The Impact of Expertise Dynamics on Human-Agent Collaboration

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Fritz Becker, M.Sc.

aus Münster

Tübingen

2024

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	14.07.2025
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	Prof. Dr. Markus Huff
2. Berichterstatter/-in:	Prof. Dr. Martin Butz

Wenn man jemand nicht versteht, so hält man ihn in der Regel für dumm. - Carl Jung (1996, p. 152)

Acknowledgments

I would like to express my gratitude to all those who supported me throughout my academic journey.

First, I have to thank my partner, Miriam. I cannot say if her emotional support and motivation or her direct support in structuring my thesis were the most significant help. The thesis would not be as it is without her.

I want to thank my supervisors, Prof. Dr. Markus Huff and PD Dr. Jürgen Buder for the balance between independence and guidance during my thesis. Special thanks to Jürgen Buder for the discussions whenever a new idea popped into my head.

Further, I want to thank my colleagues. A thesis is always a struggle, but it is more entertaining if we struggle together. Special thanks go to Dr. Nadia Said who uniquely helped me with her insight into academic careers which we discussed over coffee.

Finally, I want to thank my parents, Jutta Becker and Martin Schlüter, who patiently supported me throughout my studies.

Table of Contents

Figures

Tables

Summary

Navigating social environments demands an understanding of other agents' mental states. Understanding an agent's decision-making process helps an individual human to anticipate how its actions impact the shared environment, its proficiency at adapting the environment to its purposes, and the alignment of its intentions with their own. Arguably, anticipating agents' decisions is becoming more important due to the proliferation of autonomous systems and the growing human population. However, humans cannot perceive other agents' mental processes directly. They must either infer the process by observing the agent's behavior or relying on their experience with similar agents or relying on the testimony of others. Judging an agent by experience with superficially similar agents might lead to prejudice and the miscalibration of trust. However, applying knowledge from experience or the agent's reputation can help individual humans navigate social environments with a sense of security. It is more challenging to apply such information to newer types of agents such as chatbots, autonomous cars, or recommender systems as many human judges have little first- or second-hand experience with them.

In this dissertation, I investigate how humans build a mental representation of an artificial agent's decision process to judge its expertise and trustworthiness. In the first study, I explain how humans can anticipate the actions of an agent by using their own task expertise as a source of information about the decisions a rationally acting agent makes. In the second study, I explore the role of an observer's expertise in judging an agent's expertise. Given only information about the agent's decisions and lacking information about the agent, human judges are unable to accurately assess the agent's expertise if they know less about the task than the agent does. In the final study, the alleged and factual expertise of the agent is manipulated, and I examine how trust in the agent develops during an interaction. Before interacting with the agent, the human's trust in it depends on the agent's reputation, but their first-hand experience of the agent overwrites that expectation during the interaction.

A central point of discussion is the applicability of research on interactions between humans to interactions between humans with artificial agents and the other way around, the generalizability of the research on human interactions with artificial agents to human interactions with any agent (e.g., humans, animals, or technology). Additionally, I discuss the relevance of information about the agents that goes beyond their pure actions on an observer's ability to accurately perceive the agent's expertise. More and varying information on the agent will likely moderate the effects of observer expertise; exploring this hypothesis could be an interesting research question that builds on the results of this dissertation.

Zusammenfassung

Um sich in sozialen Umgebungen zurechtzufinden, muss man die mentalen Zustände anderer Agenten verstehen. Das Verständnis des Entscheidungsprozesses eines Agenten hilft einem einzelnen Menschen, vorherzusehen, wie sich die Handlungen des Agenten auf die gemeinsame Umgebung auswirken, wie gut der Agent die Umgebung an seine Zwecke anpassen kann und wie die Absichten des Agenten mit denen der beobachtenden Person übereinstimmen. Aufgrund der Verbreitung autonomer Systeme und der wachsenden menschlichen Bevölkerung wird es immer wichtiger, die Entscheidungen anderer Agenten zu antizipieren. Menschen können die mentalen Prozesse anderer Agenten jedoch nicht direkt wahrnehmen. Sie müssen den Prozess entweder durch Beobachtung des Verhaltens eines Agenten ableiten, sich auf Erfahrungen mit ähnlichen Agenten oder auf die Aussagen anderer verlassen. Die Beurteilung eines Agenten anhand von Erfahrungen mit oberflächlich ähnlichen Agenten kann zu Vorurteilen und einer Fehlanpassung des Vertrauens führen. Die Anwendung von Erfahrungen mit ähnlichen Agenten oder das Einbeziehen des Rufs des Agenten kann einzelnen Menschen jedoch dabei helfen, sich in sozialen Umgebungen mit einem Gefühl der Sicherheit zurechtzufinden. Es ist schwieriger, solche Informationen auf neuere Sorten von Agenten wie Chatbots, autonome Autos oder Empfehlungssysteme anzuwenden, da viele menschliche Beurteilende wenig Erfahrung aus erster oder zweiter Hand mit ihnen haben.

In dieser Dissertation untersuche ich, wie Menschen eine mentale Repräsentation des Entscheidungsprozesses eines künstlichen Agenten aufbauen, um dessen Aufgabenwissen und Vertrauenswürdigkeit zu beurteilen. In der ersten Studie zeige ich, wie Menschen die Handlungen eines Agenten vorhersehen können, indem sie ihr eigenes Aufgabenwissen als Informationsquelle über die Entscheidungen eines rational handelnden Agenten nutzen. In der zweiten Studie untersuche ich die Rolle des Aufgabenwissens einer beurteilenden Person bei der Einschätzung des Aufgabenwissens eines Agenten. Da menschliche Bewertende nur Informationen über die Entscheidungen des Agenten und keine Informationen über den Agenten selbst haben, können sie das Aufgabenwissen des Agenten nicht genau beurteilen, wenn sie weniger über die Aufgabe wissen als der Agent. In der letzten Studie werden das angebliche und tatsächliche Aufgabenwissen des Agenten manipuliert und ich untersuche, wie sich das Vertrauen in den Agenten während einer Interaktion entwickelt. Vor der Interaktion mit dem Agenten hängt das Vertrauen des Menschen in den Agenten vom Ruf des Agenten ab, aber seine Erfahrungen aus erster Hand mit dem Agenten überschreiben diese Erwartung im Laufe der Interaktion.

Ein zentraler Diskussionspunkt ist die Anwendbarkeit der Forschung über Interaktionen zwischen Menschen auf Interaktionen von Menschen mit künstlichen Agenten und umgekehrt die Generalisierbarkeit der Forschung zu menschlicher Interaktionen mit künstlichen Agenten auf menschliche Interaktionen mit beliebigen Agenten (z. B. Menschen, Tieren oder Technologie). Darüber hinaus diskutiere ich die Relevanz von Informationen über die Agenten, die über ihre reinen Handlungen hinausgehen, für die Fähigkeit einer beobachtenden Person, die Expertise des Agenten richtig wahrzunehmen. Mehr und variierende Informationen über den Agenten werden wahrscheinlich die Auswirkungen der Expertise des Beobachters abmildern; die Untersuchung dieser Hypothese könnte eine interessante Forschungsfrage sein, die auf den Ergebnissen dieser Dissertation aufbaut.

Chapter 1 - Introduction

Human population density is increasing (Goldewijk, 2005; Lewis & Maslin, 2015), and artificial systems are becoming more and more agentic (Sundar, 2020). Never before was our environment influenced as much by other agents (Al Shamsi et al., 2022; Kaur et al., 2023). To anticipate developments in our environment, understanding these agents is more critical than ever (Thornton et al., 2019).

Humans have a long evolutionary history of performing tasks such as navigating an environment with other agents. Since *Homo sapiens* emerged, we have worked and moved in groups, adjusting our senses and cognitions to the actions of others (Warren, 2018). Furthermore, each human grows up in a specific environment, learns about it, and adapts to it (Bronfenbrenner, 1977). When the modes of transport were enriched by horses and cars, societies and individual humans adapted to the change. They learned that a horse has agency but can be somewhat controlled by humans, and cars can move at speeds that test the limits of human reaction times. Society adopted these technologies, and individuals learned to incorporate them into their mental representations of the environment. A new shift is underway: autonomous artificial agents are entering our environment. Autonomous cars can behave differently than human drivers and endanger individuals who have not yet learned how to predict their actions. Humans must also adapt to the new artificial agents in their environment in areas such as image generation, writing, and consumption. Misunderstanding these agents and miscalibrating their trust in them leads to misuse.

Understanding an agent means knowing the agent's beliefs about its environment, goals or desires, abilities, and perceptions (Hellström & Bensch, 2018). In this dissertation, I present research on how humans use their understanding of an environment to represent an artificial agent mentally. After this introductory chapter, I present research that addresses the question of how people represent the decision-making process of an agent. I debate whether knowledge of the agent or knowledge of the agent's environment is more relevant to anticipating

the agent's decisions and present two experiments that answer the question. In Chapter 3, I transfer the results of Chapter 2 to show how the agent's perceived expertise depends on the observer's expertise. I present two conceptually similar studies, one well-controlled experiment, and an ecologically more valid quasi-experiment. In the last empirical chapter, I present data on how the mental representation of the agent's abilities shapes the observers' trust in it. Before having first-hand experiences with the agent, the observer's trust is dependent on the reports of others. However, first-hand experience with agents who perform well and those who perform badly leads to a quick recalibration of the trusting behavior. In the final chapter, I discuss the results that were collected. I pay special attention to the possibility of generalizing the results the results from strictly artificial agents to other agentic entities such as humans and to the fields in which the results could be applied.

Agency

This dissertation's primary goal is to understand the mental processes involved in the perception and mental representation of agents and their abilities. However, agents and agency have branching research histories. Computer science and social psychology differ in their definitions of agency. Although the motivations for using the terms "agent" and "agency" and how they are used are different in the two disciplines, the are conceptually similar. In this section, I present how I derive my understanding of the term "agent" from both traditions.

In computer science, an agent can be defined as a system that can perceive its environment and act on it to achieve its goals (Franklin & Graesser, 1997; Maes, 1993). An agent and its environment are two tightly coupled dynamical systems (Beer, 1995). An agent's environment changes in reaction to the agent's actions and independently (Barandiaran et al., 2009). The agent perceives the environment and uses its actions to manipulate it into an more favorable state. The definition of an agent applies to many kinds of systems. In the field, it is usually used to reference a computer system. However, it also applies to humans or animals and, in extreme cases, even to a thermostat or bacterium (Franklin & Graesser, 1997).

In social psychology, agency and communion (Bakan, 1966) are the two fundamental dimensions into which humans categorize other entities, such as other humans, groups, artificial agents, and animals (Abele et al., 2016; Imhoff & Koch, 2017). In this context, agency refers to the qualities of an agent that help it reach its goals, while communion refers to the qualities needed to establish and maintain social relationships (Abele et al., 2016). Following this definition, “agent” describes an entity with the qualities needed to obtain its goals but provides no information about its social skills.

For the aims of this dissertation, a combination of the definitions used in computer science and the social sciences is relevant. Accordingly, an agent is an entity capable of perceiving and influencing its environment that observers can identify as agentic. In neither definition is a specific form of the entity necessary. Agents are usually humans or artificial systems but, depending on the context, a pet, a nation, or even a large human-agent team such as a company can be seen as a single agent.

Only in very constructed scenarios is an agent alone in its environment. A human’s environment usually is shared with other agents. We humans are aware of our intentions and can estimate the future state of the environment based on our decisions. However, the intentions of other agents are hidden from us (Apperly, 2011). Given the importance of other agents for an individual’s well-being, humans have developed mechanisms to perceive the intentions of others and anticipate the consequences of their actions. These mechanisms are researched in the field of social cognition.

Social Cognition

In this section, I will present the concepts of social cognition that are the most relevant for this dissertation on social perception. I start by giving a broad overview of the field of social cognition. Then I discuss the divide of social theories into theory theories and simulation theories. These high-level theories represent two differing views on how mental images of other agents are created. Important concepts for this dissertation are grounding and theory of mind, which I explain

later in this section. Following the explanation of the relevant concepts of social cognition, I discuss their applicability to artificial agentst.

Social cognition is an agent's mental representation of other agents in its environment. In environments where agents are not alone, an agent can observe the behavior of other agents to develop mental representations of their cognitions. Understanding agents and oneself in the context of other agents is the subject of the large field of social cognition (Moskowitz, 2005). This discipline deals with the accuracy and the processes of how people perceive others (Higgins & Bargh, 1987). The field is unique in that it gives an interactive perspective to traditional cognitive science in which the mental processes of agents are researched in isolation, mostly ignoring interactions with other agents. Additionally, it enriches social psychology by describing human behavior that involves hidden mental processes (Moscovici, 1988).

Scientists and laypersons observing an agent's behavior are faced with similar problems when they try to explain an agent's mental processes: The cognitions of others are not directly observable (Premack & Woodruff, 1978). An observer needs to overcome this limitation to create accurate mental representations of an observed agent. Having accurate representations of agents helps observers make predictions about their environment and react faster and more appropriate to changes.

An important current discussion in cognition deals with what aspects of cognition that use abstract, symbolic representations and what aspects use mental simulations of the environment. These concepts are called amodal and modal cognition (Kaup et al., 2024). Grounded cognition theory postulates that cognition is based on modal representations of the environment. Thus, it opposes the common stance that knowledge is an abstract system of amodal symbols on which cognitive computations are conducted that exist independently of the environment and senses (Barsalou, 2008).

Theory Theory and Simulation Theory

In the field of social perception, the discussion about modal and amodal processes is continued. *Theory theory* and *simulation theory* are philosophical and conceptual theories of how people represent others mentally. Later in this section, I will present some examples of applied theories that can be categorized as belonging to one of these two encompassing theories.

The amodal theory theory postulates that people have implicit theories about cognitive processes in general and specific cognitions of a person (Moskowitz, 2005). The main claim of the theory is that humans form theories to explain the behavior of others. The process is similar to the scientific process in the sense that people form theories from observations they make. The theories are then revised should their predictions not hold up in reality.

In contrast, the main claim in simulation theories is that people perceive others by adopting the perspectives of others (Barsalou, 2008). They imagine what they would perceive and decide if they were a different person with different experiences (Ruby & Decety, 2001).

The distinction between simulation theory and theory theory seems very artificial. The two mechanisms are not mutually exclusive. Both can be useful ways of processing social information and have been successfully tested in experiments (Apperly, 2008). Certain situations may permit the observer to use general implicit theories about cognition (theory theory), while other situations invite the use of a more specific prototype of how an agent solves a cognitive task (simulation theory).

As theory theory and simulation theory are fundamental and theoretical, applied theories can be categorized as belonging more to one or the other. In the following subsections, I will discuss the process of grounding between two agents and the theory of mind theory. The grounding process is an example of a concept traditionally categorized more into the simulation theories (Barsalou, 2008), while theory of mind is traditionally read as a theory theory, as the name implies (Apperly, 2008). Both theories can explain how other agents' mental states can be approximated by an observer.

Grounding

Grounded cognition describes a model of cognition that describes how an agent perceives and mentally represents their environment. The process of *grounding* refers to the formation of a shared mental representation of an environment between two agents (Barsalou, 2008; Horton & Keysar, 1996). Since two agents might perceive different subsets of an environment, they must form a common ground to communicate successfully (Nohara-LeClair, 2001). In grounded communication, the speaker communicates in a way that takes the knowledge and perspective of the addressee into account (Horton & Keysar, 1996). The more two agents share the same beliefs about an environment, the easier it is for them to find common ground and communicate successfully. The degree of common ground in communication is as important as the agents' linguistic competence (Bishop & Adams, 1991).

Experimentally, grounding is researched using the *referential communication task* (Horton & Keysar, 1996; Krauss & Glucksberg, 1977). In this task, a director is asked to indicate a figure on a grid of similar figures to a matcher. The challenge for the director is to efficiently describe the target figure and prevent the matcher from confusing it with similar-looking figures. Over time, the efficiency of communication and the accuracy the matching improve (Nohara-LeClair, 2001).

Common ground could naturally exist between experts performing a task. Given the bounds of a task, they may make the same decisions independently, even without prior grounding. It follows that experts might have a reasonable idea of the decision processes of other experts without effortfully mentalizing others' decision processes. Theory of mind is the process of mentalizing the mental states of others to anticipate and explain their observable behavior (Frith & Frith, 2012). As opposed to grounding, which is concerned with successful communication through aligning the mental models of the shared environment (Van der Velde, 2015), theory of mind refers to an observer's ability to mentalize the mental model of an agent.

Theory of Mind

The term *theory of mind* was originally coined by Premack and Woodruff (1978) to describe the ability of chimpanzees to impute mental states to themselves and others. They called it a theory of mind since the mental states are not directly observable, so an observer must form and test theories about the agent's mental states to accurately predict the agent's behavior. As such, theory of mind was originally a theory theory (Moskowitz, 2005; see also the introduction of the Section "Social Cognition"). However, today, it has become so broadly applied that it is also discussed in terms of simulation theory (Andrews, 2008; Apperly, 2008).

A *mental state* refers to all hidden experiences of an agent that dictate their behavior. The experiences that are included vary among different definitions of *mental state*, but goals and beliefs are always included in some way (Baker et al., 2009; Premack & Woodruff, 1978; Wellman et al., 2001). Theory of mind has been intensively researched in developmental psychology. It describes the developmental stage at which children can discriminate between their own and other's minds (Baron-Cohen et al., 1985; Wellman et al., 2001). Since then, it has developed to be almost synonymous with social perception in general. Modern research on theory of mind is concerned with how observers can represent the mental states of agents. Behavioral experiments explore the ability of observers to infer the mental states of agents by observing their behavior (Baker et al., 2009).

The most used experiment in theory of mind research is the false-belief task (for a review, see Wellman et al., 2001). In one such task, the observer sees an agent using an object. The agent places the object in one of two containers and leaves the scene. In the absence of the agent but, crucially, in the presence of the observer, the object is transferred to another container. When the agent returns, the observer is asked to point out the container in which the agent will search for the object. The observer has to understand that the agent holds a counterfactual and different belief than they do to answer correctly. Young children fail this task, picking the container that they know the object is in (Wellman et al., 2001). Starting from the age of 4-5 years, healthy

children generally pass the test. But, even adults can fail if they are distracted from the task (Newton & de Villiers, 2007). The false-belief task is criticized as being unspecific to theory of mind. False-belief tasks may require skills that go beyond theory of mind, and theory of mind might not necessarily lead to the ability to reason about false beliefs (Bloom & German, 2000; Schlinger, 2009).

Given the dependence of theory of mind research on false-belief tasks and the criticism of the task, new experiments have been developed to research the representation of other's mental states. Some of these experiments use an artificial agent that navigates an artificial environment. Based on the agent's path and pauses, the participant can make inferences about the agent's goals, decision-making, and knowledge (Baker et al., 2009, 2017; Berke & Jara-Ettinger, 2021).

In studies of theory of mind, the observer usually has full information about the agent's task, while the agent has varying, usually less, information about the task. The observer is asked to infer the agent's mental state using information about the task and information about the agent's behavior. The results of these experiments show that healthy adults are generally able to infer the mental states of the presented agent with a high degree of accuracy (Aboody et al., 2021). While usually the agent's behavior is experimentally manipulated and thus the source of variation, the observer's task knowledge is never manipulated in theory-of-mind scenarios. The likely explanation is that researchers implicitly agree that the observer needs to be fully informed about the task to make inferences about the agent's mental states. However, it is worth investigating explicitly the consequences of an observer's lack of knowledge of how they perceive an agent.

How an agent's decision process is mentally represented by human observers is still unclear (Schaafsma et al., 2015). One possibility is that the observer observes all the agent's decisions and infers a complex mental model of the latter's decision process. However, such a process would require considerable data and much effort from the observer.

It would be more efficient for the observer to imagine a rational agent and use the information they have on the shared environment to explain a specific observed agent's behavior. Current computational models call this process inverse planning (Baker et al., 2009; Berke & Jara-Ettinger, 2021). In this theorized process, the observer assumes that the agent plans their actions rationally using the information they have about the environment. Through such a process, the observer can draw conclusions about the agent's knowledge of the environment by looking at the actions that do not lead optimally to the agent's goal. Rather than encoding each observed decision, the observer can gain a broad understanding of the agent's knowledge. By using this high-level information about the agent, the observer can understand the agent's decisions.

Artificial Agents

Theories of social cognition are generally applied to interactions between humans. In some cases, interactions between humans and animals or machines are excluded from the subject of social cognition (Flavell & Miller, 1998). However, artificial systems continue to become more autonomous, agentic, and difficult to differentiate from humans. Some systems even show emergent evidence of having a theory of mind themselves (Kosinski, 2024). Why should a person not apply their social cognitive apparatus to non-human agents as well? When interacting with computers, people apply social norms to their behavior (Nass et al., 1994). The experiment was initially conducted when personal computers were rarer, so it is impossible to say whether the participants perceived the computer as a social actor or instead defaulted to known schemas due to the novelty of the interaction.

In social cognitive experiments, it is common to use artificial agents as material to research the cognitions of people who interact with the agents under controlled conditions. For example, the aforementioned false-belief task is usually conducted using pictures and a story (Scott & Baillargeon, 2017; Van der Wel et al., 2014). The agents are illustrated in various ways—for example, depicted in sketches or represented by toy figures or images of humans (Huemer et al., 2023; Scott & Baillargeon, 2017). In path-taking experiments, the agents are usually depicted

by an avatar in an artificial environment (Baker et al., 2009; Berke & Jara-Ettinger, 2021). While there might be differences between how humans represent other humans as opposed to artificial agents or protagonists of a story, the research practice seems to be to ignore the differences and assume that the processes generalize adequately.

Theory of Machine Mind

Even though using non-human agents, avatars, and stories for the research of theory of mind is common, the question of whether people apply theory of mind to non-human agents is rarely discussed. There are good arguments to be made for and against theory of mind use for artificial agents. In this subsection, I will discuss the arguments that can be made for and against the transferability of theory of mind research on human-human interaction to human-agent interaction.

The “computers are social actors” paradigm claimed that people apply social norms to interactions with computers (Nass et al., 1994). The results could not be replicated later using desktop computers exactly like in the original setup (Heyselaar, 2023). But the effect was frequently supported using novel technology of the time (Ho et al., 2018; S. Lee et al., 2019; Srinivasan & Takayama, 2016). It seems that novel technologies rather than computers themselves are treated as social actors (Gambino et al., 2020). A reason could be that when people lack experience with technology, they mindlessly apply the interaction schemata to which they are accustomed (Nass & Moon, 2000). Once the novelty wears off, they can apply interaction schemata that are more adapted to the specific technology (Gambino et al., 2020). The same could be true for theory of mind: people apply their assumptions about how minds work in general until they develop a more specific representation of the specific agent’s mind.

An argument against theory of mind use for artificial agents is that the decision processes of these agents might be fundamentally different from those of humans. How could a human observer comprehend the actions of a machine if machines have other motivations, use other sensors, and have different decision processes? The following two answers to this question

support the use of theory of mind in such situations. First, to accurately predict an agent's behavior, the observer does not need to reproduce the agent's exact decision-making process. If the agent is acting rationally, the observer can imagine how a rational agent should act given the information the observed agent has (Baker et al., 2009, 2017). Second, human decision processes can be as incomprehensible for a human observer as those of artificial agents. Humans report having a better understanding of human experts' decision processes than those of artificial agents, but if they are asked to explain the expert's or the agent's decision process beforehand, the rating of understanding is similar (Bonezzi et al., 2022). This shows that humans tend to overestimate their understanding of other humans, possibly due to their similarities and shared human experiences. However, the decisions in question are not informed by shared experiences but by the unique expertise of the agent in question. For example, participants were convinced that the decisions of a medical expert were more understandable to them than those of an algorithm. Only after being prompted to describe the decision process did the participants rate their ability to explain the decision of an algorithm and that of a human expert similarly (Bonezzi et al., 2022).

Another argument against a theory of machine mind is that people tend to mistrust automated decision-making (algorithm aversion; Burton et al., 2020; Dietvorst et al., 2015). This effect is discussed in depth in a later section. However, not all factors of trust are necessary to understand an agent's behavior (R. C. Mayer et al., 1995). The agent has to be reliable in their decisions, although it does not need to be benevolent toward the observer. The role of an agent's ability and observable performance for their trustworthiness are special and will be discussed in depth throughout the dissertation.

Some empirical evidence exists that supports the idea that humans readily use theory of mind with artificial agents such as robots (Hellström & Bensch, 2018). For example, people can anticipate a robot's intentions through mindreading (Ono et al., 2000). However, the mentalizing of an artificial agent is influenced by an egocentric bias even more than the mentalizing of a human

agent's process (Husemann et al., 2022). Neurologically, more theory of mind was measured for agents that appeared more human-like. The activation in the relevant regions was highest for human agents, but it existed for artificial agents too (Krach et al., 2008).

In conclusion, there are hindrances to a theory of machine mind, but none are insurmountable. The differences between human and artificial agents generally result in quantitive differences in a human observer's ability to anticipate their decision. A theory of machine mind for artificial agents is possible although it might differ slightly from what we have learned about a theory of human mind.

Connection to Dissertation

In this dissertation, I discuss how humans build mental representations of artificial agents. Traditional social cognitive theories suggest how humans represent other humans. However, the transferability of social cognitive findings to artificial agents is not fully established. I will provide evidence that the actions and decisions of artificial agents can be mentally represented by humans and will discuss the similarity of this representation to the representation of humans.

Competence

Many theories of social perception categorize the perception of others into two dimensions: often one dimension is related to the individual ability of the agent and the other to its emotional interrelatedness with other agents. I have already discussed agency and communion in the "Agency" section. Competence and warmth are closely related dimensions of social perception (Cuddy et al., 2008, 2009). Competence is theorized to be a facet of agency (Abele et al., 2016). In general, the concepts of agency and competence are related to social perception. In this section, I will discuss my understanding of expertise as an agent's competence in performing a specific task. Following a definition of expertise, I will describe the role of knowledge of expertise, focusing on the distinction between implicit and explicit knowledge. Finally, I will apply

the concept of expertise in a social context, explaining how humans can estimate their expertise in comparison to that of others.

Expertise

Expertise is the knowledge and cognitive skills of agents for solving a specific task. Experts distinguish themselves from novices through experience, knowledge, and problem-solving ability (Herling, 2000). Experience increases the expertise of an agent over time if the agent can infer information about their environment through interacting with it. However, a crucial piece of knowledge can increase expertise spontaneously. What exactly makes an expert is highly dependent on the task. To answer trivia questions, an expert needs extensive semantic knowledge of the field. In chess, on the other hand, a plethora of skills are necessary, such as pattern recognition, speed, and imagination (Reynolds, 1992; Van Der Maas & Wagenmakers, 2005).

