & Tedrake, R. (2025). A careful examination of large behavior models for multitask dexterous manipulation [arXiv:2507.05331]. *arXiv*. https://doi.org/10.48550/arXiv.2507.05331

Tejeda, H., Kumar, A., Smyth, P., & Steyvers, M. (2022). Ai-assisted decision-making: A cognitive modeling approach to infer latent reliance strategies. *Computational Brain Behavior*, *5*(4), 491–508. https://doi.org/10.1007/s42113-022-00157-y

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory Cognition*, *39*(7), 1275–1289. https://doi.org/10.3758/s13421-011-0104-1

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking Reasoning*, *20*(2), 147–168.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136. https://doi.org/10.1016/0010-0285(80)90005-5

Trinh, T. H., Wu, Y., Le, Q. V., He, H., & Luong, T. (2024). Solving olympiad geometry without human demonstrations. *Nature*, *625*(7995), 476–482. https://doi.org/10.1038/s41586-023-06747-5

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and ai are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, *8*(12), 2293–2303. https://doi.org/10.1038/s41562-024-02024-1

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need [arXiv:1706.03762]. *arXiv*. https://doi.org/10.48550/arXiv.1706.03762

Vélez, N., & Gweon, H. (2019). Integrating incomplete information with imperfect advice. *Topics in Cognitive Science*, *11*(2), 299–315. https://doi.org/10.1111/tops.12388

Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, *11*, 2837–2854.

Wallace, M. B., Sharma, P., Bhandari, P., East, J., Antonelli, G., Lorenzetti, R., Vieth, M., Speranza, I., Spadaccini, M., Desai, M., Lukens, F. J., Babameto, G., Batista, D., Singh, D., Palmer, W., Ramirez, F., Palmer, R., Lunsford, T., Ruff, K., & Hassan, C. (2022). Impact of artificial intelligence on miss rate of colorectal neoplasia. *Gastroenterology*, *163*(1), 295–304.e5. https://doi.org/10.1053/j.gastro.2022.03.007

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2020). Superglue: A stickier benchmark for general-purpose language understanding systems [arXiv:1905.00537]. *arXiv*. https://doi.org/10.48550/arXiv.1905.00537

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding, 353–355. https://doi.org/10.18653/v1/W18-5446

Wang, G., Li, J., Sun, Y., Chen, X., Liu, C., Wu, Y., Lu, M., Song, S., & Yadkori, Y. A. (2025). Hierarchical reasoning model [arXiv:2506.21734]. *arXiv*. https://doi.org/10.48550/arXiv.2506.21734

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models [arXiv:2206.07682]. *arXiv*. https://doi.org/10.48550/arXiv.2206.07682

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models [arXiv:2201.11903]. *arXiv*. https://doi.org/10.48550/arXiv.2201.11903

Wickens, C. D., Clegg, B. A., Vieane, A. Z., & Sebok, A. L. (2015). Complacency and automation bias in the use of imperfect automation. *Human Factors*, *57*(5), 728–739. https://doi.org/10.1177/0018720815581940

Wijk, H., Lin, T., Becker, J., Jawhar, S., Parikh, N., Broadley, T., Chan, L., Chen, M., Clymer, J., Dhyani, J., Ericheva, E., Garcia, K., Goodrich, B., Jurkovic, N., Karnofsky, H., Kinniment, M., Lajko, A., Nix, S., Sato, L., & Barnes, E. (2025). Re-bench: Evaluating frontier ai rd capabilities of language model agents against human experts [arXiv:2411.15114]. *arXiv*. https://doi.org/10.48550/arXiv.2411.15114

Williams, L., Carrigan, A., Auffermann, W., Mills, M., Rich, A., Elmore, J., & Drew, T. (2021). The invisible breast cancer: Experience does not protect against inattentional blindness to clinically relevant findings in radiology. *Psychonomic Bulletin Review*, *28*(2), 503–511. https://doi.org/10.3758/s13423-020-01826-4

Williams, L. H., & Drew, T. (2019). What do we know about volumetric medical image interpretation?: A review of the basic science and medical image perception literatures. *Cognitive Research: Principles and Implications*, *4*(1), 21. https://doi.org/10.1186/s41235-019-0171-6

Wolfe, J. M. (2001). Asymmetries in visual search: An introduction. *Perception Psychophysics*, *63*(3), 381–389. https://doi.org/10.3758/bf03194406

Wolfe, J. M. (2021). Guided search 6.0: An updated model of visual search. *Psychonomic Bulletin Review*, *28*(4), 1060–1092. https://doi.org/10.3758/s13423-020-01859-9

Wolfe, J. M., Kosovicheva, A., & Wolfe, B. (2022). Normal blindness: When we look but fail to see. *Trends in Cognitive Sciences*, *26*(9), 809–819. https://doi.org/10.1016/j.tics.2022.06.006

Wood, T. J. (2014). Exploring the role of first impressions in rater-based assessments. *Advances in Health Sciences Education*, *19*(3), 409–427. https://doi.org/10.1007/s10459-013-9453-9

Wright, A. G. C., Lukowitsky, M. R., Pincus, A. L., & Conroy, D. E. (2010). The higher order factor structure and gender invariance of the pathological narcissism inventory. *Assessment*, *17*(4), 467–483. https://doi.org/10.1177/1073191110373227

Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.-C., Liu, Z., & Wang, L. (2023). The dawn of lmms: Preliminary explorations with gpt-4v(ision) [arXiv:2309.17421]. *arXiv*. https://doi.org/10.48550/arXiv.2309.17421

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, *93*(1), 1–13. https://doi.org/10.1016/j.obhdp.2003.08.002

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. https://doi.org/10.1006/obhd.2000.2909

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models [arXiv:2305.10601]. *arXiv*. https://doi.org/10.48550/arXiv.2305.10601

Yax, N., Anlló, H., & Palminteri, S. (2024). Studying and improving reasoning in humans and machines. *Communications Psychology*, *2*(1), 1–16. https://doi.org/10.1038/s44271-024-00091-8

Yi, S., Liu, Y., Sun, Z., Cong, T., He, X., Song, J., Xu, K., & Li, Q. (2024). Jailbreak attacks and defenses against large language models: A survey [arXiv:2407.04295]. *arXiv*. https://doi.org/10.48550/arXiv.2407.04295

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making, 295–305. https://doi.org/10.1145/3351095.3372852

Zonca, J., Giampino, A., Cherubini, P., & Reverberi, C. (2025). The advice gap: Adaptive yet suboptimal advice integration. *OSF*. https://doi.org/10.31234/osf.io/4qkje_v1