I conceptualize expertise as the probability of making the optimal decision needed to reach a goal within an environment and the minimization of the errors that occur (M. Mayer et al., 2023). Experts can use their exhaustive knowledge of their environment to anticipate the effect of their options and decide which would bring them closer to their goal. The ideal expert will always make the best possible decision. While agents with great but not perfect expertise might make errors, the deviations from optimal decisions are rarely large. In contrast, novices lacking knowledge of an environment need to explore it (Daley, 1999; Hills et al., 2015). Each decision they make offers additional information about the environment, but the decisions themselves do not necessarily bring the novice closer to the goal. Each option is equally likely to be picked by the ideal novice. On average, this leads to large deviations from optimal decision-making.

Take the example of two gambling machines, machine A winning 60% of the time and machine B winning only 40% of the time. In this case, expertise could be defined as the probability of an agent choosing machine A. A rational agent with avaliable information about the winning chances would always choose the machine that is more likely to win. A novice with no information

would have to test both machines several times to get an estimate of their chance of winning. As the learning novice explores the chances of both machines, they build expertise through their experience and, after a few trials, start to exploit the more efficient machine A. As the novice builds knowledge about the machines through experience it develops expertise at the task of choosing the winning machine. There are other ways to gain knowledge, such as receiving information from other agents.

Knowledge

The expertise of an agent depends heavily on how well the agent knows its environment. I refer to the information about an environment that is available to an agent as *knowledge*. Knowledge is the information agents can use to anticipate future states of an environment. Usually, the knowledge of human agents originates in the person's memory, but it can also be stored externally in the form of notes or graphs. Knowledge can be separated into two major types: explicit and implicit (Reber, 1989). Explicit knowledge is what people generally associate with the word *knowledge*. It refers to the manifest and verbalizable information an agent has. This knowledge can usually be put into words, drawn visually, or communicated in some other way (Berry & Broadbent, 1984). Implicit knowledge refers to the tacit information possessed by an agent that influences its decisions and actions without the agent's complete awareness (Dienes & Berry, 1997). Even though implicit knowledge can guide a person toward task success, they would still be unable to provide the correct reasons for their actions.

There are various empirical approaches to separating implicit and explicit knowledge. Study participants trained on examples of artificial grammar could differentiate between examples belonging to the grammar and those that did not. However, most of the participants were unable to infer and report the grammar accurately. Therefore, the authors concluded that implicit knowledge guided the participant's decision on whether or not an example belonged to the grammar (Reber, 1967).

In dynamic decision tasks, participants learn to control an environment that is guided by a hidden, usually noisy function. In an example of a dynamic decision task, the participants learned to control the environment to achieve desirable outcomes over several trials (Gibson, 1996). Yet, as with the study of artificial grammar, most participants were unable to report the environment's hidden function despite demonstrating control over the environment (Gonzalez et al., 2017).

In tasks that use sequence learning, the participants are asked to react appropriately to stimuli. Usually, the stimuli appear randomly, but occasionally they are presented in a set sequence. Research participants react faster to stimuli if they have already experienced the sequence. In a study of sequence learning, most participants were unable to report the sequence, so the authors concluded that they must have learned it implicitly (Curran & Keele, 1993).

In the empirical Chapters 2 and 3, I manipulate the participants' knowledge to research how their available information influences their ability to judge an observed agent's expertise and anticipate their actions. Due to similarities between anticipating the actions of an agent and the states of a dynamic system, I specifically test the implicitness of the participants' acquired knowledge in Chapter 2.

Meta-Cognitive Judgment of One's Knowledge

An important skill of humans is to be able to judge their knowledge. An accurate judgment of the accuracy of one's knowledge allows a person to focus their attention on new information and devote fewer cognitive resources to known information. Furthermore, accurate metacognitive judgment helps people know when to voice their knowledge and when to stay silent to avoid spreading potential misinformation (for example, in group discussions or sequential collaboration).

Accurate metacognition is useful but may be difficult to acquire. People with little domain knowledge of facts about an environment seem to have particular difficulties in assessing how little they know (Dunning et al., 2003; Kruger & Dunning, 1999). The discussion of people's

ability to estimate their expertise is dominated by a hypothesis made by Kruger and Dunning (1999). They claim that at least some task knowledge is needed to accurately judge one's competence at a task. A complete absence of task knowledge leads to an underestimation of the task's complexity and consequently to an overestimation of one's expertise in performing it. In essence, Dunning and Kruger claim that the accuracy of metacognitive judgments about one's task performance is dependent on one's task performance (McIntosh et al., 2019). This hypothesis is supported by considerable empirical research (Dunning et al., 2003; Ehrlinger et al., 2008).

However, the results fail to replicate in independent labs; instead, novices show metacognitive insights equivalent to those of experts (M. Mayer et al., 2023; McIntosh et al., 2019, 2022). Additionally, the Dunning-Kruger effect can be explained by an statistical artifact through data simulation. Counterintuitively, the less accurate the measurement is, the larger is the effect (Nuhfer et al., 2016, 2017). Novices that by definition, have little expertise at a specific task and thus perform badly, are aware of their incompetence and behave accordingly (M. Mayer et al., 2023).

As opposed to the claims of Kruger and Dunning (1999), task information might not be the only information people can use to judge their task performance. Even bad performers might use introspection to gain metacognitive insight through feelings of uncertainty, judgment of learning, and knowledge of their task experience. However, when a person is not judging their own expertise but rather the expertise of an observed agent, the methods of introspection are not available. In the perception of others, task information likely plays a larger role in the judgment of one's own expertise. In the next section, I will discuss the observer's perception of an agent's expertise and its consequences for the predictability of their actions and perception of trustworthiness.

Connection to Empirical Work

Factors relating to competence, such as expertise and knowledge, as well as experience, performance, and ability, can be discerned theoretically. However, differentiating between the factors might not be as easy in practice. In the second experiment in Chapter 2, I compared the task experience of the participants. I assumed that participants with more experience knew more about the task (knowledge) and consequently were more able (ability) to perform (performance) the task better. In short, they had more task expertise. In the first experiment in Chapter 3, I varied the information the participants received about the task. I assumed that participants who received more task information could use this information and thus perform better at the task and have more expertise. In theory, a more experienced or more informed participant may not necessarily be more knowledgeable about the task. However, on average, a difference—likely very pronounced—is present, so the theoretical distinction between experience, information, expertise, and knowledge has little practical importance in the experiments I present. But it is important to keep in mind that the connections between the different aspects of competence are merely assumed and not specifically tested in the empirical work.

In conclusion, the perception of task expertise and other factors of competence are essential factors in interpersonal perception. Specifically, the willingness to rely on a partner in a collaborative task depends on how capable the partner is perceived to be. In the next section, I will discuss an observer's trust in an agent, focusing on the relevance of perceived agent expertise.

Trust

Trust is a crucial factor in collaborative tasks. The perceived competence and expertise of the agent, among other factors, affect whether or not an observer decides to trust it. To trust means to rely on the trustee's actions and thus be vulnerable to errors or malicious actions of the trustee due to uncertainty about their goals and abilities (R. C. Mayer et al., 1995).

The construct of trust is frequently used to explain willingness to depend on another agent in interpersonal relations. It describes how strongly an agent, the trustor, is prepared to rely on another agent, the trustee, to reach a goal. The trustor has to evaluate three attributes of the trustee: ability, benevolence, and integrity (R. C. Mayer et al., 1995). The agent's ability refers to the skill and knowledge they have that are necessary to fulfill the task satisfactorily (see competence). Benevolence is the willingness to act in the trustor's best interest. Finally, integrity describes the agent's adherence to its principles and the acceptability therof. Essentially, the trustor evaluates whether the trustee is able and willing to reliably perform the task without direct surveillance according to the trustor's needs. However, like mental states (see Section "Social Cognition – Theory of Mind"), these attributes of the trustee are not immediately perceptible.

The judgment of trust or trusting behavior is a product of two major factors. First, the trustor might have a more or less trusting disposition. Second, the trustee's attributes, as perceived by the trustor, influence the trust that is expressed (R. C. Mayer et al., 1995). An accurate calibration of trust is crucial to a successful collaboration (Cancro et al., 2022). When the observer's perception of the trustworthiness of an agent exceeds the agent's actual trustworthiness, the observer overtrusts the agent and might delegate tasks that the agent will not perform well. Undertrust occurs when the perceived trustworthiness is lower than the actual trustworthiness. Undertrust can lead to inefficient collaboration if the observer does not delegate tasks that the agent would perform well. Both over- and undertrust result in inefficient teamwork between two agents (De Visser et al., 2020).

Due to the increasing prevalence of automated processes in the human environment, there has been an increased interest in trust in automation (i.e., Chiou & Lee, 2016; Hancock et al., 2011; Hoff & Bashir, 2015; J.-Y. Jian et al., 2000; Kaur et al., 2023; Kohn et al., 2021; J. D. Lee & See, 2004; Vinanzi et al., 2021). A crucial difference between humans' trust in other humans and their trust in autonomous systems has been identified. Humans avoid relying on automated systems if they see them make errors (Burton et al., 2020; Dietvorst et al., 2015). This

“algorithm aversion” is present even if the algorithm is performing objectively and noticeably better than the human, suggesting that the human participants are calibrating trust poorly (Dietvorst et al., 2015). A possible reason for this aversion could be that a system’s error can reflect an error in programming and will be repeated in future situations. In contrast, human errors can be forgiven as momentary lapses. Another explanation is that humans prefer to trust other humans because they feel they can understand the reasoning behind human decisions better than that behind algorithmic decisions, even though the decision processes of humans are as hidden from observation as algorithmic decisions are (Bonezzi et al., 2022).

Even though the majority of research supports algorithm aversion (Burton et al., 2020), a recent study showed that people prefer relying on algorithmic decisions to trusting human decisions (Logg et al., 2019). Attitudes toward algorithmic decisions are likely to change over time, as people become increasingly familiar with specific tools and automation in general. The participant's familiarity with the technology influences any study on humans' attitudes toward automation. This dependency makes it especially difficult to compare studies across time, since novel technology from decades ago is outdated and replaced by new versions, potentially multiple times.

In conclusion, trust is a complex construct that significantly influences human-agent collaboration. While humans are adept at assessing the trustworthiness of other humans, they often struggle to calibrate trust in artificial agents. This lack of experience can lead to undertrust, with humans failing to leverage the capabilities of AI systems, or overtrust, when they rely on AI without critical evaluation. An accurate understanding of an agent’s expertise is necessary to calibrate trust. This dependence makes trust a great measure of expertise assessment.

Perception of Agents

Identifying the abilities and knowledge of others is crucial if one’s environment is manipulated by agents other than oneself. Since humans are social animals, they constantly enrich each other's environment. Stereotypes and experiences give rich information about

individual humans and allow the formation of attitudes based on superficial features before an interaction (Cuddy et al., 2009; Fiske, 2018). The information might be false and need revision (Frith & Frith, 2012; Galinsky & Moskowitz, 2000), but it allows humans to navigate a social environment with confidence and a sense of security (Imhoff & Koch, 2017). Artificial agents are a new category for which most people's stereotypes and schemas have not yet fully developed. The continuous improvement of technology makes it even more difficult to form a coherent image of this specific group of agents.

In this dissertation, I am driven by the goal of understanding how human observers form a mental representation of an agent based on the agent's actions within their task and how this is limited by the observer's task expertise. I am specifically interested in the social perception of artificial agents. However, I discuss how some of the findings can be generalized to interactions between agents of any kind.

In this section, I will introduce the three main questions regarding the perception of agents in general and artificial agents specifically, which I intend to answer empirically throughout this dissertation. Under what conditions can an observer predict the actions of an agent? How well do observers with varying expertise judge the expertise of another agent? How does the perception of expertise influence the observer‘s trust in the agent?

Prediction of Actions

Implicit prediction of an environment's future state is theorized to be a fundamental process of perception (Bubic et al., 2010; Clark, 2013; Friston, 2010; Friston et al., 2003). Prediction errors lead to heightened attention and learning to avoid repeating the error and to gain a more accurate representation of the physical world. The anticipation of future states becomes more difficult when other agents impact them. In such cases, the observer has to model the actions of the other agents as well as the behavior of the environment.

Humans similarly learn about others, by prediction and error reduction (Joiner et al., 2017). However humans do not consciously and continuously predict future events and actions of

others. Rather, they have a more abstract theory of mind that represents the mental states of other agents. Healthy human observers can argue about the mental states of agents in terms of the agent's goals, knowledge, and beliefs—even beliefs that the observer knows to be false (Wellman et al., 2001). Using this information, the observer can make predictions about the behavior of the agent by computing how a rational agent would act based on the observed agent's presumed beliefs and goals.

According to the definition I offered earlier, experts are more likely than novices to make optimal decisions when performing tasks in their area of expertise and less likely to make major errors. The restriction of the options makes an expert's decision-making more predictable than that of a novice. The shared level of expertise gives experts a common ground on which they individually base their decisions. However, novices who try to predict the actions of an agent do not have the same or similar knowledge as the expert by definition. With less task knowledge, they might not be able to use this predictability to their advantage. To accurately predict the actions of an expert agent, the observer has to have at least the same task knowledge as the agent. If they do have at least the same experience and knowledge about the task as the agent, they can infer the agent's knowledge accurately. Only then is it possible for the observer to mentalize the agent's decision-making process. Either they have the same task information that lead to the same decisions, or the observer knows more than the agent and imagines the decisions an agent would make when lacking crucial information about the environment.

In conclusion, predicting an agent's decisions in a task is only possible if the observer and the agent are both experts and thus agree on the best practice for completing the task or when the observer knows more than the agent about the task and also knows the agent's approximate task knowledge.

Perception of Expertise

Accurately judging the competence of agents in one's environment is important for successful communication and accurate trust and resource allocation (Nickerson, 1999).

However, the competence of an agent is not directly perceptible and has to be inferred from several sources of information. The introspection that gives even novices meta-information about their performance is not available when they are judging the expertise of others (see Section "Competence – Knowledge"). Given the importance of accurate expertise judgment, human societies have created titles, labels, and degrees that ideally reflect the expertise of an agent in a certain area. In direct interaction, an expert might take into account what a human agent looks like, even though that is unreliable information (Nisbett & Wilson, 1977). However, in online interactions and interactions with artificial agents, such sources of information are rarely available, and the observer has to make a judgment based only on the agent's observed actions.

Lucassen and Schraagen (2011) created a model that categorizes the features people use to judge whether an information source is trustworthy into source, surface, and semantic features. Source features refer to the characteristics of the information source, such as titles and degrees. Surface features are features of the information source that do not relate to the content. Since the authors used online articles to validate the model, some of the surface features they identified were the writing style of the articles and the length of the texts. By semantic features, the authors referred to the content of the texts. Notably, only 6.7% of novice observers reported to use semantic features in developing their rating of trustworthiness and rather base their judgments on source and surface features. While expert judges also use source and surface features for their assessment of a text's trustworthiness, a significantly higher percentage (38.2%) claim that their judgment is based on the semantic features of the text (Lucassen & Schraagen, 2011, p. 1239).

The categories of the model are transferable from text-based expertise judgment to observation-based judgment. Imagine observing an agent solving a jigsaw puzzle. The agent and its behavior can provide information resembling that supplied by a text. The source information is similar: you might have information about the agent, and you can see its physical appearance, which offers additional information. Surface information of observable behavior is subtler; it could

include the agent's speed, confidence, and body language in picking up, moving, and placing pieces. Semantic information refers to the observable decisions that are made. Does the agent sort the pieces by color? Do they start on the outer edge? What piece do they pick up next? Such cues can give information to experienced puzzlers. Observers that are novice puzzlers, however, might have difficulty interpreting these cues.

Novice judges of expertise are unable to use the semantic features due to a lack of domain expertise (Lucassen & Schraagen, 2011). I claim that the observer's domain expertise is a fundamental limitation when judging an agent's expertise based on semantic features. This limitation is a problem, especially in online settings, where the source features are hidden or even falsified due to anonymity. Surface features can also be manipulated and might not give valid information about the agent's expertise. In Chapter 3, I test the hypothesis that the expertise of the observer limits their ability to make accurate judgments about the expertise of an artificial agent based on the "semantic features" of their behavior.

Judgment of Trust

Agents such as humans and animals have existed throughout human history. Over millennia, people learned the schemata and norms of these agents and developed stereotypes about them passed on by other individuals of their environment and continuous first-hand experience. However, new types of agents have emerged in everyday life in the form of autonomous technologies with varying degrees of agency, such as recommender algorithms, targeted marketing, and large language models. These artificial agents evolve more rapidly than biological agents and do not necessarily have a physical form or even a continuous existence. Such aspects of artificial agents prevent people from acquiring an accurate model of these agents that could inform their expectations, leading to a more malleable attitude toward novel technology than toward humans (Buder et al., 2024). The inaccuracy of their expectations makes it difficult for people to calibrate their trust in artificial agents accurately even when they observe the artificial agent outperforming a human agent repeatedly (Burton et al., 2020).

Although people have far less experience interacting with artificial agents than humans, they are not uninformed. The media and the companies behind artificial agents present information about an agent before first-hand experience becomes available. However, this information can be unreliable and even manipulative, which is why I am interested in the effects of an agent's reputation on its use and trust calibration.

I have already discussed the most important factors involved in how observers reach a judgment of trust about an agent. Perceptions of reliability and expertise are commonly mentioned as determinants of trust in an agent. Trust judgments are influenced by the situation, features of the trustor, and perceived features of the trustee (R. C. Mayer et al., 1995). However, as I discussed above and will test in the empirical chapters, the accuracy of an observer's prediction of an agent's future actions and the perception of the agent's expertise are limited by the observer's task expertise. An inability to appraise these important features accurately might hinder an accurate trust allocation.

Proposed Process and Empirical Tests

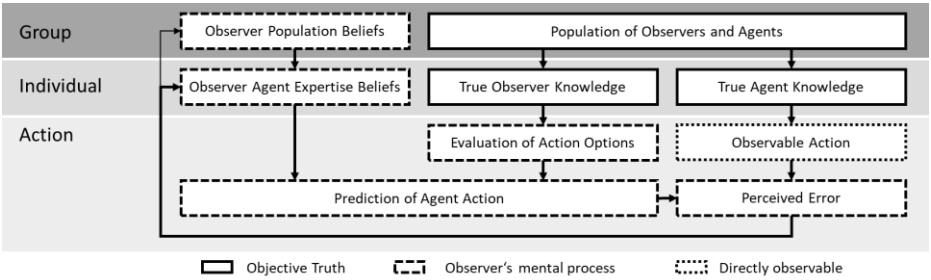
The perception of an agent's mental state is a difficult task. In this section, I propose a process of inference that could allow an observer to judge expertise and trust from the repeated observation of an agent's action (Figure 1). The process can explain the dependence of the perception accuracy on the observer's task expertise. Furthermore, I propose three methods for measuring the variables and test the hypotheses derived from this process (Figure 2).

Expertise is distributed in a population of individuals, and two individuals acting as an agent and an observer are drawn from this population. The observer wants to estimate the agent's knowledge of a task; they start by remembering what they know about the population in general. When engaging with the agent, this knowledge can be further informed by the source features of the agent, such as size, gender, reputation, or ethnicity (Lucassen & Schraagen, 2011). The estimated population's expertise distribution and stereotypes give the observer an initial anchor

for estimating the agent's expertise. This anchor can be adjusted by observing the agent's performance (Tversky & Kahneman, 1974).

Figure 1

Theoretical Process of Inference from an Agent's Actions



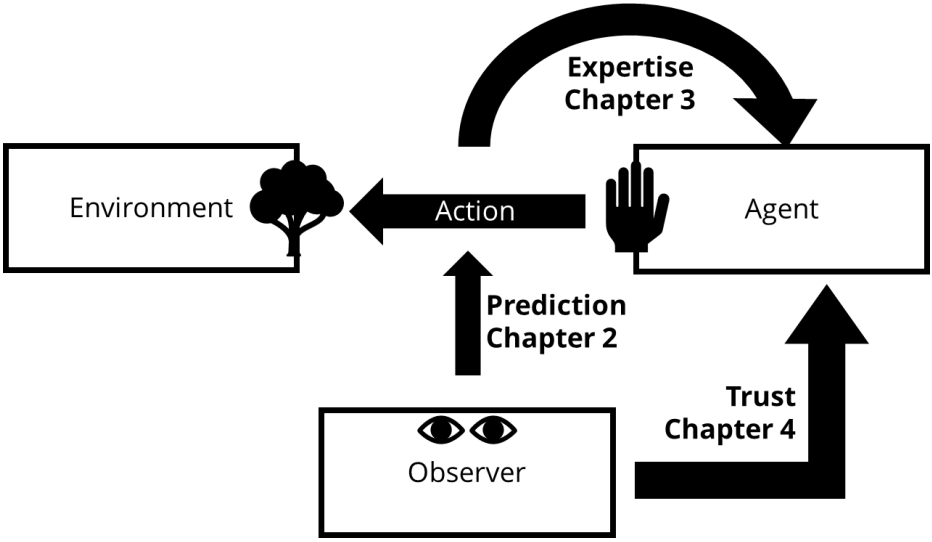
The process of adjusting the initial estimation starts with the observer evaluating the options of the agent by performing the task mentally and evaluating the expected outcomes (Figure 1). The greater expertise of the observer aids in the formation of an accurate evaluation of the agent's options. Next, the observer can view the agent's decision. By evaluating the options and the observed action, the observer can assess the severity of any errors the agent might have made. The perceived severity of the agent's errors then informs the estimation of the agent's expertise. Additionally, the observer can estimate the likelihood of the observer picking each option from the estimated expertise of the agent and the evaluation of the agent's options. This allows the observer to make predictions about the agent's decisions and use prediction errors to update their internal model of the agent's knowledge and decision process to improve their expertise estimation further.

In this dissertation, I use three measures to test hypotheses derived from this process of inference. First, I test how human observers create an individual mental model of an agent's decision-making by observing its actions. Second, I examine how the accuracy of expertise judgments is dependent on the observer's expertise. Finally, I test the effect of the observer's

perception of expertise and their prior assumptions on their trust and willingness to depend on an agent (Figure 2).

Figure 2

Empirical Test of the Process' Elements



The prediction of action begins with the observation of the agent and its environment. The observer generally assumes the agent to be rational, which means it has a goal and chooses actions to the best of its abilities to achieve that goal (Baker et al., 2009, 2017). If the agent's goal is unknown, the observer must first identify the goal. Given a rational agent, all its actions should manipulate its environment to achieve that goal. An observer with knowledge of the environment can, over time, recognize the trend of the agent's manipulation and eventually identify the goal. If the agent's goal is known, the observer can predict the agent's actions by identifying the rational decisions the agent can make with the limited information and time the agent has (Berke & Jara-Ettinger, 2021). To do this, the observer has to have at least the same information and resources

as the agent and needs to know what information the agent and the observer share. In Chapter 2, I present two experiments that show the dependence of observers on their understanding of the environment for action prediction.

As in the prediction of action, the observer has to assume a rational agent to judge the expertise of an agent at a task or in an environment. Given the agent's goal and the necessary information about the environment, the observer can judge the fitness of every single decision the agent makes to reach the goal. Together, these decisions can be used to assess the expertise of the agent. In Chapter 3 of the dissertation, I present two studies of the effect of the observer's knowledge on the accuracy of their judgment of an agent's expertise.

When an observer has identified an agent as rational and understood its goal and its knowledge of the environment, the observer can finally decide whether to trust the agent. However, making such an informed decision requires the observer to have observed the agent's actions repeatedly. In real life, such decisions have to be made based only on the observer's prior information, such as knowledge about the population, stereotypes, and the agent's reputation. In Chapter 4, I present an experiment on how the agent's reputation affects the observer's trust and willingness to rely on the agent before the interaction and first-hand observation.

Through the studies I present in this dissertation, I aim to explain and test key elements of the process by which a human observer understands an agent's decision process and the consequences for their perception of the agent's expertise and trustworthiness. The main focus of these studies is on the observer's knowledge of the environment and the effect of the varying knowledge of the environment on their perception of the agent.

Chapter 2 – Agent Behavior Prediction

The second chapter is a collaborative manuscript of Fritz Becker (first author), PD Dr. Jürgen Buder (second author), and Prof. Dr. Markus Huff (third author). It has the title “To understand my decisions, you must understand my task: Expertise effects in theory of mind.” At the time of the thesis’s release, the manuscript is under review at Acta Psychologica.

Abstract

Human behavior is frequently constrained by the behavior of other agents (increasingly artificial agents like machines and software). In a cooperative setting, each individual needs to understand the partners' intentions and corresponding actions to plan their actions adequately. Misunderstandings have adverse effects and diminish the efficiency of cooperation. We created a game-based experiment to study the effects of understanding a partner's actions in a cooperative setting. In this game, the players depend on each other's ability to make good decisions to succeed. In two studies (N = 87, N = 281), we collected data on the understanding of an artificial agent, operationalized as the ability to predict its actions and the skill at the task itself. The participants improve at predicting the agent's actions and at their subtasks in the game over time. Following a misunderstanding, the participants' performance at their subtask was worse, as measured by their actions' quality, speed, and efficiency. Results of Study 2 suggest that the improvements in predicting the agent's actions are likely the result of an improved understanding of the game rather than an improved understanding of the agent.

Introduction

Most of the time, environmental changes are not under an individual's control alone. In fact, the opposite is true: human behavior is massively influenced by the behavior of other agents in the environment, whether it is a boss, a partner, or a cat, but also the central heating that can autonomously decide and take action (Franklin & Graesser, 1997). Understanding and correctly predicting how these agents behave can help individuals achieve their goals. But this is a daunting task, as other agents' decisions, emotions, desires, and plans are often obscured from direct perception. However, humans have developed abilities to find regularities and to predict and anticipate agent behavior.

Predicting agent behavior

Several bodies of research have addressed the question of how humans succeed in predicting the behavior of other agents, and several psychological concepts that capture this ability have been proposed, such as *Theory of Mind* (Baron-Cohen et al., 1985; Wellman et al., 2001), *perspective-taking* (Epley et al., 2004), *mentalizing* (Van Overwalle & Vandekerckhove, 2013), *folk psychology* (Churchland, 1989), and *mindreading* (Apperly, 2011). For example, theory of mind describes the developmental step of children to understand that the mental states of others might differ from their own. It is also used for a person's general ability to track the knowledge and decision-making of other agents (Aboody et al., 2021; Baker et al., 2017; Berke & Jara-Ettinger, 2022). In the most frequently used experiment of theory of mind, the false belief task, an agent places an item in a box and leaves the room. Subsequently, the item is moved to another box, and the agent returns to the room. The participant (observer) is asked to predict in which box the agent will search for the item (Wellman et al., 2001). Crucially, in scenarios like the false belief task, there is always a knowledge gap between an observer who has complete knowledge of the events and the agent who usually has limited knowledge about the state of the environment. But what will happen if this knowledge gap is reversed so that an observer has limited knowledge and is tasked with predicting the behavior of a fully informed agent?

This question can be answered from literature investigating how humans learn to predict or control rule-based systems. For instance, work on artificial grammar learning showed that people are able to learn a grammar just by being exposed to positive examples of said grammar (Reber, 1967). Over repeated trials with feedback, they improved their judgments, thus showing signs of learning. However, in most studies, this did not result in explicit, verbalizable knowledge, which is surprising, given the improvement the participants showed during the task. This kind of learning is called *implicit learning* (Reber, 1989). Implicit learning has also been studied with dynamic decision tasks. An example is the *sugar production factory* (Gibson, 1996). Participants are asked to control a fictional factory that produces an output of sugar based on the input of a number of workers from the participant, the last trial's output, and a random factor. After a few rounds, the participants consistently brought the factory to the level they were told to aim for. However, their explanations of how the factory operates were inaccurate and varying, again suggesting that implicit learning and explicit learning can be disassociated. In repeatable but noisy tasks with feedback, like the ones mentioned above, humans frequently learn implicitly (Conway et al., 2010). Through implicit learning, people can pick up on statistical regularities in noisy environments (Reber, 1967) and can use them to their benefit. Important aspects of human knowledge acquisition are claimed to be learned implicitly, like language, motor skills, or social norms (Perruchet & Pacton, 2006). Contrary to explicit knowledge, implicit knowledge is neither verbalizable nor manipulable (Weinberger & Green, 2022), but it is very quick to be recalled and sensitive to signals obscured by randomness. Results like these lead to the belief that a system that learns from trial and error must be separate from a verbal and rule-based system (Sun et al., 2005), a distinction that integrates well with other dual-process theories (Kaufman et al., 2010).

There is reason to believe that predicting the behavior of social agents will follow similar learning patterns to the prediction of rule-based systems like artificial grammars (Nass & Moon, 2000). For instance, mentalizing might not always be a fully conscious and effortful process. It might be the case that some processes of mentalizing are automatic and below the threshold of

consciousness. The specific role of implicit processes in mentalizing is currently subject to discussion. There are compelling arguments both for (Schneider et al., 2017; Van der Wel et al., 2014) and against (Kulke et al., 2018) a relevant role of implicit processes during mentalizing. One possible reason for this ambiguity might be the focus on the false belief task as a test of theory of mind (Bloom & German, 2000). In any case, the notion that human observers will gain an implicit understanding when repeatedly predicting the behavior of an agent provides a working hypothesis for our Study 1. In this study, participants with low expertise about a task had to repeatedly predict a choice that a computer agent would make and received feedback about the accuracy of their prediction. Based on the literature on implicit learning, we expected that participants' prediction performance about agent behavior would improve over time, at least when the agent is performing according to some rules. To capture this improvement, we compared a condition in which the low-expertise participant had to predict an agent who made sound, rule-based decisions (high agent expertise) to a condition in which a low-expertise participant had to predict an agent who made random decisions (low agent expertise). We expected learning to occur in the high agent expertise condition. However, we also expected that participants would have difficulties explicitly verbalizing the rules that the agent uses.

The role of observer expertise

Provided that low-expertise participants improve in predicting the behavior of a high-expertise agent over time, the exact mechanism of how this occurs needs to be clarified. Two competing hypotheses could be made. The first hypothesis is built on the notion that observers will form a mental representation of an agent (its knowledge, its motives, its behavioral preferences) that becomes increasingly detailed over time. However, several lines of research suggest that humans do not build full-fledged (and costly) mental representations about the mental contents of other agents but rather make do with more parsimonious heuristics. For instance, it has been argued that the false belief task simply requires registering a mismatch between one's own knowledge and an agent's knowledge rather than an observer's formation of a

detailed mental content of an agent (Deschrijver & Palmer, 2020). Similarly, research on perspective-taking suggests that the ability to take the perspective of an agent does not necessarily involve putting oneself into an agent's shoes. Rather, it seems that observers start by putting an agent into their own shoes (egocentric anchoring, Tversky & Kahneman, 1974; Reynolds, 1992) and go through stepwise adjustments until the mental content of oneself and inferred mental content of an agent are in alignment (Epley et al., 2004). This leads to a second hypothesis about learning to predict an agent: it could be the case that improved prediction ability reflects gains in expertise about the task rather than expertise about the agent. In fact, task expertise in itself comes along with benefits that might explain the increasing ability to predict the behavior of an agent. A hallmark of task expertise is that experts can narrow down the space of all possible decisions to the few ones that promise success in the task (Daley, 1999). Therefore, an expert will deviate less from objectively correct decision-making. This, in turn, makes their behavior more predictable to an observer, provided that the observer also has similar task expertise. In tasks where there is a single optimal decision, two experts should make the same decisions, and given they know of the other's expertise, expect the other to make the same decision. These learned skills allow experts to mutually become more predictable and reliable partners in a cooperative task. If there is no single objectively correct decision, an expert can at least narrow the options of the other expert down and make a better prediction than a novice could. In these situations, anchoring the prediction in one's own decisions is a sensible strategy for action prediction (Nickerson, 1999). In other words, in situations where the observer and agent have similar task expertise, observers might forgo the effort of mentalizing by asking themselves, "What would I do in this situation?" and build their agent prediction upon this premise. By gaining task expertise, novices might become better at anchoring their agent prediction on their own decision-making. In sum, improved agent prediction might reflect a gain in expertise about the agent or a gain in expertise about the task. To the best of our knowledge, the role of observer task expertise has not been investigated in the context of agent behavior prediction.

Study 2 disentangles the two competing hypotheses about agent prediction (mentalizing vs. task experience) by experimentally varying observer expertise. If the prediction of agent behavior depends on a detailed mental representation of the agent (mentalizing), both observers with low and high task expertise should exhibit an improvement in subsequently predicting agent behavior over repeated trials. However, if task experience is the key to agent prediction, an observer with high task expertise should be able to accurately predict the behavior of a high-expertise agent from the beginning of their interaction, and this performance would not markedly increase over repeated interactions with the agent. Conversely, the prediction of an observer with low task expertise should improve over time (reflecting a stronger gain in task expertise). In sum, by manipulating the task experience of an observer, it is possible to investigate to what extent the prediction of agent behavior shows signs of mentalizing.

The present studies

Two preregistered studies combined research on implicit learning and research on mentalizing by having participants repeatedly predict the behavior of an agent. To do so, participants collaborated with a bot in a modified version of the video game Tetris (inspired by Kulms et al., 2015, 2016). The collaboration involved two steps per trial: In the first step, participants saw four Tetris-like pieces, of which the bot would have to pick one. The task of the participant was to predict which piece the bot would choose, after which visual feedback about the bot's actual choice was provided. In the second step, the participant was then required to put the bot-selected piece into the playing field. This was repeated for 200 trials in both studies.

Study 1

The goal of Study 1 was to apply an implicit learning paradigm to social contexts (i.e., moving from predicting a system to predicting an agent). Participants either interacted with a smart bot (which was guided by rules similar to those of an experienced Tetris player) or with a random bot (which picked pieces arbitrarily). In line with extant literature, we hypothesized that prediction behavior increases over trials when interacting with a smart (vs. random) bot.

Furthermore, we also predicted that the increase in prediction performance was implicit rather than explicit (Gonzalez et al., 2017). Based on the notion that implicit learning unfolds in a cognitive system that is based on trial and error, we also preregistered that false predictions slow down subsequent participant behavior in a trial (i.e., longer time and more observable actions in placing the selected piece into the playing field).

Methods

Participants

A power analysis indicated that 80 participants are sufficient to detect a learning effect of $\hat{\beta} \geq 0.05$ with a power of $1 - \beta = .8$. We used the Prolific Panel to acquire participants who were compensated with 6.25£ (8.58$) for an estimated 50-minute study. The only requirement for participation was to speak English fluently and be older than 18, causing the sample to be diverse concerning nationality and occupation. The final dataset included 87 participants (29.89% female, 70.11% male, mean age: $M = 26.74$ years).

Game Environment

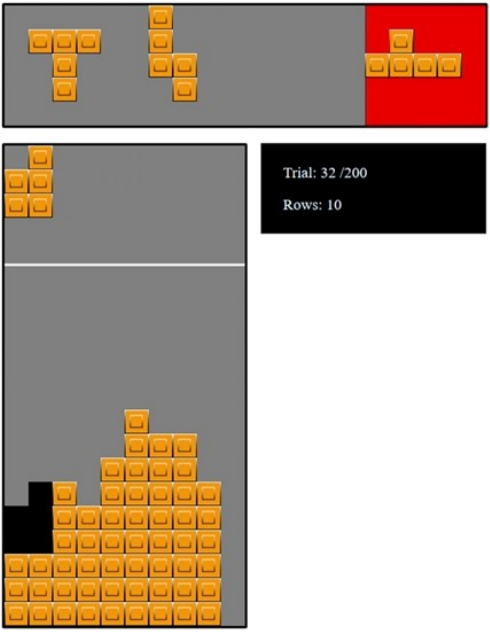
The experiment's stimulus material was inspired by the game Tetris (Pajitnov, 1984). As in the original game, the players built a structure using differently shaped pieces. Deviating from the original game (in which the game pieces - tetrominoes - were formed from all possible combinations of four squares of the same size), we used pieces from the combination of five squares. This increased the game's complexity, and the game appeared novel to the participants. As opposed to the original game, the pieces did not fall on their own. After having decided on a rotation and position, the participants dropped the piece all the way by pressing a key. This led to another difference from the original game: players could not horizontally "slide" a piece under existing structures. Like in the original game, a row was cleared when it was filled with squares without any holes. Removing multiple rows at once did not give bonus points in our version, though.

Each trial consisted of two phases. In the choice phase, a bot chooses a piece out of four randomly generated ones. At the same time, individual participants were instructed to pick a piece that they believed the bot would choose. In the placement phase, participants received the piece that the bot chose and placed it on the structure. The received piece implicitly provided feedback on the choice phase.

The piece that the bot selected was highlighted in green if it matched the prediction and highlighted in red in case of an incorrect prediction (Figure 3). Following the placement phase, complete rows were cleared automatically and counted if necessary. Then, the subsequent trial started with the choice phase. During the experiment, the participants' choices and current game states were logged. Additionally, behaviors like the time to choose and the number of movements participants made in a phase were logged on each trial.

Figure 3

State of the experiment following a false prediction.



Note: The piece predicted by the participant is highlighted red (for wrong), and the piece chosen by the bot is available to be placed on the structure below by the participant.

Material

For the test of implicit knowledge, participants were presented with ten new trials in which they had to predict the bot's choice again. Participants' explicit knowledge was tested with ten statements about the bot that were either correct or incorrect (Appendix A1). For example: "The algorithm tries to keep the number of holes small." This item is wrong for the random but true for the smart bot. "The algorithm picks a random element." This is true for the random and false for the smart bot. A questionnaire of mutual understanding was developed specifically for this study. Examples from the questionnaire are "I have the impression that I can understand the

algorithm's decision" and "The actions of the algorithm were confusing to me" for a negatively worded item (Appendix A2). The complete mutual understanding questionnaire consisted of 8 items and had a Cronbach's alpha value of $\alpha = .91$.

Procedure

The entire study was conducted online. At the start of the study, participants were informed about the study's goals and the data security policy. After providing informed consent, we collected demographic data. Participants received all information about the game environment, including the game's mechanics. The participants were informed that a bot would play with them. To avoid unintended expectation effects, we did not instruct participants about the specific nature of the bot partner (i.e., that there are two versions – a smart and a random bot). They were further instructed to try and predict the bot's actions as well as they could, even though it had no direct influence on the game score. Following the instructions, they completed 200 trials of the experiment, with the opportunity to take a break after every 50 trials.

Next, we assessed the participants' implicit and explicit knowledge of the bot. Next, participants were asked to freely report their impression of the rules by which the bot acted. Afterward, they received the questionnaire on their impression of mutual understanding between the bot and themselves. After the study, the participants were fully debriefed about the intentions and manipulation of the study. Finally, they were given an opportunity to retract their data.

Design and Variables

A univariate between-subjects design with *bot-type* as the independent variable was realized. The participants either played the game with a smart bot that selected fitting pieces based on the rules to minimize the number of holes and the height differences between columns or with a random bot that randomly selected the pieces.

As dependent variables, we used the *accuracy of prediction*, the *time* and *number of actions* for a placement, and the *quality of the placement* judged by the smart bot. These variables are used to quantify the participant's task performance. The quality of the placement is a binary

variable that was one if the participant placed the piece as the smart bot would have and zero for a different placement that is judged worse by the bot, respectively.

The results from the implicit and explicit knowledge tests are mean scores. The same is true for the subjective rating of mutual understanding.

Analyses

The analyses were conducted using the lme4 package and its extension lmerTest in R (Bates et al., 2015; Kuznetsova et al., 2017; R Core Team, 2023) and followed the preregistration (https://aspredicted.org/F6N_NX8). The data and scripts used for analysis are made public at https://osf.io/cu53e/. We used a logistic mixed model to analyze the influence of trial (amount of experience, logarithmic) and bot type on the ability to anticipate the bot's actions. As random effects, we included the random intercept and slope per participant.

We fitted linear mixed models with bot-type (smart vs. random), trial, and correctness of the prediction (main effects and interactions) as predictors separately for the three dependent variables. As preregistered, two of the outcomes were the time participants needed to place the piece and the number of actions it took to place a piece. Placement quality was added as a third outcome and analyzed using a logistic mixed model. Even though we preregistered random slopes and random intercepts for individuals over time, convergence issues forced us to settle for random intercepts for individuals. The placement time of two participants' trials exceeded 50 minutes. Those were excluded from further analysis.

Participants' implicit and explicit knowledge about the bot was analyzed using a mixed model, with bot-type (smart vs. random) and knowledge test (implicit vs. explicit) as predictors and the test score as the outcome.

We analyzed the mutual understanding questionnaire exploratorily. We present a simple t-test on mutual understanding for the two experimental groups.

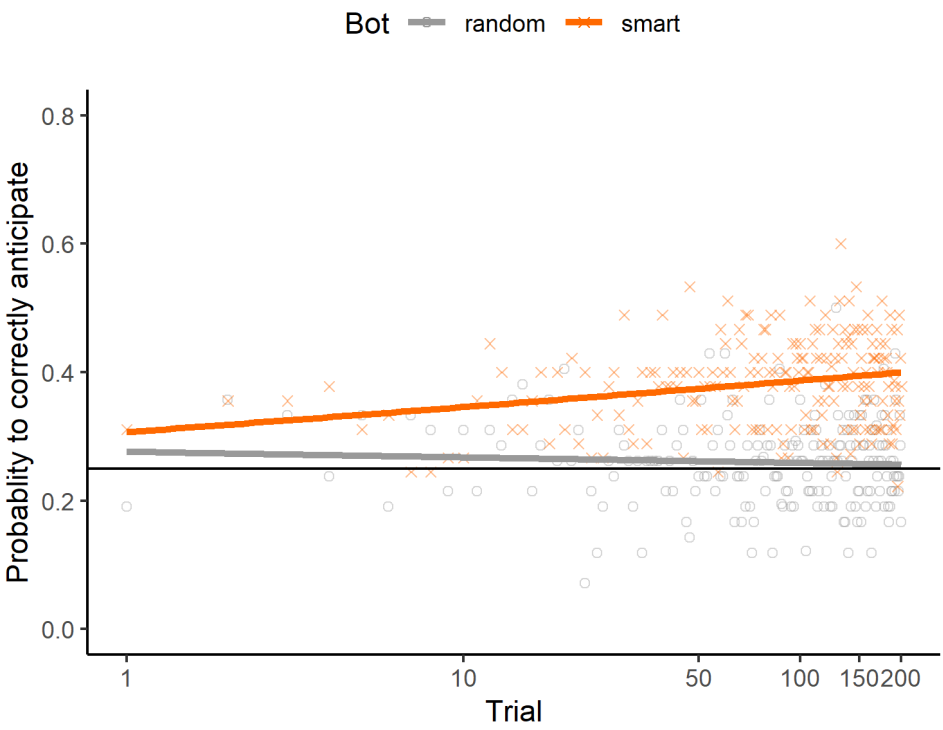
Results

Prediction

The analysis of the probability to correctly predict the next decision of the bot showed an interaction effect of the type of bot (random–smart) with the logarithm of trial, $\hat{\beta} = 0.10$, 95% CI $[0.02, 0.17]$, $z = 2.62$, $p = .009$. This suggests an ability to learn from the behavior of a bot if it is acting goal-directed (Figure 4).

Figure 4

Learning curves of the participants with the smart bot and with the random bot.



Note: The constant black line at y = 0.25 represents the guessing probability.

Task Performance

We used three measures for the performance of the participants: the time they needed to find a fitting place for the piece, the amount of actions they needed to place the piece, and whether the placement quality was the same as the smart bot would have done.

Time to Place. Modeling the "time to place" showed a significant main effect of trial, $\hat{\beta} = -0.02$, 95% CI $[-0.02, -0.02]$, $t(17292.34) = -18.40$, $p < .001$, correctness of prior prediction, $\hat{\beta} = -1.70$, 95% CI $[-2.22, -1.19]$, $t(17295.58) = -6.49$, $p < .001$, and type of bot, $\hat{\beta} = -2.28$, 95% CI $[-3.26, -1.30]$, $t(110.86) = -4.55$, $p < .001$. Thus suggesting that the participants got faster over time, were faster if they collaborated with the smart bot, and were faster after a correct decision. Furthermore, the interaction effect of trial and correctness was significant, $\hat{\beta} = 0.01$, 95% CI $[0.00, 0.01]$, $t(17296.03) = 2.75$, $p = .006$, as was the interaction effect of trial and bot type: $\hat{\beta} = 0.01$, 95% CI $[0.01, 0.01]$, $t(17292.79) = 5.33$, $p < .001$ (Figure 5A), indicating a convergence effect of training. Neither the interaction between bot type and the correctness of the prediction nor the three-way interaction were significant.

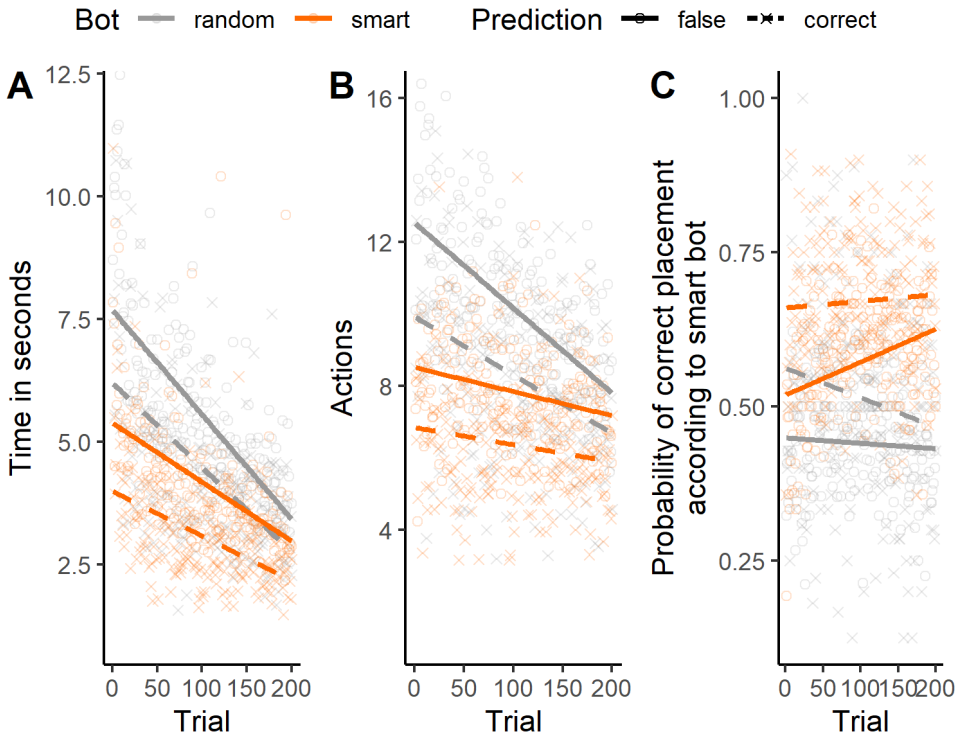
Number of Actions to Place. We found a similar pattern of results for the number of actions the participants needed to place the piece as for the time. The correlation between the time and the number of actions it takes to place a piece is strong, $r = .70$, 95% CI $[.69, .71]$, $t(17362) = 129.46$, $p < .001$. This correlation is plausible since it takes more time to place a piece if more actions are taken. As with the time it takes to place a piece, the number of actions it takes is influenced by the trial, $\hat{\beta} = -0.02$, 95% CI $[-0.03, -0.02]$, $t(17286.79) = -12.80$, $p < .001$, correctness of prior prediction, $\hat{\beta} = -2.59$, 95% CI $[-3.40, -1.78]$, $t(17291.01) = -6.28$, $p < .001$, and type of bot, $\hat{\beta} = -3.94$, 95% CI $[-5.32, -2.56]$, $t(119.30) = -5.60$, $p < .001$. The interaction of trial with the correctness of prediction, $\hat{\beta} = 0.01$, 95% CI $[0.00, 0.01]$, $t(17291.58) = 2.07$, $p = .038$, and with the bot type, $\hat{\beta} = 0.02$, 95% CI $[0.01, 0.02]$, $t(17287.35) = 6.28$, $p < .001$, were significant

again. Yet the interaction of bot-type and prediction and the three-way interaction of bot-type, prediction, and trial were not significant (Figure 5B).

Placement Quality. As a third indicator of the alignment between the bot and the participant, we collected data on whether the participants in both conditions placed the bot-selected piece as the smart bot would have done. The correlation of this measure with time to place, $r = -.10$, 95% CI $[-.12, -.09]$, $t(17362) = -13.84$, $p < .001$, and number of actions, $r = -.09$, 95% CI $[-.10, -.07]$, $t(17362) = -11.69$, $p < .001$, is much smaller than those variables' correlations; thus, it is a useful alternative measure. The main effects of bot-type, $\hat{\beta} = 0.31$, 95% CI $[0.03, 0.60]$, $z = 2.14$, $p = .032$, and correct prediction, $\hat{\beta} = 0.43$, 95% CI $[0.23, 0.63]$, $z = 4.18$, $p < .001$, were significant for the measure of quality, too. This shows that participants were better at placing the piece when it was the one they predicted. This effect was again independent of the bot-type, $\hat{\beta} = 0.08$, 95% CI $[-0.20, 0.35]$, $z = 0.55$, $p = .584$. The main effect of trial was not significant, $\hat{\beta} = 0.00$, 95% CI $[0.00, 0.00]$, $z = -0.80$, $p = .423$, but the interaction of trial and bot type was significant $\hat{\beta} = 0.00$, 95% CI $[0.00, 0.00]$, $z = 4.31$, $p < .001$. These effects (Figure 5C) show that only participants collaborating with the smart bot got better at placing the piece over time, whereas participants who collaborated with a random bot did not improve.

Figure 5

Three Graphs illustrating the interactions of Bot and Trial with A Time to place, B Actions to place, and C Quality of the placement.



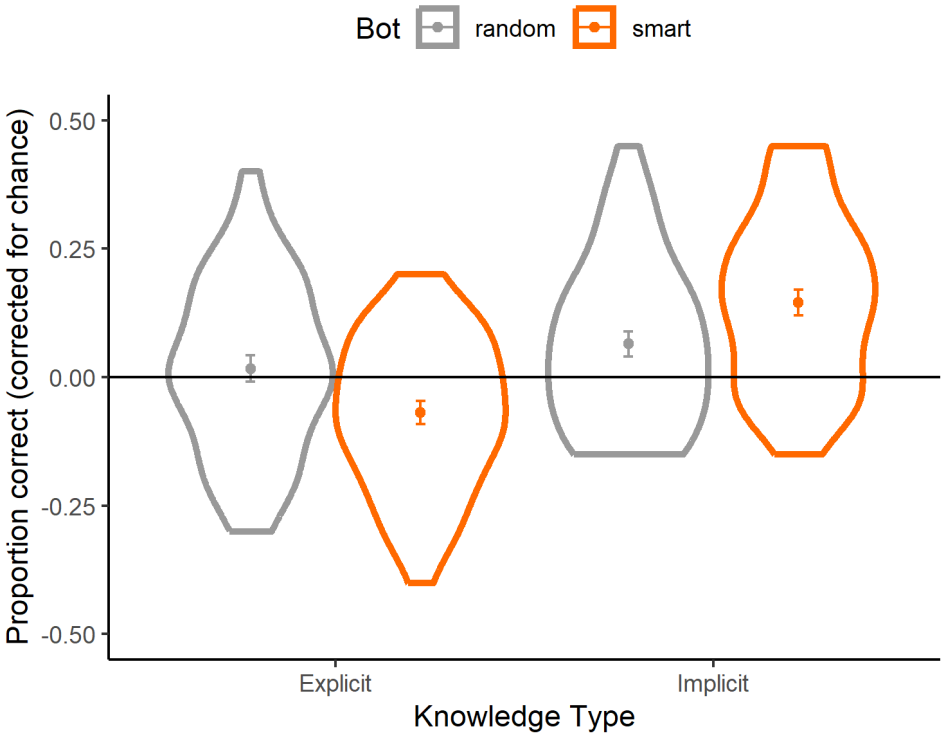
Knowledge tests

The analysis of the test data revealed a main effect for the bot type. Overall, participants with the smart bot were slightly worse on the tests: $\hat{\beta} = -0.09$, 95% CI $[-0.15, -0.02]$, $t(165.83) = -2.48$, $p = .014$. This effect was qualified by the interaction between bot type and test category: $\hat{\beta} = 0.17$, 95% CI $[0.08, 0.25]$, $t(84.54) = 3.65$, $p < .001$. The tests themselves did not have a significant main effect on the result, $\hat{\beta} = 0.05$, 95% CI $[-0.02, 0.11]$, $t(85.02) = 1.48$, $p = .142$. Post-hoc tests revealed that participants who interacted with the smart bot were significantly

better on the implicit tests - $\Delta M = 0.08$, $t(166) = 2.32$, $p = 0.022$ - but were significantly worse on the explicit test: $\Delta M = -0.09$, $t(166) = -2.48$, $p = 0.014$ (Figure 6).

Figure 6

Chance corrected probability to correctly answer a knowledge question as a function of knowledge type (explicit, implicit) and bot type (random, smart).

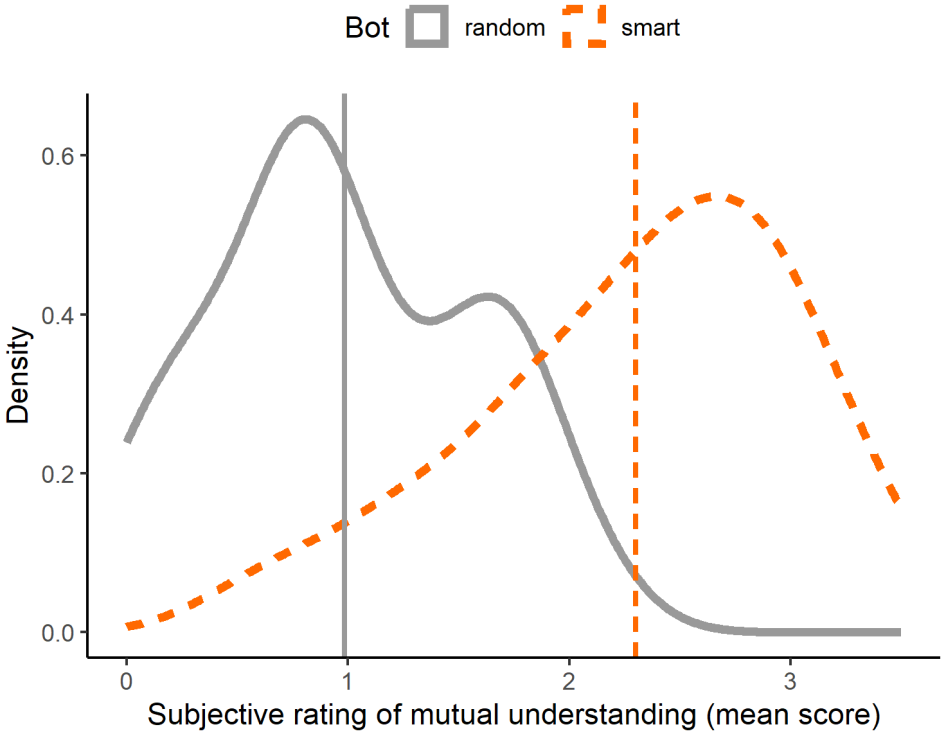


Subjective experience of mutual understanding

The t-test for the mutual understanding questionnaire showed a significantly higher rating of mutual understanding for the smart bot than for the random bot, $\Delta M = -1.32$, 95% CI $[-1.60, -1.04]$, $t(83.55) = -9.42$, $p < .001$ (see Figure 7 for a density plot).

Figure 7

Distributions of the average mutual understanding rating of the different bots. Vertical lines are the means of the respective distribution.



Discussion

The first study supports the hypothesis that repeatedly predicting the behavior of an agent in a novel task follows principles of implicit learning. Over 200 repetitions, the participants improved in predicting the actions of a goal-directed agent. This effect was not shown for a control group interacting with a random bot. The group interacting with the predictable bot had an advantage in the tests of implicit knowledge but a disadvantage for explicit questions. They were

able to predict the bot's behavior in novel situations, showing that they did learn the same general behavioral patterns of the bot, but they did not identify the correct explanations of the rules guiding the bot's behavior. These results suggest a more implicit rather than explicit learning process for the participants in the smart bot condition. However, this does not eliminate the possibility that participants would have made better explicit judgments if they had been prompted differently.

Correct predictions of the bot's behavior had a positive effect on participants' performance in their part of the cooperative task. Following a correct prediction, they were faster and more efficient in their moves and placed the piece more in accordance with the smart bots' judgment. These results occurred for both bot types. The interaction effects with time show that the advantage of the smart bot and a correct prediction shrink the more experience the participants gain.

An important question is left open by this study. It is not possible to tell what the determining factor is for the prediction of the agent's behavior, the task knowledge, or the ability to mentalize. The task is novel to the participants; while they might have experience in the traditional game, the alterations render them useless. So, the participants started the task with no task knowledge and no knowledge of their partner. Since the smart bot only picks well-fitting pieces, an improvement in predicting its actions can be indicative of both more knowledge of the task and a better mental image of the partner. The second study addresses this question by manipulating the task experience of participants.

Study 2

While Study 1 focused on the question of whether the prediction of social agents follows a logic similar to the prediction of non-social systems, it was not possible to explore the second overarching question of whether an increase in prediction performance is due to mentalizing due to increased task experience, or a combination of both. This aspect was specifically addressed in Study 2. Here, participants always interacted with a smart bot (i.e., a bot with a relatively high amount of task knowledge). The experimental manipulation was about the task experience of

participants. Before interacting with the bot, half of the participants individually played an unrelated game (Snake condition), whereas the other half of the participants individually played the Tetris version (Tetris condition). We hypothesized that prediction performance would improve over trials in the Snake condition. However, we had competing hypotheses about the predictive performance in the Tetris condition. If the game experience is the determining factor for being able to predict the bots' actions, we no longer expect an improvement in the predictions of the experienced group. If the ability to specifically understand the agent determines the prediction performance, we expect experience in the game to have no influence on the prediction performance. In this case, the experienced group will also improve the quality of their predictions due to an improvement in understanding the agent.

Methods

Participants

As preregistered (https://osf.io/8xemj/) and calculated through a power simulation, our target sample size was 300 participants. We collected a sample of 330 participants, of which 49 were unusable due to errors in data transfer. So, the following analyses were conducted on a sample size of 281. The average age of the sample was $M = 32.80$ ($SD = 11.13$) years old. The gender was roughly distributed equally, with 46.62% identifying as female, 52.31% as male, and 1.07% as other. Despite the exclusions based on missing data, participants were still distributed roughly equally among the experimental groups, with 142 (50.53%) participants in the Snake condition and 139 (49.47%) in the Tetris condition, respectively.

Game environment

The environment was the same as in Study 1. All participants played with the smart bot.

Material

The tests of explicit and implicit knowledge were improved over the first study. The number of explicit questions was extended to 15 questions, and the implicit questions were altered in their phrasing (Appendix A1). Instead of continuing to identify the piece the bot would

pick, one piece out of four was highlighted, and the participants were asked whether the bot would pick this item or not. This allows for a better comparison with the explicit questions, in which signals and distractors were also used. Additionally, we added a questionnaire consisting of 9 statements on possible assumptions about the agent's general behavior and evaluated the participants' certainty of the statements. Statements like "The bot acts goal-directed." "The bot tries to get a good score." and "The bot is unhelpful." were rated by the participants on a 5-point scale from 1: certainly false to 5: certainly true (Appendix A3). Prior to the experiment, the Cronbach's alpha of this questionnaire was $\alpha_{pre} = .55$. Following the experiment, the internal reliability was higher $\alpha_{post} = .83$.

Procedure

Study 2 was an online experiment, with the procedure being slightly altered from Study 1. We again had two groups of participants, but both played the game with the smart bot as a partner. In contrast to Study 1, we manipulated the experience the participants had with the game environment itself. One group of participants played 100 trials of our version of Tetris; the other group played a game of Snake during that time. Prior to and after their interaction with the smart bot, the participants answered the 9-item questionnaire on their assumptions of the agent's behavior. Then, in accordance with Study 1, they answered implicit and explicit questions and a questionnaire on their perceived understanding of and by the bot.

Design and Variables

A univariate between-subjects design with *experience* as the independent variable was realized. The participants either played the game alone before starting the game with a bot, or they played Snake, a game that is similar in complexity but holds no information about Tetris. As dependent variables, we again used the *accuracy of prediction* as an indicator of successful mentalizing. To quantify the task performance of the participant, we measured *time* and *number of actions* for a placement and the *quality of the placement* as judged by the smart bot. The quality of the placement is a binary variable that is one if the participant placed the piece as the smart bot

would have and zero for any different placement, respectively. The results from the implicit and explicit knowledge tests are simple mean scores. The same is true for the subjective rating of empathy.

Analyses

As preregistered (https://osf.io/8xemj), to identify a difference in the speed in which participants in the two conditions learn to predict an agent's actions, we conducted a logistic mixed regression with the logarithm of time and experimental condition as fixed effects and correctness of the response as the outcome. Random intercepts were calculated for each participant. To identify the cost of false predictions, we created linear mixed models with three outcomes: Time to place a piece, amount of actions to place a piece, and the quality of the placement. As independent variables, we used the outcome of the prior prediction and time. Random intercepts were calculated for each participant. We used a t-test to identify a difference in sensitivity for implicit vs. explicit knowledge. We used the control group (Snake) as the reference group for our analyses. The data and scripts used for analysis are made public at https://osf.io/cu53e/.

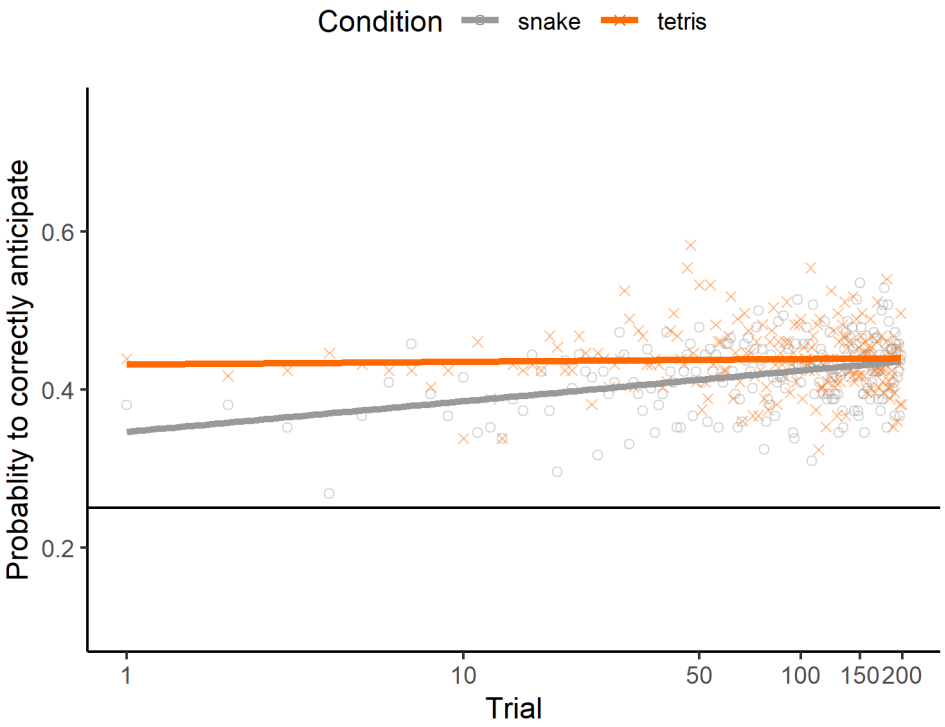
Results

Prediction

The analysis of the experimental data on the correctly anticipated trials by trial and experimental group shows several effects. The effects of the logarithm of trial, $\hat{\beta} = 0.06$, 95% CI $[0.03,0.09]$, $z = 4.67$, $p < .001$, and the condition, $\hat{\beta} = 0.37$, 95% CI $[0.20,0.54]$, $z = 4.18$, $p < .001$, are significant main effects. Notably, the interaction effect of group and trial negates the main effect of trial $\hat{\beta} = -0.07$, 95% CI $[-0.10, -0.03]$, $z = -3.67$, $p < .001$. This means that the control group showed a significant improvement over time as opposed to the experimental group, $\hat{\beta} = -0.01$, 95% CI $[-0.03,0.02]$, $z = -0.51$, $p = .611$ (Figure 8).

Figure 8

Learning curves of the participants with Tetris experience vs. task-irrelevant Snake experience.



Note: The constant black line at y = 0.25 represents the guessing probability.

Task Performance

We analyzed the participants' placement skill in three ways. The speed in time and in number of actions, and the quality of the placement are measured by the overlap with the bot placement.

Time to Place The time to place a given piece decreased linearly over the trials, $\hat{\beta} = -11.62$, 95% CI $[-13.15, -10.09]$, $t(55924.60) = -14.90$, $p < .001$. The participants were faster at

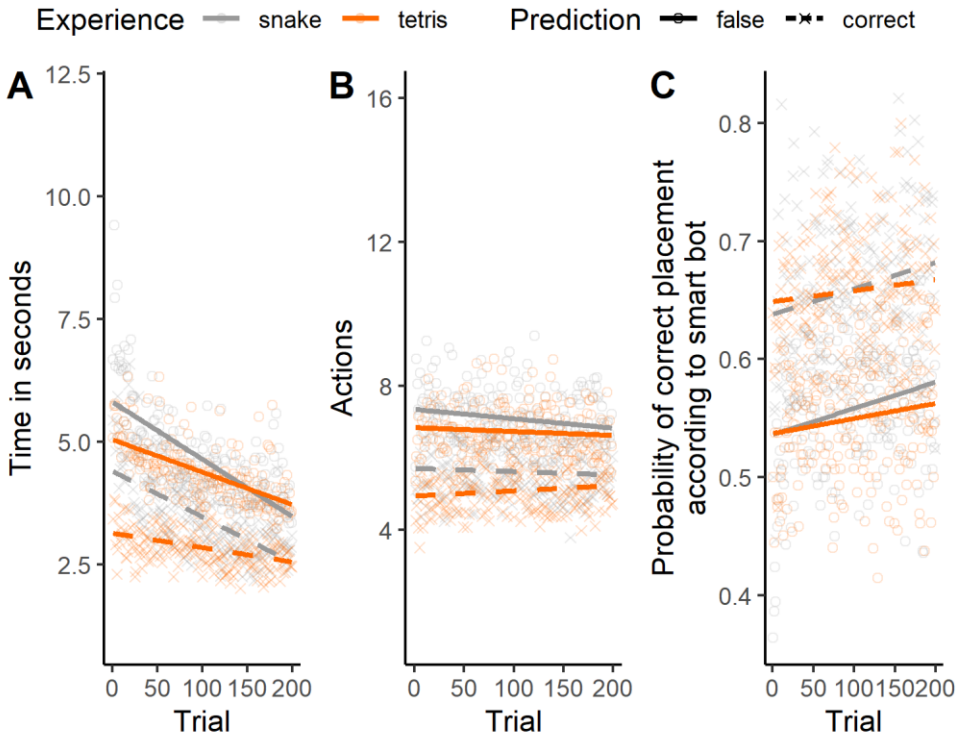
placing the piece after a correct prediction, $\hat{\beta} = -1{,}447.53$, 95% CI $[-1{,}724.79, -1{,}170.27]$, $t(55961.37) = -10.23$, $p < .001$, and also when they were more experienced at the task, $\hat{\beta} = -807.91$, 95% CI $[-1{,}321.55, -294.26]$, $t(445.79) = -3.08$, $p = .002$. The interaction of prediction and experience was significant, $\hat{\beta} = -430.39$, 95% CI $[-822.39, -38.38]$, $t(55959.81) = -2.15$, $p = .031$, as was the interaction of trial and experience, $\hat{\beta} = 4.66$, 95% CI $[2.46, 6.85]$, $t(55926.09) = 4.16$, $p < .001$. Again, this pattern indicates a convergence over time. The interactions of trial and correct prediction, $\hat{\beta} = 1.83$, 95% CI $[-0.55, 4.21]$, $t(55940.08) = 1.51$, $p = .132$, and the three-way-interaction, $\hat{\beta} = 1.93$, 95% CI $[-1.44, 5.30]$, $t(55942.98) = 1.12$, $p = .262$, were not significant (Figure 9A).

Number of Actions to Place. As in Study 1, the number of actions was similar to the time the participants took to place a piece. It also decreased over time, $\hat{\beta} = 0.00$, 95% CI $[0.00, 0.00]$, $t(55924.59) = -2.82$, $p = .005$. The participants were also more efficient at placing the piece after a correct prediction, $\hat{\beta} = -1.72$, 95% CI $[-1.99, -1.44]$, $t(55961.38) = -12.19$, $p < .001$, but the experience did not have a significant effect, $\hat{\beta} = -0.49$, 95% CI $[-1.00, 0.02]$, $t(445.82) = -1.88$, $p = .060$, nor did the interactions (Figure 9B).

Placement Quality. The overlap of the participants' placement with the bot's intended placement was greater after correct predictions, $\hat{\beta} = 0.37$, 95% CI $[0.26, 0.47]$, $z = 6.92$, $p < .001$, and got better over time, $\hat{\beta} = 0.00$, 95% CI $[0.00, 0.00]$, $z = 4.60$, $p < .001$. Again, the effect of experience $\hat{\beta} = 0.04$, 95% CI $[-0.17, 0.26]$, $z = 0.40$, $p = .689$, and the interaction effects were not significant (Figure 9C).

Figure 9

Three Graphs illustrating the interactions of Expertise (Snake vs. Tetris) and Trial with A Time to place, B Actions to place and C Quality of the placement.



Implicit and Explicit Sensitivity

The analysis of the test data showed no significant advantage for implicit sensitivity over explicit sensitivity, $\Delta M = -0.09$, 95% CI $[-\infty, 0.02]$, $t(534.56) = -1.28$, $p = .100$.

Exploratory Analysis

Assumptions

We see the certainty in the assumptions about the bot changed following the interaction with it, $\hat{\beta} = -0.53$, 95% CI $[-0.59, -0.48]$, $t(4777.46) = -20.54$, $p < .001$. However, the certainty in the participant's assumptions did not have a meaningful impact on the ability to predict the bot's actions, $\hat{\beta} = 0.01$, 95% CI $[-0.08, 0.10]$, $z = 0.26$, $p = .798$.

Discussion

Study 2 replicated the learning effect of an inexperienced group interacting with a smart bot from Study 1. Predictive performance increased over time when the task experience was low. In contrast, participants with experience started with a high prediction rate and did not improve over time. These results indicate that these participants did not mentalize but rather tried to predict the best-fitting piece. Some minimal mentalizing might have happened early in the interaction when the participants had to identify their partner as able and cooperative in the game.

The positive effect of a correct prediction could be replicated. Following a correct prediction, participants were again faster and better at placing the piece, and they also improved over time. Task experience only had a significant effect on the time the participants took to place the piece, not on the efficiency of the movements or on the quality of placements. The interaction effects on time show that a false prediction is more costly for an experienced player and that inexperienced players catch up in speed to the more experienced players.

In Study 1, we presented the benefit of implicit testing of the participants' understanding of the bot. These results were not replicated in Study 2. There are several possible explanations for this. The tests in Study 2 were more conservative in a few ways. For this test, there was no random control group as there was in Study 1; both groups interacted with the same bot. The tests themselves were altered from Study 1; the implicit test was changed so it fits a signal detection analysis. However, this also means that the participants had to make a bigger transfer, which has been shown to be difficult with implicit learning. The explicit test was improved by adding more items. This should only benefit the reliability of the test and not have an impact on the outcome.

General Discussion

In two studies using a novel experiment task, we examined people's ability to predict the decisions of an artificial agent (Kaufman et al., 2010). We found that participants improve in predicting an agent's decisions over time, but only if the agent is acting goal-directed and the participant is new to the task. The implicitness of the learned model is unclear. In the first study, we found some indication of it but failed to replicate these results in a methodologically improved second study. Further, we found that following a correct prediction, the participants were able to place the piece better and more efficiently. Finally, we found a benefit of the smart bot in the participants' subjective impressions of mutual understanding.

The experiment is useful as a codependent, repeatable task. The game environment and continuous development over time give it a high external validity. The results can likely be generalized to other cooperative scenarios. Another benefit is the quick repetition of trials, which leads to a fine time resolution of the measured variables. Also, the motivation of the participants to play games is likely higher than in comparable non-game studies. On the downside, the measures have to be chosen carefully in order to not break the flow of the study. Additionally, the measurement can cause confusion for the participants. The participant has to allocate cognitive resources between the given task and the measurement. This is a two-edged blade as it hides the purpose of the study from the participant but also introduces a source of noise.

Using this task, we uncovered two conditions that must be met for a learning process towards the behavior of an agent to happen at the time of cooperation in the game. First, perhaps trivially, the behavior of the agent has to be predictable. A randomly acting agent cannot be predicted; thus, people cannot improve in predicting its actions. Second, the observer must be inexperienced at the task itself. A person experienced at the task no longer improves in identifying fitting pieces and apparently also does not improve in identifying the specific behavioral rules the agent acts on.

If mentalizing required forming a picture of the strategies, beliefs, and desires of an agent, repeatedly seeing how an agent performs should have led to an increase in prediction performance, even for experienced players. The fact that experienced observers did not improve noticeably speaks against mentalizing. One can conclude that participants in the Tetris condition of Study 2 solely relied on their task experience and made their predictions as if the agent used a strategy close enough to their own. One important conclusion that can be drawn from this is that making good predictions does not necessarily require mentalizing, and it can also be accomplished by having task experience.

For experienced observers, it seems to suffice to rely on their experience to make reasonably accurate predictions. However, for inexperienced observers, this is not necessarily true. They improve at predicting the agent's actions. This improvement would be plausible given either only an improved understanding of the agent, of the task, or both. For instance, the anchoring-and-adjustment hypothesis suggests that the extent of mentalizing depends on the similarity between the observer and agent. Applied to our scenario, one might argue that experienced participants of Study 2 did not have to mentalize (high similarity), whereas inexperienced participants had to mentalize (low similarity).

The results reveal an interesting limitation of people's ability to take on others' perspectives: In many cases, predicting an agent's decisions and having task experience are inextricably intertwined. Someone with high experience can make better predictions about agent behavior (even without mentalizing) simply by being able to produce the most plausible behavior in a given situation. In contrast, someone with little experience will only be able to predict an experienced agent's behavior by gaining experience. It might be that some task experience is a necessary condition for a good prediction of behavior.

Future studies could further elucidate this principle by systematically varying observer experience and agent experience in a single design. We would predict a limiting effect from the observer's experience, which states that properly assessing an agent is largely dependent on the

observer's experience. For instance, an accurate assessment of agent expertise is only possible for an observer that has the same or higher expertise. Conversely, without getting feedback, an observer will not be able to accurately assess the expertise level of an agent with objectively higher expertise. In other words, observer expertise acts as a threshold for making accurate assessments about agents. It is possible to accurately assess an agent with lower expertise, but it is not possible to accurately assess an agent with higher expertise unless the situation affords feedback (which makes it possible to learn from the agent with higher expertise). One should note that the predicted experience contingency effect could be conceived as an interpersonal equivalent to the Dunning-Kruger effect (Dunning, 2011): People don't know what they don't know. Those who have low task knowledge/experience are least capable of properly assessing their own (Dunning-Kruger) as well as others' (experience limit) performance.

Our studies have shown that participants improve their predictions over time if their task experience is low and the agent is predictable. However, the results on the implicitness of the learned material are inconclusive. In Study 1, we see a clear interaction between the type of test and the type of agent. This indicates a more implicit knowledge of the smart agent. Interestingly, the participants also felt a better understanding of this agent even though the results of the explicit knowledge test were worse than those of the participants with the random control agent. These results could not be replicated in the second study. The test in the second study was more conservative in two senses. First, the tests were improved, the explicit test was longer to cover more possible explanations, and the implicit test was altered to better fit a signal detection task. While this allows a better comparison between the tests, it also requires more transfer by the participant, which is known to be difficult in implicit tasks (Dienes & Berry, 1997; Seger, 1994). Secondly, the bot was no longer varied, so the effect had to show without a control group. These results are representative of the ongoing discussion on the amount of automaticity in mentalizing processes (Kulke et al., 2018; Schneider et al., 2017). Looking only at the results of Study 1, one

could come to the conclusion that the participants implicitly mentalize about the agent. However, Study 2 raises doubts about this interpretation.

However, we consistently found a prediction effect. The participants performed better on their task when the prediction they made for the bot's choice was correct. On average, they were faster, more efficient in their movement, and placed the pieces better. We identified two plausible explanations for this phenomenon. It is plausible that the participants invested mental effort into the piece they predicted. They likely already had a position in mind where they intended to place the piece. But this effort was then wasted when they received a different piece and had to redo the search for a good placement. Further, mental capacities might be bound by the processing of the wrong prediction and the integration of the newly gained information into the knowledge about the agent. Both processes are not mutually exclusive and can happen simultaneously. The studies are not equipped to make claims about the cognitive processes leading to these results, and further investigations in this direction might be fruitful.

The role of feedback is especially important in these studies. In the case of the smart bot, the feedback on a false prediction holds much information for the observer. On the one hand, it gives information about the specific behavior of the agent. On the other hand, since the agent is performing well, it gives information about how to perform well on the task. Further examination of the specific effects of feedback in a behavior prediction task over time would certainly be interesting. However, it would also be difficult to pinpoint because some form of feedback is always inherent to the behavior of the agent once its behavior is observable.

The two studies presented here have shown great promise, but there are some limitations that need to be addressed. Mentalizing processes could be facilitated by the perception of a social presence (Morgan et al., 2022). Thus, playing online with an artificial agent could hinder the participants from engaging in mentalizing processes. Finally, the measurements might be imperfect. The quality of placement has no strictly objective measure, so we used a measure that was available. As already mentioned, the measure of explicit knowledge is imperfect. Any

amount of statements about the agent's behavior does not capture the full space of possible explanations. Thus, an argument can always be made that the participants might have gained explicit knowledge and would have been able to express it if they were prompted differently (Shanks & St. John, 1994). The results are not limited by ethnic or national constraints. The experiments were conducted online; thus, the sample is skewed towards people with internet access and who are competent in internet use.

To conclude, in the two presented studies, we investigated the ability of people to anticipate an agent's behavior in a codependent cooperative setting. Unexpectedly, mentalizing processes seem to not play an important role in these situations. The predictive ability seems to hinge on the ability to make judgments on the task itself and, following that, to predict that the bot will pick the best available piece that one could think of in a specific situation. While the results from the two studies suggest such conclusions, they were not designed to fully manipulate the information an observer and an agent have about a set of tasks. Further research in this direction is necessary and might unravel natural limitations to the understanding of others' behavior.

Chapter 3 – Expertise Judgments

The third chapter is a collaborative manuscript of Fritz Becker (first author), PD Dr. Jürgen Buder (second author), and Prof. Dr. Markus Huff (third author). It has the title “Greatness recognizes greatness: When only Experts make Accurate Expertise Judgments of Others.” At the time of the thesis’s release, the manuscript is under review at the Journal of Experimental Psychology: Human Perception and Performance.

Abstract

Knowing what someone else knows is important in shared environments. However, the knowledge of others is not immediately perceptible. Luckily, an agent's behavior within a task leaves clues for an observer to estimate their expertise. A task expert will make the best decisions to reach their goal, while a novice will make decisions seemingly at random. An expert observer should be able to evaluate the decisions of an agent by comparing the agent's decision to the ideal one. On the other hand, a novice observer does not know the ideal decision and, thus, cannot compare the agent's decision to it.

We test the hypothesis that with increasing expertise, people can more accurately assess the expertise of others. The results of the two studies we present are consistent with this hypothesis. The first experimental study showed the expected interaction of agent and judge expertise in a task with limited external validity. In the second study, the results of the first study were replicated in a more generally applicable task. We discuss possible moderators of this effect, such as conflicting expertise cues and additional metacognitive information communicated by the agent. Moreover, we outline the consequences of expertise imbalance for identifying imposters in expert fields.

Introduction

Imagine watching a chess game of two anonymous players with unknown ratings. How do you come to a conclusion about the capabilities of the players (Reynolds, 1992)? You will likely try to find the best move yourself and compare it to the moves that the players make. If you are a decent player who can reliably find the best move in the position and see the players deviating from it, the severity and frequency of the errors give you a lot of information about the players' abilities. However, if the players are masters, you might quickly find the game in a confusing position where you cannot find the best play and consequently have difficulties evaluating the moves of the players. It is then difficult to say how much better the players are than you.

In an environment shared with other agents, estimating somebody else's expertise at a task relevant to oneself is an important skill (Kumar et al., 2023). However, expertise is not immediately perceptible, and people need to interpret cues of expertise to come to an estimation. The cues for an agent's expertise can be difficult to process for observers with little task knowledge. In this paper, we explore the influence of observer and agent expertise on the observer's judgment of the agent's expertise.

Expertise

Expertise is the specialist knowledge, cognitive skills, and abilities an agent needs to perform well at a specific task (Collins & Evans, 2009; M. Mayer et al., 2023). Experts deviate less from optimal decision-making than novices. Further, agents who are identified as experts are seen as more credible and trustworthy (Rieh & Danielson, 2007).

Expertise self-evaluation

Existing research is focused on identifying the expertise cues people use to identify their own expertise and to compare their expertise to the general public (Koriat, 2008; Tullis, 2018). People can use metacognitive cues for the evaulation of their expertise, such as a feeling of knowing and confidence in one's decisions to be optimal. These cues use interoception; they are not available in the estimation of another person's expertise unless verbally or non-verbally

communicated. The discussion on the metacognitive insight novices have, as opposed to experts, is ongoing (Kruger & Dunning, 1999; McIntosh et al., 2019; Nuhfer et al., 2017).

Expertise evaluation of others

When evaluating the expertise of others, knowledge can be a curse (Birch, 2005; Tullis & Feder, 2022). If the judge has no information about the agent, estimations are anchored in one's own knowledge. However, if given appropriate information observers can track the beliefs and desires of agents. This "theory of mind" is believed to be an ablility of only highly developeted social agents (Kosinski, 2024; Penn & Povinelli, 2007). In the following paragraphs, we discuss the information that is necessary for an accurate expertise estimation.

Curse of Knowledge

Previous work specifically on the expertise estimation of others has generally focused on the "curse of knowledge" (Birch, 2005; Tullis & Feder, 2022), which suggests that under conditions of cue sparsity, perspective-taking is especially difficult for experts as they lack information on the agent necessary to take their perspective. For this reason, the expert's expertise judgment of others is biased towards their own knowledge (Bromme et al., 2001). However, when information on an agent's attributes or task behavior is missing, it is practically impossible for people to estimate the expertise of others accurately. Since they do not have insight into the agent's cognition, the judges are forced to use their own experience as a proxy to come up with plausible estimates (Tullis, 2018; Tullis & Feder, 2022; Tullis & Fraundorf, 2017). By observing the agent that is to be judged, the observers can intergrate the information they have on the task and the information they get from the agent's behavior in order to make more accurate and less biased expertise judgements.

Theory of mind

Even when given cues, estimating the knowledge of others is not an easy task. First, the observer has to be aware that others might hold information different from ones own. Then the observer needs to keep track of the information the agent holds and the differences to their own

knowledge (Deschrijver & Palmer, 2020). Theory of mind is an agent's ability to infer the mental states of others from their behavior (Baron-Cohen et al., 1985; Wellman et al., 2001). Using their theory of mind, observers are able to infer the hidden knowledge of agents by observing their behavior (Aboody et al., 2021; Baker et al., 2017; Berke & Jara-Ettinger, 2021). This even allows the observers to make accurate predictions about the future behavior of the agents. However, in these experiments, the observers usually have complete information about the task. Only the information on the agent's knowledge is hidden and needs to be inferred. A popular example is the false belief task, where some fact about a situation is changed in the absence of the agent, leading them to have a false belief about the situation. Given the developmental ability, the participant knows about the change and can infer what the agent will do (Bloom & German, 2000; Huemer et al., 2023; Kulke et al., 2018). Experiments for adults have more complex tasks, such as inferring the food preferences of an agent based on its information search behavior (Baker et al., 2017), the interpretation of the agent pausing in a maze due to a distraction vs. actively thinking (Berke & Jara-Ettinger, 2021), or the inference of an agent's knowledge based on the agent's choice of a search task (Aboody et al., 2021).

In these scenarios, the observer has complete knowledge of the agent's task, while the agent's task knowledge can be limited. The observer is the task expert, and the agent knows less. In the present studies, we show this is a special situation in which an accurate mentalization of the agent's decision-making process is possible. Conversely, when the agent has more task knowledge than the observer, understanding the agent's decisions is more complicated, bordering on impossible. It seems obvious that people cannot understand decisions based on hidden information. However, this logical statement has not been shown experimentally, even though it might have severe consequences for our understanding of expertise dynamics in social perception.

Expertise cues

To correctly evaluate an agent's expertise, the observer needs cues, otherwise the estimation is likely biased towards their self-estimation (Tullis & Feder, 2022). However, not all cues are equally valid and interpretable. Building on the 3S model of information trust, we suggest three categories of cues for expertise estimation based on agent decisions (Lucassen & Schraagen, 2011). In their model of information trust, Lucassen and Schraagen (2011) identify three categories of cues people use to come to a validity judgment of the textual information. The categories are source features, surface features, and semantic features. The source features are features that do not refer to aspects of a text itself but rather to attributes of the source, such as the author's or the website's credibility. Surface features of a text refer to the style and the language used in the text. It can also refer to the use of images or references to other literature. On the other hand, semantic features represent the content of the information. Basing trust judgments on semantic features is the most effortful and requires domain expertise of the judge (Lucassen & Schraagen, 2011).

Lucassen and Schraagen (2011) base their categorization on textual information. For the purpose of the paper, we similarly categorize the information from an agent's observable behavior. The information source is the agent, and as such, the attributes (source features) of the agent can give cues to their expertise. Examples of surface features of agent behavior are the confidence and the results of the agent's actions. The semantic features are the agent's decisions themselves. Like semantic features of a text, the decisions give information to task experts who can evaluate the quality of a decision based on their own task knowledge.

The agent's attributes and the surface information of decisions can be interpreted by observers, regardless of their expertise. The information from an agent's attributes (i.e., source features) can be unreliable or even invalid. For example, an academic title, age, height, and attractiveness are cues that judges regularly use to inform their expertise judgments (Forgas & Laham, 2016; Nisbett & Wilson, 1977). The reliability and validity of these cues varies and

depends on the task. The surface features of behavior might also be invalid or even manipulated. For example, a chess student who always plays confidently into the same trap in the opening appears confident, but it is no cue for expertise. On the other hand, the decisions of an agent trying their best at a task gives valid insight into their expertise (i.e., semantic features). Due to the varying validity of agent attributes and surface information we focus on expertise evaluation based on the semantic features of a decision.

Present studies

In the present studies, we focus on expertise estimation based on decision cues while keeping information on the agent's attributes and surface information uninformative or constant. As with semantic features of written text, the observer needs expertise to interpret the agent's decisions (Bower et al., 2024; Lucassen & Schraagen, 2011). Expert observers can compare the agent's decision to what they think would be the best decision in the current situation to judge the severity of the agent's error. This estimation will be closer to the true error of the agent than that of a novice. Per definition, a novice does not know the best decision and has no or an inaccurate anchor to base their estimation.

Existing results on theory of mind are based on observers having full information about the agent's task. We expand on that by varying the expertise of the observer and of the agent. In the first study, we experimentally manipulated the task knowledge of participants by showing them different subsets of rules using artificial grammar. In the second study, we measured participants' knowledge through a trivia questionnaire. We then showed the participants the products of agents with differing expertise and asked them to estimate the number of rules the agents knew and the number of answers the agents answered correctly, respectively.

We claim that judges can accurately mentalize and predict the expertise of those agents who know less about the task. Conversely, it is not possible for judges to accurately assess the degree of expertise of agents with superior knowledge. Due to the relational nature of the assessment process, the accuracy of expertise judgments rests on the judge's skills (with higher

expertise leading to higher accuracy) and the agent's skills (with higher agent expertise leading to lower accuracy).

Study 1

In the first study, participants were familiarized with a subset of an artificial grammar's rules (Reber, 1967). Using this knowledge, they were tasked with estimating the expertise of agents with varying knowledge about this grammar. The agents produced We experimentally tested two hypotheses. In the preregistration, we phrased them as follows:

H1: If an observer knows more about a task compared to an agent, evaluation of an agent's skill in performing the task will be more accurate.

H1a: On trials where the observer knows more about the task than the agent, the error will be lower.

H1b: The slope of error increase for each rule the agent knows is flatter the more rules the observer knows (an interaction effect of observer and agent knowledge on estimation error).

H2: If an observer knows more about a task compared to an agent, they are more confident in their evaluation of the agent's skill.

Methods

The methods, hypotheses, and analyses of Study 1 were preregistered under https://aspredicted.org/S6G_15V. The preregistration, materials, data, and analyses are openly accessible under https://osf.io/gbt64/.

Simulation

To anticipate the size of the effects, the necessary sample size, and the resulting power of the analyses, we simulated it. The base rationale we used for the simulation was that if the judge's knowledge k_j was more or equally high as the agent's knowledge a_i, the estimation was exact.

$$d(k_j, a_i) = \begin{cases} 1, & k_j < a_i \\ 0, & k_j \geq a_i \end{cases}$$

If the judge knows less, they must guess between the highest and the lowest possible rules known by the agent. We informed the participants that the highest amount of rules and, consequently, the highest possible expertise was $u = 6$. The lower bound is the number of rules known to the judge k_j. For the simulation, we set the guessed expertise to the midpoint between the upper and the lower bound $g(k_j, u)$.

$$g(k_j, u) = \frac{(k_j + u)}{2}$$

The agent knowledge estimated by the judge $\hat{a}_i$ was simulated using the following equation:

$$\hat{a}_i = d(k_j, a_i) \cdot g(k_j, u) + \big(1 - d(k_j, a_i)\big) \cdot a_i + \gamma_j + \sigma_{i,j}$$

with $\gamma_j \sim N(0, \sigma_a^2)$, $\sigma_{i,j} \sim N(0, \sigma_b^2)$.

In order to anticipate the effect and sample size we can expect from this process, we use the linear mixed model we use to test Hypothesis 1b because it is likely the least powerful test.

$$Y = \beta_0 + \beta_1 k_j + \beta_2 a_i + \beta_3 k_j a_i + \gamma_j + \sigma_{i,j}$$

again with $\gamma_j \sim N(0, \sigma_a^2)$, $\sigma_{i,j} \sim N(0, \sigma_b^2)$.

Using this simulation, we found that 200 participants are sufficient to find a significant interaction effect of agent and judge knowledge at power $1 - \beta > 0.8$.

Participants

We collected data from 240 participants via the Prolific panel. The participants were paid 3£ for an expected 20-minute experiment. Three participants withdrew their data at the end of the study, so the final sample consisted of 237 participants. The average age of the participants was $M_{age} = 30.46$ ($SD_{\text{age}} = 10.22$). Their self-identified genders were $n_{male} = 157$, $n_{female} = 77$, $n_{non\text{-}binary} = 2$, one preferred not to answer.

Material

The material was created using a simple artificial grammar (Figure 10A). The grammar was built with the intention of being able to fully describe it with six verbal rules (Table 1). We phrased the six rules on the basis of the grammar as unambiguously as possible. We ordered the rules so that each new rule maximally diminishes the room of possible strings. The order of the rules was important, as a participant who is shown rule 3 will also always be shown rules 1 and 2. We decided on this in order to create a strictly increasing expertise. The strings were created by building the grammar's graph and traversing its edges. Since the graph is looping, and we wanted the strings to terminate, we limited the graph traversal so that each node can only be visited three times at maximum. The strings following fewer rules were created by adding more edges (less restrictions) to the graph (Figure 10B). With the added freedom for graph traversal, strings can be created, violating the higher rules while still adhering to the lower ones. The rules are ordered in a way that maximizes the information of each new rule (Appendix B1).

Figure 10

Artificial Grammar Graph for A all six Rules and B rules 1-4

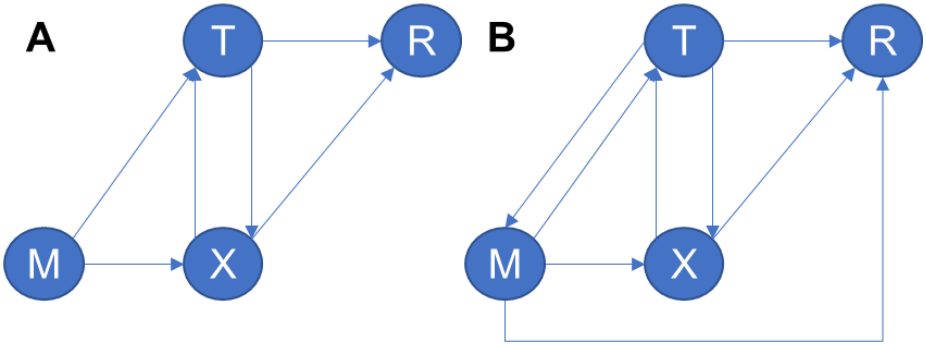


Table 1

Rules and Example Strings

Rule	Example
0.	IEYICCWP
1. A correct string may only consist of the following characters: M, R, T, X	RXMTXMR
2. A correct string always starts with M	MTTRTRMRMX
3. R can only and must always be at the end of a correct string	MXTMXTTMR
4. After an X, only R or T may follow	MMXTXTTR
5. After a T, only R or X may follow	MMMR
6. After an M, only T or X may follow	MXTXTXR

Procedure

Following the information, consent, and demographic forms, participants were presented with a set of rules. Depending on the condition, participants were shown 1, 3, or 5 of a list of 6 rules. The rules hidden from the participant were shown as blacked-out to highlight the total amount. In order to minimize memory effects, the uncovered and blacked-out rules were present at all stages of the experiment where the rules were relevant. Participants then familiarized themselves with their rules by identifying 14 strings as correct or false according to their available ruleset. Further, they were asked to create five strings in accordance with their rules. After familiarization, participants were instructed on the experiment. During the experiment, participants judged four agents with different agent expertise based on the sequential presentation of four strings that each agent had generated. For each presented string, the participants were asked how many rules this agent knows. After one agent's set of 4 strings, the participant was requested to give a final judgment of the rules this agent knows, together with a metacognitive estimation about their certainty for their judgment. Afterward, an interim page was shown to mark the change from one agent to the next. Following the experiment, we tested whether the participants learned some of the rules that were hidden from them. For these tests, the rules were no longer shown on the screen. First, they were instructed to identify the six true rules among 12.

Afterward, they were again given 14 strings, 2 for each rule, and were instructed to identify the ones that were true according to all rules. The study was concluded with a debriefing and the participant's last opportunity to retract their data.

Design and Variables

We varied the expertise of the agent and the judge by varying the number of rules presented to the participant (judge expertise) between subjects and the number of rules the agents used to create strings (agent expertise) within subjects. The d variables are the participant's estimation of the agent rules and their confidence in the estimation. The estimation error e was calculated as the absolute difference between the agent judgment. $\hat{c}$ and the agent's actual amount of correct answers c $e = |\hat{c} - c|$.Our experiment has a 3 x 4 mixed design. We varied judge expertise, the between-subjects factor, on 3 levels: 1, 3, and 5. Agent expertise, the within-subjects factor, varied on 4 levels: 0, 2, 4, and 6.

Analyses

The analyses were conducted using R (R Core Team, 2023). The models were fitted using lme4 (Bates et al., 2015). The degrees of freedom were estimated using the Satterwhite method implemented in the lmerTest package (Kuznetsova et al., 2017).

As preregistered, the main hypothesis was tested using two tests. First, we use a t-test testing the difference between trials with a knowledge advantage for the judge and trials with the agent having an advantage (H1a). Second, we use a linear mixed model with estimation error as dependend variable, judge, agent rules, and their interaction as fixed effects and a random intercept for participant and agent (H1b). The second hypothesis was tested with the same model specification with the participants confidence as dependent variable (H2).

Results

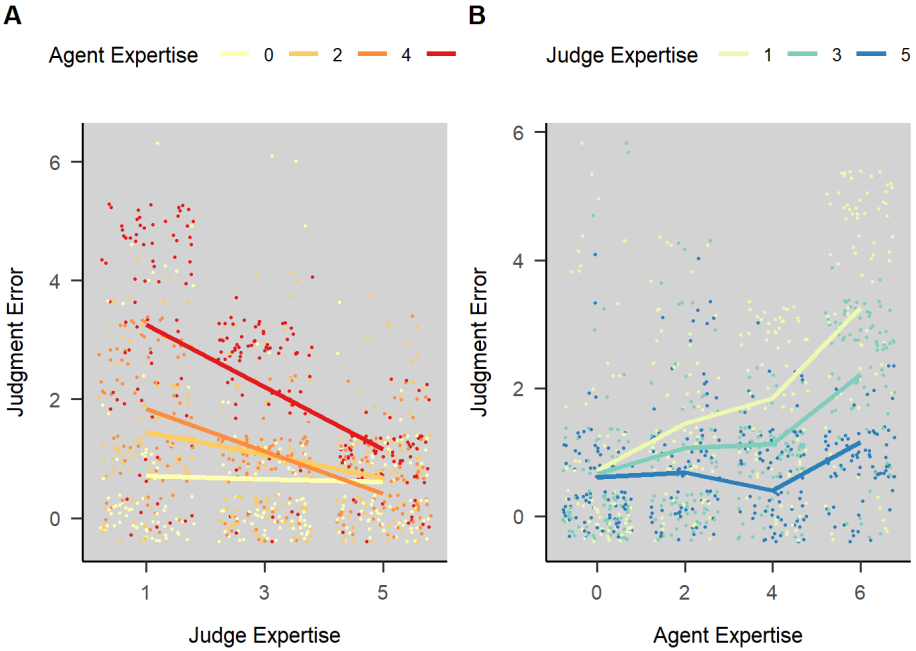
We analyzed the first hypothesis using two methods. The t-test comparison of trials where the agent knows more than the judge compared to trials where the judge knows more shows a significant advantage of knowledge for the accuracy of the judgments (mean error

difference = -1.02, 95% CI [-1.18, -0.87], $t(891.22) = -12.68$, $p < .001$; Cohen's $d = -0.83$, 95% CI [-0.96, -0.69]). This result is in line with H1a.

We tested the second part using a mixed linear model. We expected an interaction effect of the agent's expertise and the judge's expertise on the expertise estimation error. The expected interaction effect is statistically significant and negative ($B = -0.08$, 95% CI [-0.10, -0.06], $t(941) = -8.48$, $p < .001$; $\beta = -0.10$, 95% CI [-0.12, -0.08]). The effect is illustrated in Figure 11. The main effect of agent expertise is significant ($B = 0.48$, 95% CI [0.32, 0.64], $t(941) = 5.96$, $p < .001$; $\beta = 0.17$, 95% CI [0.06, 0.28]), while the main effect of judge is not ($B = -0.03$, 95% CI [-0.10, 0.05], $t(941) = -0.73$, $p = 0.467$; $\beta = -0.03$, 95% CI [-0.12, 0.05]). This result confirms H1b. Based on the results of Study 1, Hypothesis 1 should be accepted. The result shows the relevance of task knowledge for accurately judging an agents' expertise. Expertise below one's level of knowledge can be assessed accurately, on the other hand the expertise of agents above ones level has to be guessed. Note that the theoretically possible error is lower for intermediate agent expertise. Randomly guessing would lead to lower average errors for agent expertise of 2 and 4 and higher errors for 0 and 6 (Nuhfer et al., 2017).

Figure 11

Expertise Interaction Effect



From Hypothesis 1, we expect higher accuracy of expert judges that would be expressed in smaller deviations from the correct answer, resulting in a smaller standard deviation of the error. To further test the claim, we calculated an unregistered, exploratory linear model using the standard deviation of the judgment errors as the dependent variable and the judge's expertise as an independent variable. The effect of judge expertise is significant and negative (B = -0.32, 95% CI [-0.37, -0.28], $t(235)$ = -13.68, $p < .001$; β = -0.67, 95% CI [-0.76, -0.57]), giving further, albeit unregistered proof of the effect's robustness.

We further explored the interaction of judge and agent expertise. Figure 12 illustrates the individual distributions of judgment for all combinations of judge and agent expertise. Note that

the peak is equal to the agent expertise in cases of higher judge expertise and at the judge expertise in cases of lower judge expertise, indicating an anchoring-and-adjustment strategy (Epley et al., 2004). Anchoring and adjustment is a hypothesized process of perspective-taking, in which one mentally anchors in one's own experience and then adjusts stepwise towards an expected experience of somebody else. This generally leads to a bias towards one's own experience, an egocentric bias. With regard to these results, it seems that adjustment towards higher expertise is impossible since the judges lack information about the higher-order rules.

Figure 12

Expertise Judgment Distributions by Agent and Judge Expertise

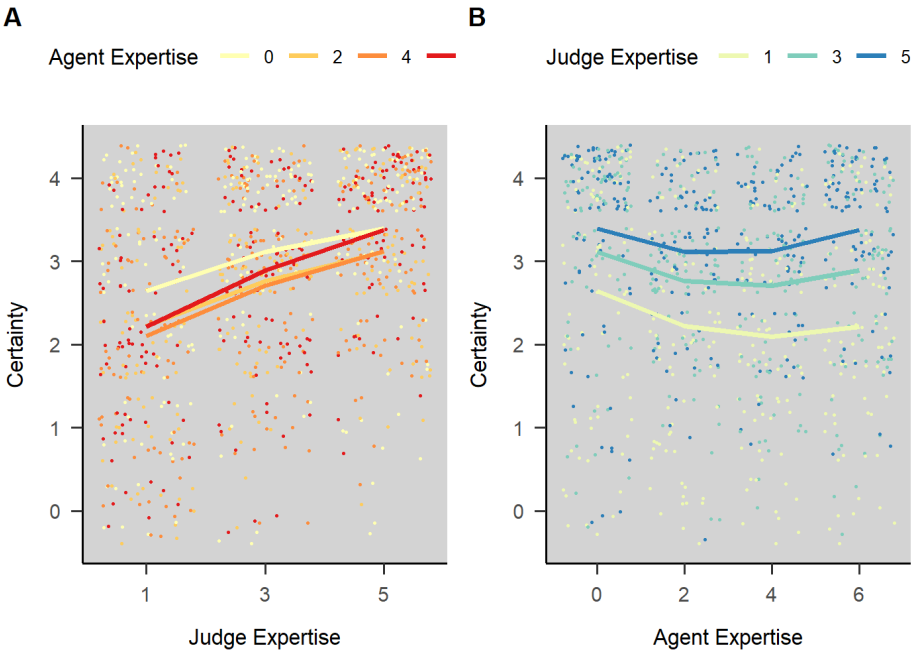
Hypothesis 2 was directed at the self-estimated confidence of the judges. We expect a higher confidence in situations where the judge knows more rules than the agent. The t-test testing the difference of certainty by whether the judges knew more than the agents (mean certainty in lower judge knowledge = 2.58, mean certainty in higher judge knowledge = 3.04) suggests a small effect (difference = 0.46, 95% CI [0.31, 0.60], $t(931.48) = 6.02$, $p < .001$; Cohen's $d = 0.39$, 95% CI [0.26, 0.52]). The result confirms hypothesis 2.

To gain a further understanding of the effect, we fitted a linear mixed model predicting the certainty from the rules available to the agent, the judge, the interaction, and random intercepts for the judge and the agent (Figure 13). The model's total explanatory power is substantial (conditional $R^2 = 0.49$), and the part related to the fixed effects alone (marginal R^2) is 0.11. The effect of agent rules is non-significant and negative ($B = -0.09$, 95% CI [-0.19, 0.01], $t(941) = -1.71$, $p = 0.088$; $\beta = -0.03$, 95% CI [-0.11, 0.05]). The effect of obs rules is significant and positive ($B = 0.19$, 95% CI [0.11, 0.27], $t(941) = 4.63$, $p < .001$; $\beta = 0.26$, 95% CI [0.15, 0.36]). The interaction effect of agent and judge rules is significant and positive ($B = 0.02$, 95% CI [0.002, 0.03], $t(941) = 2.24$, $p = 0.026$; $\beta = 0.02$, 95% CI [0.002, 0.04]).

We further collected data on the recognition of rules and strings after the experiment in order to identify if the participants inferred rules from the interaction with agents knowing more rules. Such inference would be an exciting find for social learning. However, it would make the interpretation of the results presented already more difficult since the difference in agent and judge knowledge is less defined. However, the exploratory results indicate that following the experiment, participants who were shown more rules were more sensitive towards the correct rules in a rule recognition task ($B = 0.35$, 95% CI [0.29, 0.41], $t(235) = 11.95$, $p < .001$; $\beta = 0.61$, 95% CI [0.51, 0.72]), and more sensitive towards the correct strings in a string recognition task ($B = 0.39$, 95% CI [0.35, 0.43], $t(235) = 19.26$, $p < .001$; $\beta = 0.78$, 95% CI [0.70, 0.86]). While these results are no proof of an absence of inference, which is impossible, they show that the participants with more rule knowledge can recognize and apply them better.

Figure 13

Interaction effect of agent and judge rule knowledge on judge certainty



Discussion

The first study experimentally confirms the hypothesis that novices cannot mentalize the decisions of an expert and thus lack the information to accurately assess the expert's expertise. The distributions suggest that the participants used an anchoring and adjustment strategy (Epley et al., 2004). Some influence of random guessing could be present in the data, but this does not alter the interpretation (Nuhfer et al., 2017). The hypothesis was confirmed through three separate analyses, indicating robustness against the chosen analysis method. The validity of the experiment itself is very limited, though. While there are cases in which expertise is built up in a

strictly linear fashion (e.g., learning to add before learning multiplication), knowledge is rarely as strictly linear as in this experiment. Moreover, there can be specializations within the same level of expertise that are not shared among all. For example, chess players of the same level could specialize in different openings and, as such, fail to mentalize each other's moves even though they have comparable expertise overall. This could mean that a specialized novice could even surprise a more expert player if they lack this specific knowledge. This would make expertise a necessary but not a sufficient condition for accurate expertise judgment.

Higher expertise also leads to more confidence in the expertise judgments. The confidence is well calibrated since, with higher judge expertise, the judgments become more accurate.

During the experiments, the participants did not learn the rules hidden from them by observing the performance of more expert agents. This validates the results further since they cannot stem from an unexpected improvement in the observer's task knowledge. However, the experiment is very constructed, and knowledge is rarely as linear as we created it for the experiment. To address the low ecological validity of the experiment due to the linearity of the task knowledge, we replicated the results in a second study measuring the natural variance of expertise with trivia questions instead of manipulating it.

Study 2

To show the generality of the expertise imbalance hypothesis, we replicated the findings from Study 1 using the naturally occurring knowledge differences in three trivia categories. The main hypothesis we wish to test in this experiment is that an observer needs comparatively equal to higher task knowledge to assess an agent's task knowledge accurately. To test this hypothesis, we derive two statistically testable Hypotheses:

H1: Estimation Error is inversely related to the expertise of the observer. Higher observer expertise leads to lower Estimation Error.

H2: An interaction effect exists between participant and agent expertise on Estimation Error.

Specifically, higher agent expertise results in a steeper negative slope of the effect of observer expertise on Estimation Error.

Methods

This study's methods, hypotheses, and analyses were preregistered under https://osf.io/mjhcr/. The preregistration, materials, data, and analyses are openly accessible under https://osf.io/gbt64/.

Simulation

We again simulated a possible cognitive process before starting the experiment in order to estimate the statistical power we would have. We used item response theory to simulate the judge's and agent's correct answers. The judge skill and item difficulty parameters were drawn from a standard normal distribution.

We assumed that regardless of the correctness of the answer, if the agent's answer a_i was the same as the judge's answer j_i, the judge would count it as a correct answer, with a fixed probability of $\gamma = 0.4$ to change their opinion on this item.

$$r(j_i, a_i) = \begin{cases} |1 - \gamma|, & j_i = a_i \\ |0 - \gamma|, & j_i \neq a_i \end{cases}$$

The more certain a judge is, the less likely they should judge a different answer as correct. From the equality of the answers and the probability for judges to judge opposing to their answer, we calculated the judgments per item and finally used the sum of answers judged as correct as the expertise judgment.

The model we finally used to simulate the estimated agent expertise $\hat{k}$ is essentially a binomial distribution. However, the probability of success is different if the judge and agent answer the same.

$$\hat{k} = \sum_{i=1}^{n} Bernoulli(r(j_i, a_i))$$

Using this simulation, we found out that 100 participants are sufficient to reach a power β - 1 = .85 with an effect size of at least d = 0.20.

Participants

We collected data from 200 participants via the Prolific panel. The participants were paid 5£ for a 30-minute experiment. Two participants withdrew their data at the end of the study, so the final sample consisted of 198 participants. The average age of the participants was $M_{age} = 41.45$ ($SD_{age} = 12.85$). The self-identified genders were $n_{male} = 109$, $n_{female} = 87$, and $n_{other} = 2$.

Material

For the trivia quiz, we used the existing multiple-choice questions of Coane and Umanath (2021). The questions were not categorized, so we used GPT-3.5 to give rough categories to the 421 questions. We then manually summarized the question based on the GPT categories into three major categories: history, arts & culture, and science & technology (Appendix B2). Based on the elderly sample of Coane and Umanath (2021), we calculated the difficulty of the questions and randomly chose 20 questions per category of equally distributed difficulty. We manually exchanged questions that were too US-centered to make the questions accessible to a worldwide sample.

To create trivia questionnaires filled out by others, we used the same questions. The answers were calculated using item response theory (Fox, 2010; Kelly et al., 2023). The difficulty of the questions was calculated earlier from the original questions sample. We defined the person variable to be the 10, 50, and 90 percentiles of the standard normal distribution. We refer to them as low, mid, and high expertise, respectively. Using this information, we simulated the answers of 9 agents. We repeated the random process five times until each agent correctly answered at least three questions more than the next worst agent.

Procedure

At the beginning of the experiment, the participant gave informed consent and demographic data. Following comprehension and attention checks, participants were presented

with the first trivia topic. Participants were asked to estimate their expertise on a 7-point scale and assess how many of the following 20 questions on the topic they believe to be answered correctly. They then answered the multiple-choice questions. Following the questions, they repeated the estimation of how many they got right. Then, the participants saw the agents' filled-out questionnaires. The answer chosen by the agent was highlighted in orange. For each question, participants could mark the agent's output as answered correctly or not. In the end, the participant was asked for an overall estimation of the number of correct answers and an estimation of the agent's expertise. After the three agents, the participants went on to the next topic, repeating the procedure. The topics were presented randomly. Within a participant questionnaire, the questions and their answers were presented randomly. The agents and their questions were randomized, too. Only the multiple choice options of the agents were presented in a fixed order due to limitations of the survey program, SoSci Survey (Leiner, 2023).

Design and Variables

We used a 3x3 within-design quasi-experiment. The factor agent has three levels: low, mid, and high. The other factor is the trivia topic. Its levels are history, arts & culture, and science & technology.

The most important measured variables are the participants' expertise and their judgments of the agents' correct answers. The participant's expertise was calculated as the person variable of the item response theory. The participant's judgments were transformed into a judgment error, as in Study 1. This measure has the same problems as Study 1, with restricted variability of errors for intermediate agents.

Further measures are the metacognitive measures of the participants estimating their own success at the trivia questions before and after answering them and the questions asking specifically for the expertise of the agents.

Analyses

As in Study 1, the analyses were conducted using R (R Core Team, 2023). The models were created using lme4 (Bates et al., 2015). The degrees of freedom were estimated using the Satterwhite method implemented in the lmerTest package (Kuznetsova et al., 2017).

The main hypothesis was calculated using a linear mixed model with the participant's judgment error as dependent variable, with judge and agent expertise as fixed effects, controlled for the separate topics, and participants as a random effect. Additionally we caluculated this model with different measures of judge and agent expertise. Further, we analyzed the self-estimation of expertise by the participants using a linear mixed model with participant measured expertise as predictor, controlling for the topics, with participant as random variable. We conducted this analysis for uninformed self-estimation based on only the topic before the participant saw the questions, and informed self-estimation following the questions.

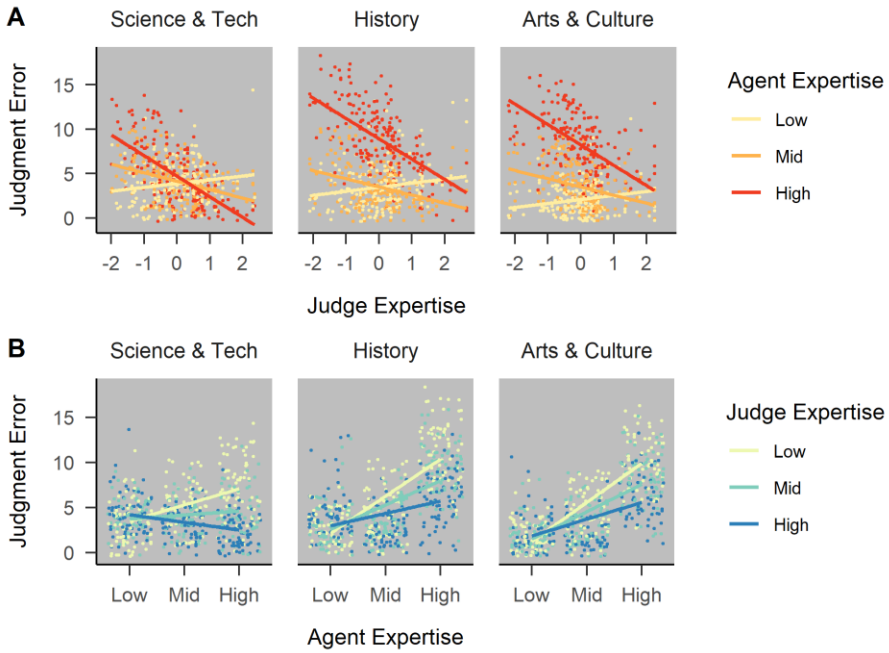
Results

The interaction effect of judge and agent expertise is significant and negative (B = -1.08, 95% CI [-1.19, -0.96], t(1773) = -18.57, p < .001; β = -0.29, 95% CI [-0.32, -0.26]). The effect is illustrated in Figure 14. The result replicates the effect from Study 1 and reinforces the claim that with increasing expertise, one can more accurately assess the expertise of others. The main effect of judge expertise is significant (B = 1.17, 95% CI [0.98, 1.35], t(1773) = 12.43, p < .001; β = 0.20, 95% CI [0.17, 0.23]), so is the main effect of agent expertise (B = -3.79, 95% CI [-4.78, -2.80], t(1773) = -7.50, p < .001; β = -0.75, 95% CI [-0.95, -0.56]). The effects of topic, dummy coded with Arts & Culture as the reference group, were non-significant. Based on this analysis, H1 and H2 have been confirmed. With a higher expertise of the observer, the judgments of the agent's expertise are more accurate. Further, this effect is moderated by the agent's expertise, as an expert agent's expertise is difficult to estimate by more judges. We continue to test the hypotheses' robustness by using different measures of expertise and expertise estimation.

We tested the robustness of the analysis in terms of the measure of expertise. For the analysis, we used the IRT person variable θ as a predictor for the model. However, directly using the sum of correct answers is another legitimate measure of expertise for the agent and judge. The results of the linear mixed model predicting the judgment error from the sum of correct answers of the judge and the agent, controlled for topic and with random intercepts for participant and agent, are presented in Table 2.

Figure 14

Expertise Interaction Effect



We further had two measures of expertise judgment. The measure we used in the earlier analyses was the error in estimating how many questions the agent answered correctly.

We measured the estimated expertise on a 7-point Lickert scale as a second measure. Table 3 presents the same model as above, with θ as predictors and the expertise Lickert scale as the dependent variable. These two analyses lead to the same conclusion as the first model: If the judge's expertise is higher, their expertise judgment of others is more accurate.

An examination of Figure 14 suggests a small but noticeable increase in errors with increasing judge expertise when judging low-expertise agents. A follow-up, exploratory analysis examining this effect reveals a significant positive relationship between judge expertise and error for the judgment of low expertise agents (B = 0.40, 95% CI [0.19, 0.62], $t(588) = 3.72$, $p < .001$; β = 0.16, 95% CI [0.07, 0.24]). The effect is comparatively small and exploratory; thus, it is to be interpreted carefully. However, it does suggest that it is difficult for experts to take the perspective of novices.

Table 2

Linear Mixed Model of Judgement Error by the sum of correct answers of judge and agent

	Estimate	Std. Error	df	t value	Pr(>|t|)
Judge corr. answers	0.43	0.04	1,768.30	9.67	0.00
Agent corr. answers	0.99	0.11	6.36	8.88	0.00
Topic: History	0.34	1.28	5.00	0.27	0.80
Topic: Science & Tech	0.17	1.29	5.00	0.13	0.90
Judge x Agent corr. answers	-0.06	0.00	1,578.03	-17.27	0.00

As in Study 1, we examined the standard deviation of the judgment errors the judges make. Again, the variance of the errors is smaller the more expertise the judge has (B = -0.57, 95% CI [-0.69, -0.46], $t(592) = -9.57$, $p < .001$; β = -0.37, 95% CI [-0.44, -0.29]). This means the errors are less random if the judge has more expertise.

In summary, these results replicate the findings of Study 1. Judges with higher expertise give more precise estimations of others' expertise. However, there is an effect indicating that experts are slightly worse at judging novices.

Table 3

Linear Mixed Model of Estimated Expertise by sum of correct answers of judge and agent

	Estimate	Std. Error	df	t value	Pr(>|t|)
Judge Theta	0.22	0.04	1,249.72	5.91	0.00
Agent Theta	0.33	0.09	5.00	3.49	0.02
Topic: History	0.34	0.24	5.00	1.41	0.22
Topic: Science & Tech	0.10	0.24	5.00	0.42	0.69
Judge x Agent Theta	0.19	0.02	1,574.20	9.13	0.00

In addition to estimating the expertise of others, the participants also estimated their own expertise by guessing the number of correctly answered questions on a topic before and after answering them. In an exploratory analysis, we related the self-estimation to the participant's expertise. The effect of expertise on the uninformed guessing is significant and positive (B = 0.32, 95% CI [0.02, 0.62], t(588) = 2.08, p = 0.038; β = 0.09, 95% CI [5.08e-03, 0.18]). The same is true for the effect of expertise on informed guessing (B = 0.36, 95% CI [0.10, 0.63], t(588) = 2.73, p = 0.007; β = 0.12, 95% CI [0.03, 0.21]). As opposed to the hypothesis of Kruger and Dunning (1999), these results rather suggest a slightly worse metacognitive accuracy with increasing expertise.

Discussion

In Study 2, we were able to replicate the result from Study 1. Experts give a more precise judgment of an agent's expertise than novices do. This result is robust towards different measures of expertise and judgments. Note that novices' error distribution is consistent with random guessing of the number of correct answers from the agent. The precision of expert judges, on the other hand, is not. In this more natural study, we found a slight decrease in judgment accuracy with higher expertise of the judge with regard to low expertise agents. The decrease is consistent with a curse of knowledge in that experts are overestimating the general population's knowledge and consequently overestimating the performance of novices. Further, the

metacognitive accuracy was adequate on all levels. Following the Dunning-Kruger effect, one would expect a lower metacognitive accuracy of novices due to a worse understanding of the task (Kruger & Dunning, 1999). In contrast to the expected double burden of novices, they are even slightly more accurate in their self-estimation. This finding is exploratory and small. However, it is consistent with prior findings on metacognition as a function of expertise (McIntosh et al., 2019). While we did not find a dual burden of self-estimation, the dual burden of the estimation of others seems very real. A novice is bad at a task and aware of their incompetence. However, they are unaware of how much more competent the agent might be. The possible consequences of this effect will be further elaborated in the general discussion.

General Discussion

Summary

In these two studies, we have shown an interaction effect of an observer's and an agent's expertise on the observer's judgment accuracy of the agent's expertise. This judgment was solely based on cues from the agent's decision-making.

In the first experiment, we yielded an interaction effect between judge and agent expertise on the expertise judgment error. The experiment was somewhat artificial, using a linear and categorical knowledge space. In reality, knowledge rarely has these features. This critique was addressed in a second study. The second study featured naturally occurring knowledge for a reduced experimental rigor. The interaction effect could be reproduced in the more naturally valid scenario. These effects are indicative of an expertise imbalance effect, in the sense that experts are able to accurately assess any agent that only has a subset of their knowledge. On the other hand, novices have to guess the expertise of anyone with a more extensive knowledge.

Interpretation of results

People cannot accurately judge others' decisions if they do not understand the task they are performing. While it might seem trivially true, this could be a hard limitation of perceiving agents in the social environment that should be acknowledged. The judgment of expertise is

becoming more crucial because of the credibility associated with it. In our society, we invent cues such as degrees or standardized exams, and people use unreliable, even unrelated cues such as attractiveness or height (Nisbett & Wilson, 1977) to make expertise judgments without having to evaluate the task performance of people who might be better at the task. We do this because of the enormous cost associated with giving tasks to and trusting less qualified agents due to the limited task knowledge of the judges. For example, a politician can claim anything without harm to their reputation if only a handful of experts know enough about the topic to uncover the lie. To these experts, the politician's lack of expertise is obvious. Others, however, might believe the politician because of his social status and general credibility. With the increasing importance of the internet, individuals have to judge the credibility of others online using only very little information about the other person (Lucassen & Schraagen, 2011).

Our results align well with those of theory of mind research. The participants were able to make accurate inferences about an agent's knowledge, given that they had a precise knowledge of the agent's task (Aboody et al., 2021; Baker et al., 2017). However, these inferences were less accurate when the observer had less knowledge about the task. Research in this area could ask questions on the effects of the observer's perfect task knowledge. In reality, theory of mind might be more difficult than in laboratory situations. People rarely have full knowledge of an agent's task. Missing knowledge, it is understandable that they are unable to ascribe such knowledge to an agent performing the task.

According to our research, deception only works via manipulation of other expertise cues (attributes and metrics). However, it is impossible to perform like an expert surgeon in an ER if one is a novice in medical tasks. If they are present, experts will be able to detect the errors of the imposter. To the best of our pop cultural knowledge about conpeople, presenting faked evidence of expertise and avoiding doing the task are the strategies they use to avoid suspicion. On the other hand, an expert has the option to hide their expertise by acting erratically and deliberately making errors. However, they have to weigh the use of hiding their expertise against

the potentially worse task results. It would be interesting to research the ability of experts and novices to hide their status and their abilities to detect imposters.

Without cues from the agent, any knowledge of the judges makes it difficult for them to take a novice's perspective (Tullis, 2018; Tullis & Feder, 2022). However, we have shown that experts quickly and accurately identify novices if given appropriate information. Even though the results of the second study suggest that taking a novice's perspective is difficult for an expert, it seems more difficult for a novice to take an expert's perspective. Within our experiments, knowledge is a blessing rather than a curse (Birch, 2005).

Our results suggest that the metacognitive accuracy for self-evaluation is well-calibrated throughout all levels of expertise. Even novices have some available information that tells them the extent of their "ignorance" (Kruger & Dunning, 1999). Information stemming from their own cognition is lacking once they need to estimate the expertise of others. The accuracy of novices might improve if they get additional information about the agent's metacognition. Following this idea, experts have a dual advantage over novices in the sense that they perform better at a task and can also assess the knowledge of others more accurately (McIntosh et al., 2019). This imbalance might have severe societal consequences. Especially in online interactions, where the information about the other is very sparse, one frequently has to judge others by the products of their decision-making. In such situations, experts might only be recognized by other experts, while for novices, experts and other novices are indistinguishable. The confusion might lead to a misplacement of trust in entities unworthy of it.

Limitation and Scope

The findings presented here are limited to judgments based purely on the information from the products of a decision-making process. We speculate that these cues can only be interpreted by observers who can mentally recreate the decisions leading to this product. Any other cue, like the speed of an answer (i.e., confidence) or the agent's reputation, would likely

moderate the results. We think that novices might rely more on such cues than an expert would if the other cues turn out to be unreliable.

The scope of the presented results is further limited to situations in which the truth can be identified. Naturally, it is more difficult to assess the expertise of people on social issues on which there are controversial attitudes, and truth is a more fluid concept. Assessments of expertise also depend on how closely the information expressed by an agent fits with one's own worldview.

Further, we have only used answers generated automatically by the agent. However, due to the nature of the studies, the participants were not made aware of the attributes of the agent. So, to them, the agent could also have been an earlier participant or a cat walking over a keyboard. Even though we did not specifically test it, we claim that the expertise imbalance exists between any two entities with an agency capable of expertise. This will be true only insofar as the judge is not aware of the attributes of the agent that we already discussed, which likely moderate the effect.

Outlook

It has been shown that novices do have metacognitive insight into their knowledge but are missing the knowledge necessary for successful task completion (McIntosh et al., 2019). In our specification of an expertise judgment of others, only the task knowledge is relevant, as the agents do not communicate their metacognitive experience. The agent communicating the certainty in their decisions, or even explaining their decision process, could eliminate the effect as the judge has more information on which to base their expertise judgment (Birch et al., 2010). On the other hand, novices might not use this information as efficiently as experts or are more susceptible to unreliable depictions of certainty. Research in this direction is necessary to test these contrasting ideas. Other cues like agent attributes, reputation, or performance metrics will likely moderate the expertise imbalance, too. We think these cues in the presence of decision cues will less strongly influence experts who can interpret an agent's decisions.

The judgment of an agent's expertise likely has consequences for other judgments. Trust in an agent is influenced by many factors, and domain expertise is one of them. The difficulties of novices in accurately assessing an agent's expertise might have consequences for trust calibration. Given that novices cannot interpret the decisions of an expert agent, other cues might be given more weight (Lucassen & Schraagen, 2011).

Experts' decisions are generally closer to the optimal decision, as opposed to novices, which is more exploratory and random. This should make an expert's decision-making more reliable and predictable. However, again, only to other experts since the observer has to know an optimal decision, too. This could help in cooperative task solving, where an accurate mental representation of the partner helps to efficiently work together.

Conclusion

In conclusion, in this paper, we established a new hypothesis on how the expertise of two people interacts with each other and supported it experimentally and quasi-experimentally. We found that with less expertise, the judgments of an agent's expertise become less accurate. We argue that the imbalance comes from the inability of novices to mentally recreate an expert's decision process. However, the effect of this expertise imbalance is limited to specific information from the agent. Pending independent replication and further validation of the hypothesis, it is a starting point for extension into research on moderating cues.

Chapter 4 – Trust Calibration

The fourth chapter is a collaborative manuscript of Fritz Becker (first author), Celine Spannagl, M.Sc. (second author), Dr. habil. Jürgen Buder (third author), and Prof. Dr. Markus Huff (fourth author). It has the title “Performance rather than Reputation Affects Humans’ Trust towards an Artificial Agent.” The manuscript was published in Computers in Human Behavior: Artificial Humans, Volume 3, 2025, 100122, ISSN 2949-8821, https://doi.org/10.1016/j.chbah.2025.100122.

Abstract

To succeed in teamwork with artificial agents, humans have to calibrate their trust towards agents based on information they receive about an agent before interaction (reputation information) as well as on experiences they have during interaction (agent performance). This study (N = 253) focuses on the influence of a virtual agent's reputation (high/low) and actual observed performance (high/low) on a human user's behavioral trust (delegation behavior) and self-reported trust (questionnaires) in a cooperative Tetris game. The main findings suggest that agent reputation influences self-reported trust prior to interaction. However, the effect of reputation immediately gets overridden by performance of the agent during the interaction. The agent's performance during the interactive task influenced delegation behavior as well as self-reported trust measured post-interaction. Pre- to post-change in self-reported trust is significantly larger when reputation and performance are incongruent. We conclude that reputation might have a smaller than expected influence on behavior in the presence of a novel tool that affords exploration. Our research contributes to understanding trust and delegation dynamics, which is crucial for the design and adequate use of artificial agent team partners in a world of digital transformation.

Introduction

In times of digital transformation, cooperative, collaborative, and delegative interaction with artificial agents such as computers, virtual bots, embodied robots, and deep neural networks is becoming increasingly common, with an expectation for the future to further increase in relevance (Baltieri et al., 2023). To successfully use agentic tools, a careful calibration of the trust a user puts into the agent is necessary (Cancro et al., 2022). Unwarranted doubts about the agents' knowledge, ability, or biases may lead to unnecessary micromanagement and under-delegation, even of agents that are highly skilled or specialized (Candrian & Scherer, 2022). Conversely, over-delegation and over-trust as a result of high reputation and overestimation of the agents' capabilities can also lead to lowered team performance (De Visser et al., 2020). Accurate calibration of trust is made even more difficult by the motivated manipulation of an agent's reputation by the companies selling such agentic systems. An inaccurate reputation of an agent might create more difficulties for an accurate calibration of trust towards it. In this article, we intend to analyze how people's trust towards an agent in a cooperative game develops over time based on the agent's reputation and performance.

Human-Agent Teams

Continuing developments in automation technology cause modern computer systems to no longer be predictable tools but rather autonomous agents. These technological advancements gave artificial systems the capability to work cooperatively with humans on complex tasks. Technological developments in robotics and artificial intelligence are leading towards work scenarios in the near future where "autonomous agents are motivated beings working alongside their human counterparts" (Larson & DeChurch, 2020, p. 2). Over the last few years, with considerable advances in machine learning and cognitive modeling as well as the increasing availability of data and advanced computational resources, studies concerning the term human-autonomy teaming are thriving (Daronnat et al., 2019, 2021; Hafizoğlu & Sen, 2018a, 2018b; McNeese et al., 2018, 2021; O'Neill et al., 2022). Contrary to prior work in which autonomous

agents were only discussed as subservient aids, agents are increasingly viewed as associates that can work independently and collaboratively (O'Neill et al., 2022). In the context of *human-agent-teams* (HAT), the teams are described as humans and autonomous agents, respectively, as individual members with distinct, unique roles, working interdependently toward a common goal (O'Neill et al., 2022). In our research, we focus on the interaction of an individual human with a non-human artificial agent on a cooperative task.

Trust

Although conceptualizations of trust have become increasingly unified in recent research, a universally accepted definition or model of trust in agents has yet to emerge. This may be due to the contextual nature of trust (Kohn et al., 2021). We understand trust as an agent's (trustor's) willingness to rely on and be vulnerable to another agent's (trustee's) actions in the face of uncertainty and potential risk, based on the expectation that the trustee will perform a particular action regardless of the ability to monitor or control the trustee's behavior (R. C. Mayer et al., 1995). Similar definitions can be found throughout research on trust in HAT (Kulms & Kopp, 2016; J. D. Lee & See, 2004; McNeese et al., 2021). For instance, trust in a particular automation technology is defined as "a human's propensity to submit to vulnerability and unpredictability, and nevertheless to use that automation" (Sheridan, 2019, p. 2). The extent of trust is both a function of trustor features (e.g., dispositional trust) and trustee features (trustworthiness). When evaluating trustworthiness, the trustor can decide, based on trustee characteristics such as ability, benevolence, and integrity, how risky placing their trust in the trustee could potentially be (Kulms & Kopp, 2016). Hence, trust affects cooperative behavior in teams and thus impacts team performance and task outcomes. Trust in humans, automation, and robots is found to be comparable and founded in shared underlying traits such as reliability, predictability, and ability (J. Jian et al., 2000; Ullrich et al., 2021). For the present study, we manipulated two aspects of the ability facet of trustworthiness: the purported ability of an agent (reputation) and the actual ability of an agent (performance).

Measuring Trust

Trust, as a latent and not directly measurable variable, is generally measured through self-report surveys and questionnaires (Kohn et al., 2021; O'Neill et al., 2022). Self-report measures are simple and have high face validity. However, limitations of self-reported measures consist of interruption of tasks through the survey, the lack of capability of surveys to capture a continuous evolution of trust due to assessment at few single time points, as well as memory failures and subjective bias (Kohn et al., 2021). Therefore, behavioral measures are increasingly utilized in research on trust in HAT (Daronnat et al., 2020, 2021; Hafizoğlu & Sen, 2018a, 2018b; Kulms & Kopp, 2019). Trust in agents can impact behavioral processes or tendencies, including risk-associated decisions (e.g., delegation, reliance, cooperation, or intervention) as well as outcome-related measures (e.g., decision and response time, combined team performance), which are considered behavioral indicators of trust (Kohn et al., 2021). These behavioral indicators of trust can be measured via observation and systematic recording. It is suggested that self-reported and behavioral measures should ideally both be employed, as different trust measures may capture distinct facets of trust (Kohn et al., 2021). For the present study, trust was measured using both self-report questionnaires and task delegation as a behavioral measure.

To measure trust, many studies employ game-based frameworks of HAT interaction (Correia et al., 2018; Daronnat et al., 2020, 2021; Kulms & Kopp, 2016). Examples of these are computer-simulated interaction scenarios with virtual agents, e.g., in military situations like missile shooting (Daronnat et al., 2021, 2022), dependence on a virtual robot in emergency evacuation situations (Robinette et al., 2017), or reliance on a pet feeding robot during absence (Ullrich et al., 2021). Moreover, trust games with lower-risk scenarios are used, such as collaborative word-transcription tasks (Hafizoğlu & Sen, 2018a, 2018b) or a collaborative Tetris puzzle game (Kulms et al., 2015; Kulms & Kopp, 2016, 2019). We used a Tetris-like game in the present study, too, in order to create a low-risk yet engaging situation in which the participants can meaningfully interact with an agent.

Reputation

Whether humans actually decide to collaborate with an agent might hinge on information that is provided about the agent's capabilities and features (i.e., on agent reputation). Reputation is an important influence factor in decision-making and also a mechanism to encourage and sustain cooperative behavior (Duradoni et al., 2021; Xu, 2014). Reputation can give guidance in choosing and deploying the right tool or agent partner for a particular task or in delegating certain tasks (Hafizoğlu & Sen, 2018a). Further, reputation has been observed to be one of the most dominant influencing factors on the base level of trust as it provides knowledge about the agent's reliability and past performance prior to any direct interactions with the agent (Ullrich et al., 2021). Reputation can be manipulated via written or oral information about the agent's properties, e.g., the trustworthiness (Hafizoğlu & Sen, 2018a) or its functional performance and abilities as well as social components like ratings and use by peers (Alarcon et al., 2020). More generally, warranting theory (Walther et al., 2009) suggests that reputation information by third parties is often perceived as highly diagnostic because it is more difficult to manipulate than information from agents themselves. In the context of this study, we manipulated reputation (high vs. low) by providing different information about an agent's capability prior to interaction.

Reputation can influence participants' expectations and, thus, initial trust attitudes and behavior. In a game of trust, a positive agent reputation leads to greater trust of the human in their agent partner (Hafizoğlu & Sen, 2018a). Programmers reported higher trustworthiness for software that had better reputation indication (Alarcon et al., 2020). However, a manipulation of reputation based on customer reviews for an agent was quickly overwritten by a person's own experience with the agent (Ullrich et al., 2021). These results show that reputation does have an effect on the perceived trustworthiness of agents, but this effect might be short-lived in the wake of personal experience with them.

Performance

An agent's competence, which is demonstrated by its actual performance, is another key factor influencing trust. In contrast to reputation, which indirectly informs about the agent's capabilities, observed real-time behavior and feedback can demonstrate the actual performance of an agent (Robinette et al., 2017; Ullrich et al., 2021). Past and current personal experience with a robot can be used to deduce judgments about the robot's performance (Hafizoğlu & Sen, 2018b; Robinette et al., 2017; Ullrich et al., 2021). An agent's performance can be manipulated by the number or types of errors or randomness in the system's behavior. If humans consistently see the robot perform well, they develop general trust in the robot and positive expectations about the robot's future performance (Ullrich et al., 2021).

In experiments, the influence of an agent's performance on trust could repeatedly be shown (Daronnat et al., 2020; Kulms & Kopp, 2016; Robinette et al., 2017). The performance of an agent is among the most influential factors of trust development. It is ranked higher than environmental factors and agent characteristics (Hancock et al., 2011). Over time, the experienced performance of an agent tends to outweigh the a priori information about an agent, such as its appearance (Kulms & Kopp, 2016). Notably, not only the agent's performance but also the team's performance, which is co-dependent on human performance, has an impact on the perceived trustworthiness of the agent (McNeese et al., 2021). Overall, the evolution of trust, influenced by agent or team performance, is accumulative: it grows or diminishes in response to the positive and negative outcomes of a series of experienced events (Chiou & Lee, 2023; De Visser et al., 2020). This should be particularly evident in situations where actual performance and alleged reputation are inconsistent (e.g., a high-reputation agent has low performance, or a low-reputation agent has high performance).

Present Study

To our knowledge, in the context of HAT, there has been no research on the influence of performance-related reputation and observed performance (and the potential interaction of

these factors) on trust development and delegation, using behavioral as well as self-reported trust measures. Therefore, the present study aims to investigate the influence of these factors in a game-based, interdependent, successive task paradigm with low-risk and everyday work-life relevance. We continuously assess the delegations during the human-agent cooperation to observe the real-time development of trust and compare pre- and post-interaction self-reported trust measured via questionnaires. We expect that before any interaction with an agent, the trust towards it is shaped by its reputation. Once the person experiences the performance of the agent firsthand, the information they gained from others (i.e., reputation) will be continuously overwritten by the newly gained information. Based on these conceptual rationales, we preregistered the following hypotheses:

1) Initial trust in agents with high reputation is significantly higher than in agents with low reputation.
2) Over time, the change in trust for agents with reputation-inconsistent performance is significantly larger than the change for reputation-consistent agents.

To test these high-level hypotheses, we rephrased them to be more directly testable in the context of the two measures of trust we collect (delegations and self-report).

H1a) We expect a significantly higher mean of self-reported trust in the high reputation condition compared to the low reputation condition after the participants read the introductions of the agents but before the interaction with them.

H1b) For the delegations, we expect an interaction effect of reputation and trial that qualifies for a higher initial probability of delegating towards a high reputation agent.

H2a) We expect a significant interaction between reputation, performance, and time on self-report trust. Furthermore, a more immediate test of the hypothesis is a t-test between reputation-consistent and -inconsistent conditions on the absolute difference (change) between t1 and t2.

H2b) We expect a significant interaction between reputation, performance, and trial on the probability of delegating.

More exploratorily, we looked at the relationship between trust measures, self-report, and delegations. Regarding this relationship, we assumed that pre-interaction trust influences the delegation behavior, similar to reputation in H1b) and H2b). Further, we correlated the sum of delegations with the self-report data in order to validate both as adequate measures of trust.

Moreover, we analyzed whether effects were also dependent on participants' (in addition to an agent's) performance in the game. For instance, participant's performance in the game might also influence the trust they report towards the agent.

Methods

Preregistration and Ethic Vote

A preregistration of this study is available on the Open Science Framework (https://osf.io/6tdk8). Ethical approval for this study (filed under LEK 2023 /004) was obtained from the ethics committee of the Leibniz-Institut für Wissensmedien (IWM).

Participants

We recruited 301 participants via the Prolific platform (www.prolific.com). Participants were paid 5£ for the estimated 30-minute study (10£/h). Inclusion criteria were a minimum age of 18 years and a good understanding of written English. Our target sample was a minimum of 300 participants. We tested N = 301 participants (201 male, 97 female, 1 diverse, 2 NAs, M_{age} = 32.10, SD_{age} = 10.10). Data of 48 participants were excluded due to negligence in following the study instructions (which required at least 10 interactions with the bot), resulting in N = 253, with 61-69 participants per group (171 male, 80 female, 1 diverse, 1 NA, M_{age} = 31.83, SD_{age} = 9.96).

Experimental Paradigm

The current experiment used a mixed study design. Former research has reported that the order in which participants interact with agents of different reputations can influence trust behaviors (Hafızoğlu & Sen, 2018a). Hence, we decided to vary the factors performance and

reputation between subjects. Time was a within-subject factor, reputation and performance were between-subject factors, each with two levels (high/low), resulting in four (between-subject) groups. Participants were randomly assigned to one of the four conditions via SoSci Survey (Leiner, 2023):

Our independent variables (reputation and performance) were manipulated using two different reputation texts informing about the agent's reputed properties and two differently well-performing Tetris bots in the cooperative Tetris-like game. Additionally, we used trial in the experiment and pre-/post-interaction time as a third independent variable. Behavioral trust as a dependent variable was operationalized by task delegation behavior (i.e., trials in which participants delegated the subtask of choosing a Tetris block to the bot), and self-reported trust was assessed via questionnaire.

Materials

Reputation Texts

The written description for the high vs low agent reputation differed in their information regarding the bot's developmental stage ("fully tested and developed final version" vs. "testing phase as beta-version"), their expertise ("trained over 100 million rounds of the game" vs. "has not been trained in this game before") and reliability ("The bot will always choose" vs. "The bot may not always choose [the best fitting block]") (Appendix C1).

Game Environment

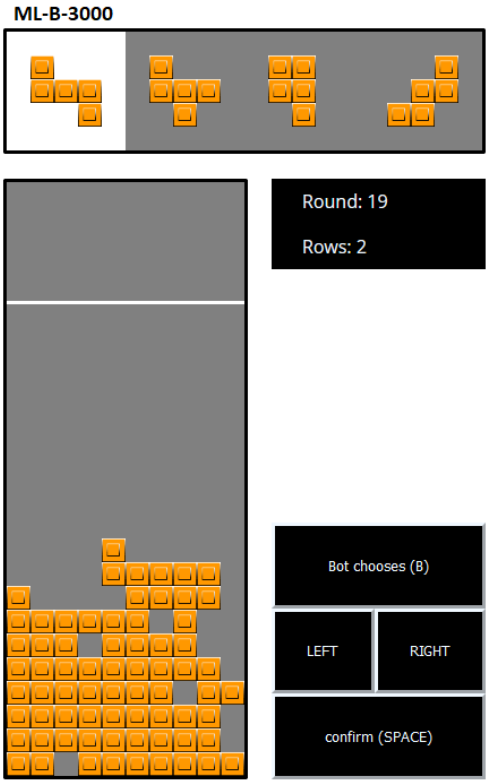
We programmed a Tetris-like game for our study (Figure 15). A similar type of game has been suggested as an interactive game framework (Kulms et al., 2015) and was used to research trust in human-agent cooperation (Buchholz et al., 2017; Kulms & Kopp, 2016, 2019). The following properties of our game differed from the commonly known original Tetris game: there was no time restriction and no dropping of pieces at varying speeds. Participants (or the bot) could choose from four selectable pieces. In order to further decrease familiarity with classical Tetris and to increase the difficulty, compensating for the increased ease by selecting a piece, the

pieces were created from five blocks (“pentominoes”) instead of four (“tetrominoes”). Like in the original Tetris game, completed rows were cleared from the interface. The goal was to complete as many rows as possible with the appropriate choice, orientation, and placement of the Tetris blocks within the given number of rounds without letting the stacked blocks exceed the white line. Exceeding the white line was considered a game failure (coded as a “game-over trial”). In case of a game over, the game restarted with no explicit damage to the game score. The high-performance bot team partners were programmed to select a Tetris block suitable for row completion consistently. In contrast, the low-performance bots were programmed to choose blocks randomly.

The cooperative task went over 100 trials. Each trial consisted of two stages. In the first stage, one Tetris piece (out of four randomly drawn pieces) had to be picked. Participants could choose to either pick themselves or delegate the pick to the bot. In the second stage, it was always the participant who had to place the picked piece into the playing field.

Figure 15

Screenshot of the Experiment



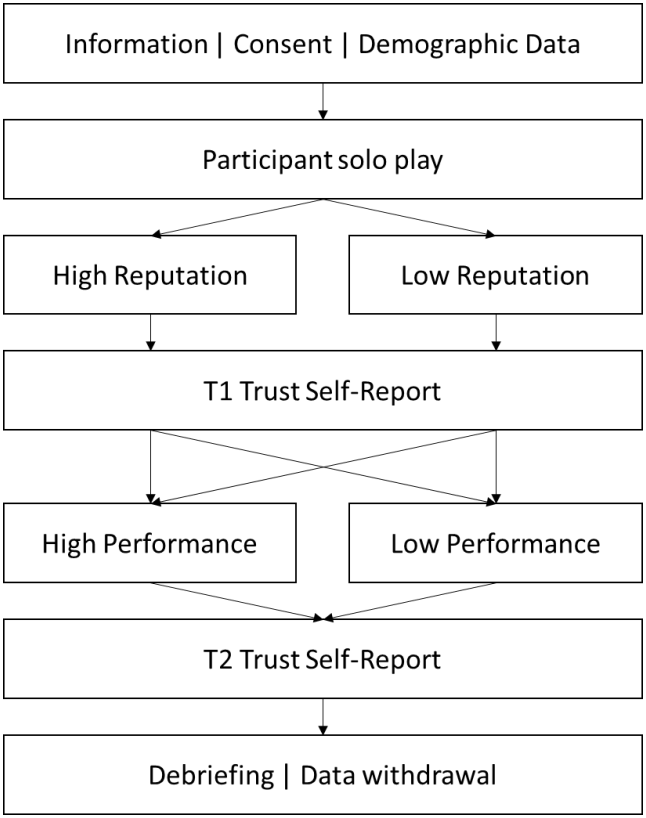
Questionnaires

For the self-report of trust, we constructed a questionnaire using items of the following trust questionnaires (Kohn et al., 2021): First, the Multi-Dimensional Measure of Trust (MDMT; Ullman & Malle, 2020), which can be used as a pre- and post-test to capture change in trust. Its subscales have a good internal consistency (Cronbach's $\alpha \geq .80$). We confirmed the internal consistency of the entire scale, $\alpha_{t1} = 0.95$ and $\alpha_{t2} = 0.97$. Second, we added items of the Source Credibility Measures (SCM; McCroskey & Teven, 1999). Third, the complete Trust in Automated

Systems Test (TOAST; Wojton et al., 2020), α_{t1} = 0.78 and α_{t2} = 0.88 (Appendix C2). The chosen questionnaires are compatible since many items between the questionnaires overlap. The construction of the current questionnaire was conducted the following way: We adopted the two factors, performance, and moral trust, from the MDMT and used the five subscales: reliability, competence, ethicality, transparency, and benevolence. Further, we added the expertise subscale to performance trust using items of the SCM (McCroskey & Teven, 1999). We chose to omit the fallback option of "does not fit" or "don't know" to make the data more feasible for our correlation analyses. We kept the 7-point Likert scale but adapted the scaling label of the original MDMT from "not at all" to "very much" to match our instruction and the scaling labels in the TOAST to "not entirely disagree" to "entirely agree." We integrated the TOAST (Wojton et al., 2020) with the three subscales: system purpose, system performance, and underlying processes to ask about trust-related system comprehension. Lastly, we added the general trust and trustworthiness items, which were often used as single-item questionnaires in HAT studies (Kohn et al., 2021). The complete trust questionnaire had a good internal consistency at both time points: α_{t1} = 0.95 and α_{t2} = 0.97. The trust scores were calculated by averaging over the subscales first and then again averaging these subscores until one single score was reached. We calculated that this procedure better represents the scale's factor structure in a confirmatory factor analysis than a single factor or a single factor per questionnaire.

Figure 16

Study Process



Procedure

First, following a brief introduction to the study, the consent, data protection declaration and data on participant demographics such as age, gender, highest level of education, and professional occupation were collected. Afterward, participants received an explanation of the Tetris game and played 50 practice trials by themselves to become familiar with the game and acquire knowledge about appropriate block choice and placement. Participants received, according to their random assignment, a written agent description (agent reputation) and then

filled out the pre-interaction trust questionnaire. For the interactive Tetris game, participants were instructed to choose the interaction with the bot at least ten times to be able to observe the bot's performance. Further, they were instructed to complete as many rows as possible within the 100 trials without letting the structure exceed the white line. Participants played 100 trials with the bot, in which the participants could either delegate the Tetris block choice to the bot or choose the block themselves. In the second step of a trial, participants always rotated and placed the blocks themselves. Completed rows disappeared, and the number of completed rows was displayed on the game surface. After the interactive Tetris game, participants completed a post-interaction trust questionnaire comprising the same items as the pre-interaction questionnaire but with retrospective phrasing and tense changes. Lastly, participants were debriefed about the study's purpose and the random assignment to the bot they interacted with (Figure 16).

Statistical Analysis

Data analysis was performed using R version 4.3.2 (R Core Team, 2023). We calculated a generalized linear mixed-effects model (GLMM) using the lme4 package (Bates et al., 2015) to estimate fixed and random effects for the correlated binary observations. The manifestation of the observed dependent variable behavioral trust was coded with 1 (choice to delegate) or 0 (choice to not delegate). We defined the fixed effects by including the main effects and interaction terms between reputation, performance, and trials and added random intercept terms for participant.

To analyze the self-reported trust per group, we calculated the trust score in accordance with the determined factor structure (for pre- and post-questionnaires) for each participant and conducted repeated measures 2×2 ANOVA with post-hoc directional t-tests for independent samples. Further, we compared each group regarding their changes in self-reported trust by the mean differences of the pre- and post-interaction questionnaire scores.

For the exploratory correlation of the behavioral and self-reported data, the sum of delegated trials per participant was correlated with the pre- and post-questionnaire data. Further,

we report a linear model of pre-interaction trust and number of delegations predicting the post-interaction trust. Lastly, we explored the effects of participant performance (yielded from the initial solo play) on the team's performance and the post-interaction questionnaire. The score for the team performance was calculated by standardizing and averaging the amount of game overs and the sum of completed rows in the interactive game. Participant performance was evaluated by standardizing and averaging the number of game overs, the sum of completed rows, their placement speed, and placement actions in the solo game. We calculated a linear regression on the team performance predicted by reputation, agent performance, and the participant's solo performance. Further, we looked at the effect of the participant's solo performance and agent performance controlled for the reputation on the post-interaction self-reported trust.

For the ANOVAs, *t*-tests, and regressions, functions of the stats package (R Core Team, 2023) were used. The effect sizes Cohen's *d* and η^2 or η_p^2 were calculated using the lsr package (Navarro, 2015) and the rstatix package (Kassambara, 2020). If possible, all tests were conducted two-sided, despite the directionally phrased hypotheses. All reported *p*-values were compared to a significance level of $\alpha = .05$, *p*-values in multiple comparisons were adjusted accordingly. Unless stated otherwise, the tests and results of the next section refer to included participant data ($N_{included} = 253$). As preregistered, incomplete surveys or disregard of instructions (e.g., less than ten bot interactions) led to exclusions of participant data ($n = 48$, 15,9%). Note that for one single participant, an extra trial on the Tetris game was inadvertently recorded by a program error. We excluded the extra trial (trial 101) and retained the remaining data since it was eligible for analysis.

Results

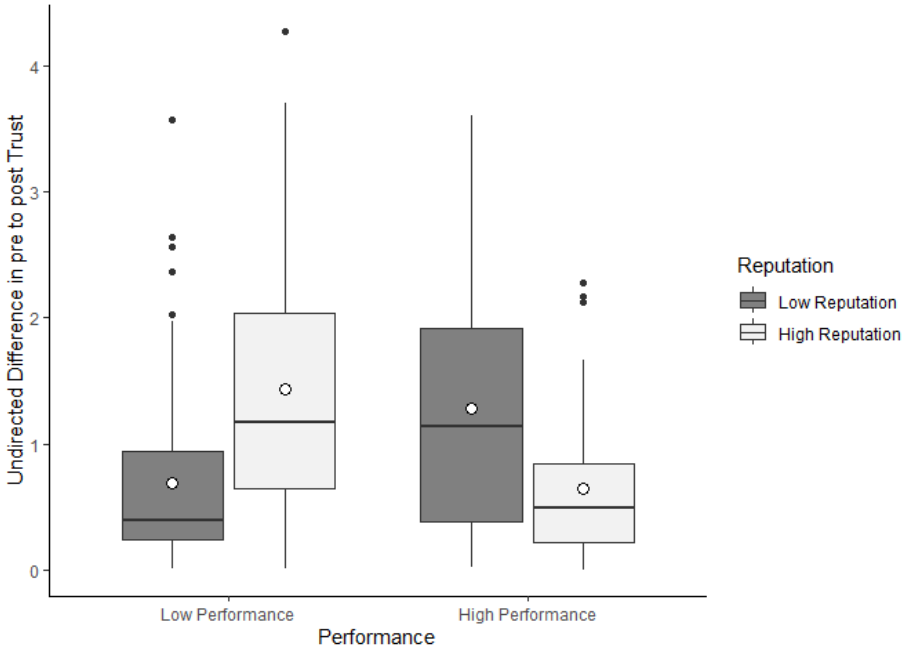
Self-Reported Trust

We collected self-reported trust data at two time points. First, after the participants read about the agent but have not yet interacted with it (t1). The second time was after the cooperative play with the agent (t2).

According to H1a) we expected an initially higher trust in a higher reputation agent than in a low reputation one. We used a t-test to test this, analyzing pre-interaction trust scores. The test revealed significantly higher trust in high reputation groups (M = 5.07, SD = 0.76) compared to low reputation groups (M = 4.09, SD = 0.94), $t(244.77) = 9.12$, $p < .001$, with a large effect size as measured by Cohens' d = 1.14. A t-test analyzing post-interaction trust scores shows significantly higher trust in high-performance groups (M = 5.27, SD = 0.96) compared to low-performance groups (M = 3.85, SD = 1.21), $t(244.97) = 10.44$, $p < .001$, with a large effect size of d = 1.30. These results show that before experience with an agent, the participant's trust is dependent on the agent's reputation. After the interaction, the performance of the agent determines the trust. Thus, based on this analysis, H1a) is confirmed.

Figure 17

Self-Reported Trust Pre- to Post-Interaction Change



Pre-Post Change

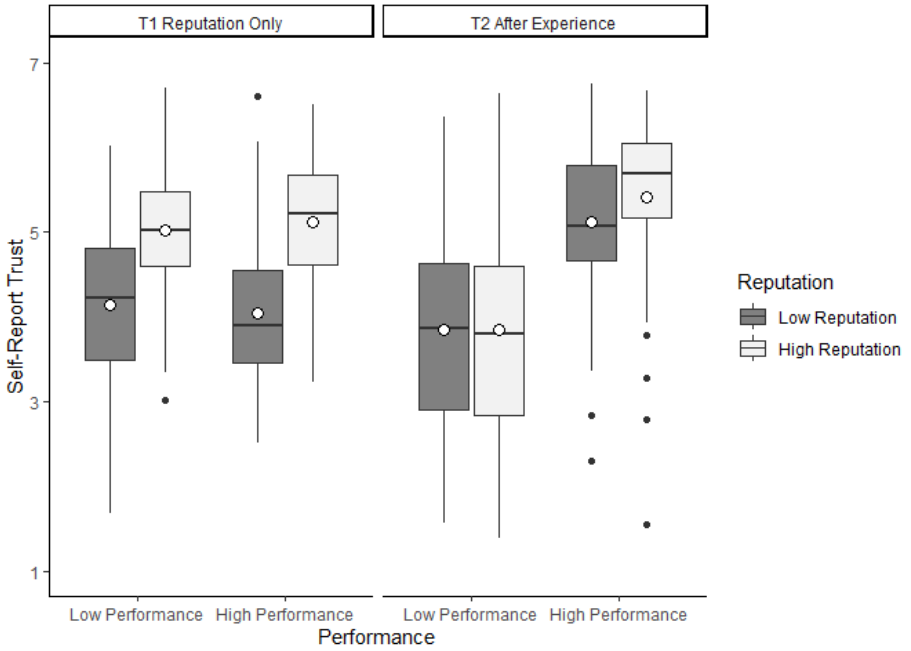
H2 states that the change of trust is greater for reputation-inconsistent conditions. The hypothesis can be tested in two ways using the self-reported data. First, we calculated the absolute difference (change) between t1 and t2 and conducted a t-test between consistent and inconsistent conditions, expecting a greater change for inconsistent groups. Secondly, we performed a repeated-measures ANOVA, expecting a three-way interaction of performance, reputation, and time.

The t-test suggests significantly larger absolute changes (Figure 17) in pre-to-post trust scores for reputation-inconsistent (M = 1.35, SD = 0.98) compared to reputation-consistent groups

(M = 0.67, SD = 0.64), t(209.78) = 6.57, p < .001, with a large effect size of d = 0.91. Regarding this t-test, H2 can be regarded as confirmed.

Figure 18

Self-Reported Trust



Note. Self-reported trust assessed in the pre- and post-interaction questionnaires. The group median is indicated by the black horizontal line, and the arithmetic mean is indicated by the white circle.

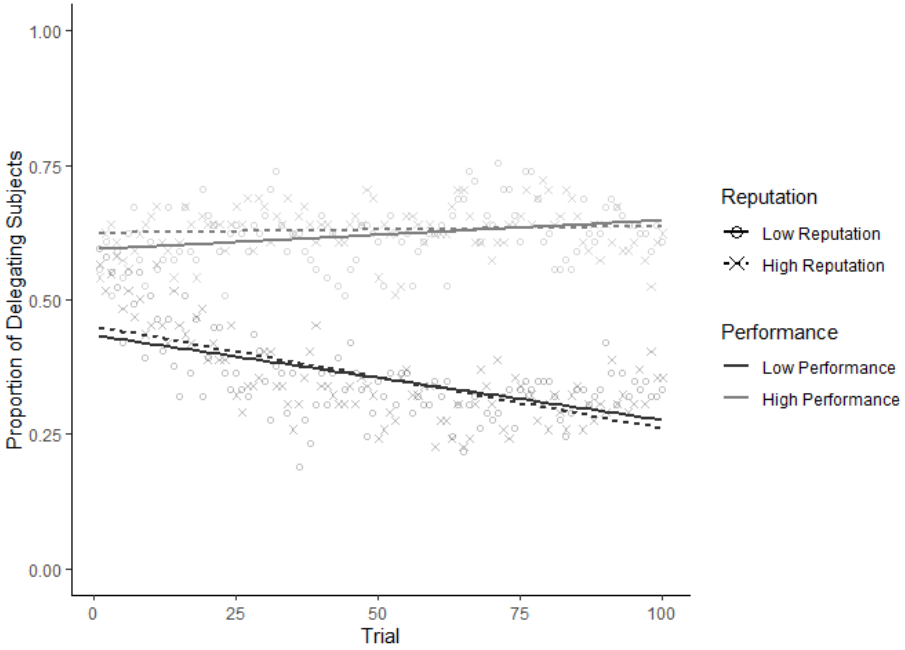
The repeated measures ANOVA, with reputation and performance as between-subject factors and assessment time (i.e., pre- or post-interaction) as within-subject factor (Figure 18), yields a medium main effect of reputation, F(1, 249) = 29.44, p < .001, η_p^2 = .11, a large main

effect for performance, $F(1, 249) = 46.99$, $p < .001$, $\eta_p^2 = .16$, and no significant main effect of assessment time $F(1, 249) = 0.09$, $p = .760$, $\eta_p^2 < .01$. There is a medium significant interaction effect between reputation and time, $F(1, 249) = 38.16$, $p < .001$, $\eta_p^2 = .13$, and a large significant interaction effect between performance and time, $F(1, 249) = 112.30$, $p < .001$, $\eta_p^2 = .31$. No significant interaction effects were found between reputation and performance, $F(1, 249) = 1.38$, $p = .242$, $\eta_p^2 < .01$, or between reputation, performance and assessment time, $F(1, 249) = 0.08$, $p = .784$, $\eta_p^2 < .01$. The main effect of reputation reconfirms H1a), reputation has a significant effect on self-report trust at t1. Due to the non-significant interaction effect of reputation, performance, and time, the results of the t-test on pre-post trust change between reputation-consistent and reputation-inconsistent groups could not be replicated. This inconsistency casts doubt on the reliability of H2a).

Dynamics of Delegation over Time

Figure 19

Development of Delegation during the Experiment



Note. The lines illustrate between-subjects variables, and trial is a within-subjects variable.

To make reliable statements about the dynamics of trust with regard to the reputation and performance of an agent, we tested the hypotheses using another measure of trust. Delegating to the agent provides evidence that participants trusted the agent to make a sensible decision in the game. The hypotheses H1b) and H2b) were tested using a generalized linear model. For H1b), we expected a significant interaction effect of reputation with time in the direction of a greater trusting behavior in early trials. For H2b), we looked for the three-way interaction

between performance, reputation, and time, indicating a differing slope between reputation-consistent and -inconsistent conditions. We expect the slopes of reputation-inconsistent conditions to be steeper than the reputation-consistent ones. Figure 19 shows the proportion of participants who chose to delegate to the bot on each trial.

Table 4

Generalized Linear Mixed Model

Predictor	*B*	SE	*z*-value	*p*-value
Intercept	1.05	0.23	4.57	<.001
Reputation	-0.26	0.32	-0.81	.417
Performance	-1.21	0.32	-3.79	<.001
Trial	0.00	0.00	0.74	.457
Reputation × Performance	0.21	0.45	0.48	.631
Reputation × Trial	0.00	0.00	1.74	.082
Performance × Trial	-0.01	0.00	-7.36	<.001
Reputation × Performance × Trial	0.00	0.00	-0.91	.363

Note. Results of the GLMM fitted on the behavioral data. Reference levels are high reputation, high performance, and trial 0. The *B*-Estimates are log odds.

We fitted a GLMM of the binomial family using Maximum Likelihood and Nelder-Mead optimizer to examine the relationship of trust with reputation and performance over the trials. Confidence intervals (95% CI) and *p*-values were computed using a Wald *z*-distribution approximation. The model included participants as a random intercept. The reference levels of the factors were high reputation, high performance, and trial 0. The model's total explanatory power is substantial (R^2 = .53), and the part related to the fixed effects (i.e., reputation, performance, and trial) alone is R^2 = .12. Within this model, the expected interactions between reputation and trial

(B = 0.003; 95% CI [-0.0004, 0.006], p = .082) and performance, reputation, and trial (B = -0.002; 95% CI [-0.006, 0.002], p = .363) are non-significant. However, the interaction of performance and trial was significant (B = -0.01, 95% CI [-0.01, -0.008], p < .001). More results are summarized in Table 4 . The interaction of reputation with trial failed to reach significance (though the effect would be significant using a one-sided criterion). Thus, the confirmation of H1b) needs to be discussed. Since the three-way interaction is not significant, H2b) was not confirmed using this model. However, we detected an effect of the agent's performance and differing developments over time regarding performance.

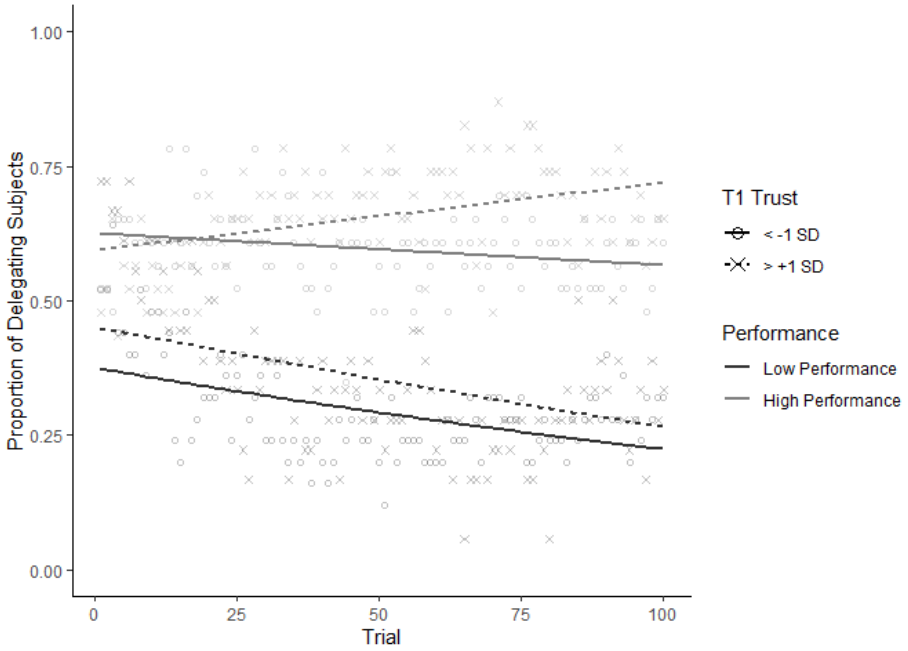
Moderating Effect of Trust on Delegation Development

To understand the influence pre-interaction trust had on the development of delegation, we exploratorily fitted a logistic mixed model (estimated using ML and Nelder-Mead optimizer) to predict delegation with t1 self-report trust, performance, and percentage of trial (time). The model included participant ID as a random effect. The model's total explanatory power is substantial (conditional R^2 = 0.53), and the part related to the fixed effects alone (marginal R^2) is 0.12. The model's intercept, corresponding to t1 self-report trust = 0, performance = 0, and time = 0, is at 0.93 (95% CI [0.61, 1.25], p < .001). Within this model, the effect of t1 self-report trust is not significant (B = -0.24, 95% CI [-0.55, 0.08], p = 0.136). The effect of performance is significant (B = -1.12, 95% CI [-1.56, -0.68], p < .001). The effect of time is significant (B = 0.20, 95% CI [0.04, 0.36], p = 0.013). The interaction effect of t1 self-report trust and performance is not significant (B = 0.38, 95% CI [-0.06, 0.82], p = 0.091). The interaction effect of t1 self-report trust and time is significant (B = 0.42, 95% CI [0.26, 0.58], p < .001). The interaction effect of performance and time is significant (B = -1.22, 95% CI [-1.43, -1.00], p < .001). The interaction effect of t1 self-report trust, performance, and time is significant (B = -0.40, 95% CI [-0.62, -0.18], p < .001). The effect is illustrated in Figure 20. The figure shows that the interaction effect is in the direction of diverging slopes for trust and parallel slopes for performance. It seems like the trusting

participants, as opposed to untrusting ones, tend to increase their delegation probability over time, but only for well-performing agents.

Figure 20

Delegation as a Function of Time, Agent Performance, and Pre-Interaction Trust



Note. The lines illustrate between-subjects variables, and trial is a within-subjects variable. The continuous variable self-report trust was dichotomized by splitting the sample into participants with t1 trust lower than -1 SD and higher than +1 SD.

Relation Between Self-Reported Trust and Delegation

To find further connections between self-reported trust and delegation, we conducted exploratory analyses correlating the self-reported trust measures with the sum of delegations. We

then calculated a regression with pre-interaction measured trust and delegation as predictors of post-interaction self-reported trust.

We found a significant positive correlation of post-interaction self-reported trust with the delegations per participant, $r = 0.41$, 95% CI [0.30, 0.51], $t(251) = 7.12$, $p < .001$. The correlation of pre-interaction self-reported trust with subsequent delegations is not significant: $r = 0.07$, 95% CI [-0.06, 0.19], $t(251) = 1.05$, $p = 0.294$. These correlations show that the pre-interaction trust holds no detectable predictive power for the number of delegations, but the sum of delegations and post-interaction trust are linked in a way that more delegations co-occur with higher post-interaction trust. Furthermore, the regression model with pre-interaction measured trust and delegation as predictors explains a significant proportion of variance in post-interaction measured trust, $R^2 = 0.27$, $F(2, 250) = 45.24$, $p < .001$, $R_{adj}^2 = 0.26$, with a residual standard error of 1.122. The results revealed significant main effects of pre-interaction measured self-reported trust, $B = .42$, 95% CI [0.27, 0.56], $t(250) = 5.77$, $p < .001$; $\beta = .31$, 95% CI [0.21, 0.42], and delegation (i.e., behavioral trust), $B = .02$, 95% CI [0.01, 0.02], $t(250) = 7.17$, $p < .001$; $\beta = 0.39$, 95% CI [0.28, 0.50], with post-interaction trust scores. The regression shows that the correlation of the number of delegations on the post-interaction trust is still relevant when controlling for the pre-interaction trust.

Task Performance and Trust

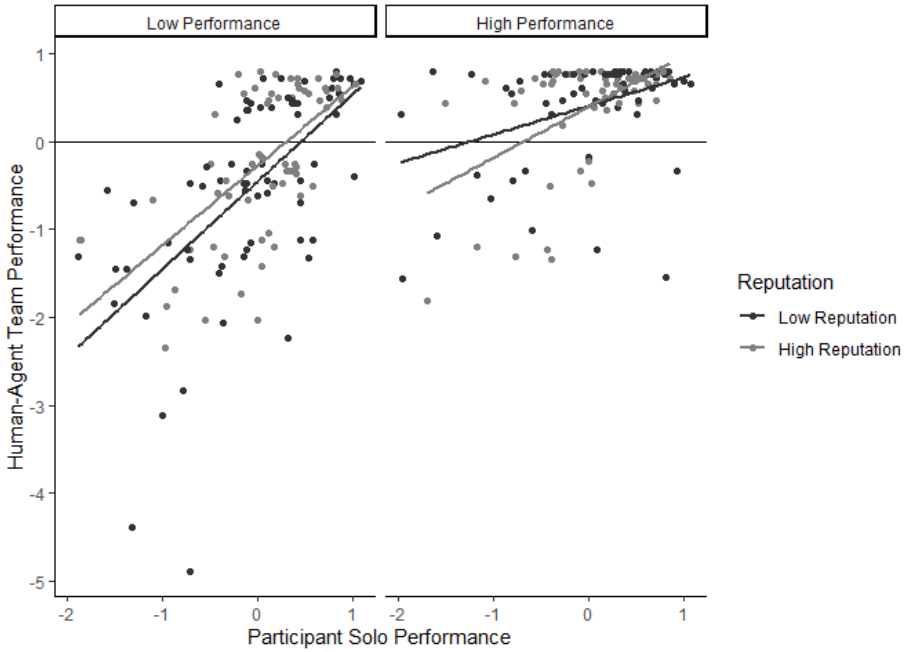
Exploratively, we measured the participant's individual and the team performance and explored their relation to each other and to post-interaction self-reported trust.

Individual Performance Effect on Team Performance

To explore the effect of the players' performances on the teams' performance, we perform a linear regression on the team's performance calculated by the mean of the standardized amount of game overs (inversed) and the completed rows. Predictors were the manipulated agent performance, agent reputation, and the participant's performance calculated from their solo play.

Figure 21

Effect of Participant and Agent performance on Team Performance



Note. Team performance is calculated as standardized sum of game overs and completed rows in a cooperative game. Participant performance is calculated as standardized movement speed and efficiency, sum of game overs, and completed rows in solo play.

In the interactive game with the bot, the team performance differed between the experimental groups regarding the agent performance B = -0.67, 95% CI [-0.92, -0.42], $t(245)$ = -5.20, p < .001, but not with regard to the reputation, B = 0.009, 95% CI [-0.24, 0.26], $t(245)$ = 0.07, p = 0.944. The participant's skill from the solo game had a significant effect on the team's performance, B = 0.58, 95% CI [0.26, 0.90], $t(245)$ = 3.62, p < .001. Yet the interaction of agent and participant performance was not significant, B = 0.33, 95% CI [-0.11, 0.76],

$t(245) = 1.47$, $p = 0.142$, neither were the other interactions. The main effects of the agent and the participant's performance show that both players needed to perform well in the game in order to succeed as a team. The undetectable interaction effect of participant and agent skill suggests no detectable further improving effect of a good player with a good agent or vice versa. The linear model is illustrated in Figure 21.

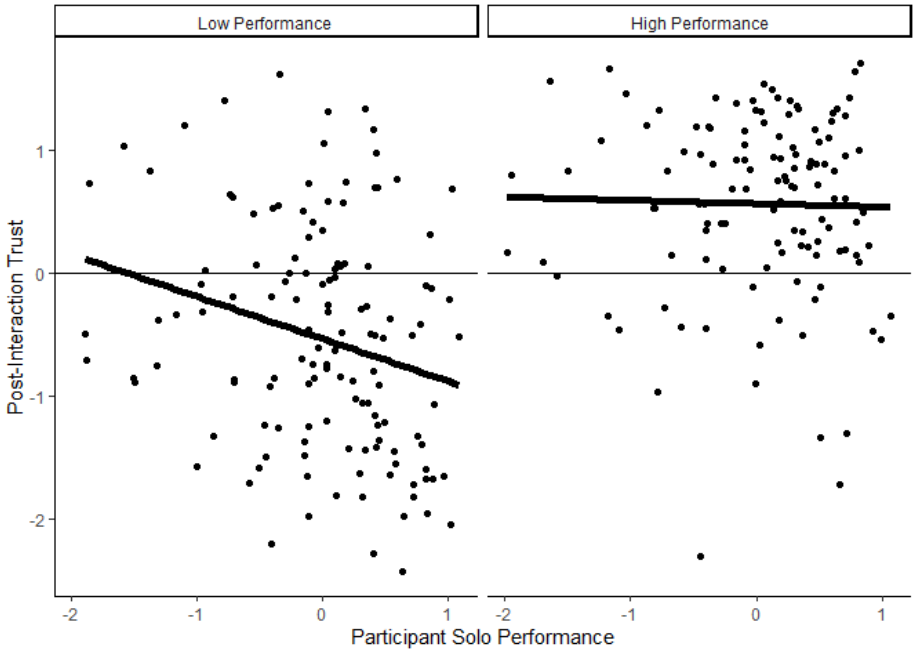
Performance Effects on Trust

We believe that a person's own expertise has a major effect on how that person judges the expertise of others. Looking at the performances of the participants and their post-interaction trust, we could see a result of this expertise effect. A good player might trust the agent less if they are still the better player and sees the flaws of the agent's game. On the other hand, a player with little understanding of the game will not be able to distinguish good from bad pieces given by the agent, and this would lead them to overtrust even bad agents. We test this idea with a linear regression on post-interaction trust. The predictors are agent performance and participant performance, which are controlled for reputation.

As we established in other models, the largest effect on post-interaction self-reported trust is the skill of the agent, $B = -1.10$, 95% CI [-1.30, -0.89], $t(248) = -10.55$, $p < .001$. Neither an effect of reputation, $B = -0.13$, 95% CI [-0.34, 0.07], $t(248) = -1.28$, $p = 0.202$, nor a main effect of the participant's solo performance is detectable, $B = -0.03$, 95% CI [-0.25, 0.19], $t(248) = -0.26$, $p = 0.794$. However, an interaction of the participant's solo performance and the agent's performance is statistically significant, $B = -0.32$, 95% CI [-0.63, -0.01], $t(248) = -2.05$, $p = 0.041$. This result suggests a moderating effect of the participant's performance on the effect of the agent's skill on post-interaction trust. As we exploratorily hypothesized, the direction of the effect suggests that better players judge the low-performing agent as less trustworthy than their worse-performing peers do (Figure 22).

Figure 22

Interaction Effect of Agent and Participant Performance on Post-Interaction Trust



Note. Participant performance was calculated as standardized movement speed and efficiency, sum of game overs, and completed rows in solo play.

Discussion

Trust is an important factor in human-agent teams. With this study, we are looking at the initial creation of trust through an agent's reputation and subsequently through its performance on a cooperative task. We measured trust via self-report and via delegation behavior, believing that initially, the trust towards the agent is dominated by reputation and then is continuously replaced by the experienced performance of the agent. The hypotheses were mostly confirmed with regard

to self-reported trust. However, the influence of reputation on the trust toward the agent was not detectable in the delegation to the agent.

Summary and Discussion of Results

In the self-report data, we identified a clear dependence of the participants' trust on the agent's reputation before they had any experience with it. The successful manipulation of trust through reputation is in line with existing research (Ullrich et al., 2021). However, as the cooperative task unfolded, the effect of reputation was no longer detectable, and the trust self-report was mostly dependent on the experienced performance of the agent. The expected interaction effect of time, reputation, and performance was not significant. Even in a similar analysis of the pre-post-differences, the interaction of reputation and performance was not significant. Yet, the t-test specifically comparing the magnitude of change in reputation-consistent to -inconsistent conditions showed significance and a considerable effect size. Thus, considering only self-reported trust data, H1a can be fully confirmed, whereas results for H2a are inconclusive. It is possible that the test of a three-way interaction was underpowered, explaining the discrepancy between the significant t-test of change and the non-significant interaction effect.

The experiment was also intended to give more context by investigating behavioral aspects of trust (delegation during the cooperative game). We expected the probability of delegating to the agent to reflect the a priori reputation of it initially and then, over time, to be continuously influenced by the participants' experience of the agents' performance. Yet, we neither detected a main effect of reputation nor an interaction with time nor a three-way interaction of reputation, performance, and time. Therefore, neither H1b) nor H2b) could be supported. With a more liberal use of directed hypothesis testing, the interaction effect of reputation and time would have reached significance. However, using our sample size and criteria, we could not detect an effect this small. Retrospectively, we have a possible explanation for this result. Considering that this scenario is low risk, and the bad performance of the agent would not lead to high costs, factors other than an agent's reputation might play a more critical role in the delegation decision

process. First, the instruction of the participants to use the agent at least ten times could have created an early tendency to use the agent despite a lower trust towards it. But more likely, curiosity was shown to be an important motivation in the intention to use autonomous technology in low-risk scenarios (Schwesig et al., 2023). This explanation is especially plausible since, in current times, people are constantly exposed to advertisements trying to influence their opinion toward a product, such as software, as in the case of this study. This habituation and the fact that the participants are aware of the experiment context might lead to low confidence in the description of the agent. The effects of risk and curiosity on trust and delegation could be investigated in further experiments by adding measures for curiosity or increasing the risk of the game by adding monetary incentives.

Even though we did not detect the effects of reputation, we did find significant and large effects of performance and its interaction with time on the delegation decisions during the experiment. We interpret these results in the sense that despite the large effect of reputation on self-reported trust, the participants practically disregard the prior information when it comes to the actual use of the agent. It indicates that the participants put a larger value on their own experience compared to the prior information from unknown sources. The effect of an agent's performance on a participant's trust in it is well documented (Daronnat et al., 2020; Kulms & Kopp, 2016; Robinette et al., 2017). However, to our knowledge, the dominance of one's own experience over the reputation has not yet been shown.

An exploratory analysis of the effect of the initial trust towards the agent on the delegation decisions suggested a moderating effect of a priori trust on the delegations over time that is different for a high vs. low-performing agent. A speculative explanation for this result, as opposed to the non-significant moderation effect of reputation, is that trust as a trait of the participants has a larger variance than the difference created by the manipulation of reputation (Hoff & Bashir, 2015). Thus, a highly trusting person might still have a lot of trust for the low reputation agent (and vice versa). Descriptively, it seems like a trusting person delegates more

over time if the agent is performing well and less if it is performing badly. On the other hand, untrusting persons do not change their delegation behavior over time for the well-performing agent but also decrease their delegations slower if the agent performs badly. This fascinating, albeit exploratory, result could be further investigated by not manipulating the trust through reputation but instead measuring the trait trust of participants.

Concerning the validity of self-report and behavioral measures of trust, the amount of delegation and the post-experimental self-reported trust show considerable correlation. This suggests that, while not perfectly, the behavioral and reported trust are connected. However, the self-reported trust before the interaction does not indicate delegation during the interaction. The relationship between delegation and post-interaction trust stays relevant even when controlling for pre-interaction trust levels. The correlation suggests validity of both measures. We conclude that delegation is a valid measure of trust within the scope of this experiment.

Exploratorily, we looked at the influence of the participants' proficiency in the game on their success in the cooperative game and the evaluation of the agent. We saw that both the participants' and the agents' performance influenced the team's success. Both influencing the result of the game suggests that the tasks were well distributed and allowed the participants to feel their own and the agent's influence in the success of the game. However, an amplifying effect was not detected. The bad performance of the agent did not significantly moderate the effect of the participant's performance on the game score. Nevertheless, the participant's performance moderated the effect of the agent's performance on the self-reported post-interaction trust. Lower-performing participants had more trust for the low-performing agent than their better-performing peers. Any explanation for this result is speculative since the analysis was exploratory and the effect is on the small side. A plausible explanation is that lower-performing participants lack the expertise to assess the agent's skill accurately and thus fail to calibrate their trust accurately. This implies that in order to calibrate one's trust towards an agent accurately, a good understanding of

the agent's task might be necessary. More research focused on these questions would be needed to confirm these ideas.

Implications of the Current Research

Trust is a widely used construct in the field of human-agent interaction (Kohn et al., 2021). It is also complex and abstract and, as such, impossible to measure directly. A person's self-report requires a deep understanding of their thoughts and feelings toward an agent. On the other hand, behavior is never solely controlled by a single theoretical construct. Ideally, both measures are connected. However, the complexity of trust as a construct and the multi-causality of behavior can lead to mismatches (Kohn et al., 2021). In this study, we found a correlation between the behavior and post-interaction self-report that suggests a moderate connection between them while still leaving room for other behavior-influencing factors. On the other hand, pre-interaction trust could not predict the number of delegations the participant would make. These results indicate a connection between trust and behavior but does not indicate equality. Yet, using pre-interaction trust we were able to detect the expected interaction effect with time on delegations that we did not find using reputation, the manipulated variable. This suggests that there is more variability in trust toward the agent than is created through the manipulation. Possibly, the participants' trait trust (specifically towards artificial agents), which was not measured in this study, has an influence on the self-reported trust and the delegations (Hoff & Bashir, 2015). A study adding a measure of trait trust is likely to shed light on the disconnect between reputation, self-reported trust, and delegation. The correlations and the regression model suggest a reliable connection between self-report trust and delegation behavior. We conclude that while trust certainly is not the only determining factor of a delegation decision, it is an important one. Thus, delegation behavior can be used as a valid measure of trust if other usage factors, such as curiosity, risk, or ease of use, are held equal or are controlled for.

Further, the research presented here can inform the marketing of novel agentic systems. We found that manipulating the agent's reputation had a large, significant effect on the

trust the participants reported. However, the usage of the agent was not significantly influenced by it. With agentic systems, the novelty of the product might already motivate users to try it, and not much convincing is needed. From the behavioral data, we could see that irrespective of the reputation, the use of the agent dropped immediately if the agent's performance was unacceptable. This dismissal of poorly performing agents was reflected in the participants' self-reported trust towards the agent. Research on algorithm aversion showed that a once-broken trust in an agent is more difficult to repair than a similarly broken trust toward a person (Dietvorst et al., 2015). It might be interesting to conduct further research on the long-term effects of a broken trust. The lack of a reputation effect on use suggests investing more resources into developing a polished agent than managing its public image. Once a potential user knows of a new development, they are motivated to test it in a safe environment, given the opportunity, despite a low expectation from the agent's reputation. Any further intention to use is only dependent on the actual performance of the agent.

Limitations and Scope

The scope is difficult to assess accurately. On the one hand, the game likely creates an intrinsic motivation in the participants to perform well. On the other hand, the novelty of interaction with an agent creates a possibly conflicting motivation to explore the agents' behavior. This experiment can certainly be generalized to other cooperative situations in which parts of tasks can be automated. But it is likely limited to low-risk situations in which the system can be tested without costly repercussions. In higher-risk situations, like medicine or military applications, the free testing of an automated system by the user might be hindered by safety concerns (Schwesig et al., 2023). Thus, the reputation might have a larger impact on the trust towards and early adoption of automated systems in high-risk situations.

The effect of reputation on pre-interaction trust was very large. While this seems like an indication of a well-working manipulation, it might also indicate a bias to please the researcher since the participants were aware of the experiment situation and might have also become

somewhat aware of the manipulation. In order to avoid this, we chose a between-subjects design so that the participants only see one description of an agent. However, it is difficult to say just how aware the participants were of their experiment's condition.

Conclusion

We conducted a multimethod experiment on the influence of an agent's reputation and performance on trust. Using two different measures of trust, we were able to understand how reputation and performance separately influence trust measured as self-report and then actual usage of the agent. The influence of the reputation on the use of an agent was smaller than we expected. Retrospectively, we explain this by the low-risk situation and the novelty of the agent, which naturally fosters the participant's exploratory behavior. The well-documented effects of an agent's performance on the trust towards it could be replicated. It seems like this effect far outweighs the effect of reputation. Even though the hypothesis on the initial influence of reputation could not be confirmed, we believe the additional analyses and discussion give valuable insights into the effects of reputation and, subsequently, of trust on behavior. While reputation has a significant effect on self-reported trust, it does not carry over to the delegation behavior. The pre-interaction trust is not only influenced by the reputation manipulation but also by the traits of the participant (Hoff & Bashir, 2015). The moderating effect on the delegations that we initially expected to come from the reputation was then found through the pre-interaction trust.

It is necessary to be aware of the complexity of human behavior. Even though trust is an important factor in people's behavior, it is not the only factor. Results from behavioral studies can be interpreted as trusting behavior but could be determined by a different factor, like a search for information or boredom. Self-report, on the other hand, has high face validity but can only detect the conscious reflections of a person's experience and must not necessarily have a consequence for their behavior. Effects of trust could be inflated if they are only detected through self-report, even though trusting the agent is not necessary in the situation. Researchers need to

consider these two points, especially when basing strong statements on only one type of measure.

Chapter 5 - General Discussion

Throughout three projects, I addressed how observers form a judgment of the expertise of artificial agents that is dependent on their task expertise. The dependence of these judgments on the observer's expertise is a novel approach that has not been extensively researched.

In the first project, I introduced the relevance of the observer's task knowledge to theory of mind research. While researchers working on classical theory of mind experiments, such as the false-belief task, usually assume a perfect understanding of the agent's task by the observer (Wellman et al., 2001), I focused on the variation in the observer's expertise in the experiments detailed in this thesis.

I collected data on the prediction of an agent's actions. The probability of a successful prediction depended on the expertise of both the agent and the observer. The prediction of a novice agent was impossible by design; it was an agent with the policy of choosing randomly. In this scenario, no degree of expertise would allow an observer to predict the agent's actions. The prediction of the actions of an expert agent was possible but depended on the participant's expertise. Novice participants improved over time at prediction, concurrently with improvement at the task. Task experts, in contrast, did not improve; they predicted at the highest level throughout the complete experiment. Thus, they did not exhibit mentalizing of the agent decision process beyond finding a piece that fit. Their early prediction of pieces that fit well implies that they assumed the agent was cooperative and had some degree of expertise in the game. The lack of improvement suggests that the participants did not intensively mentalize the process of the agent and instead used their knowledge of the task to predict what any rational agent would choose. Consequently, the expert participants must have quickly understood that the smart agent was able and willing to cooperate but did not further understand the specific decision process of the agent. The perception of the agent's ability and expertise must have played a role for the observer to predict the agent's actions.

I investigated the evaluation of an agent's expertise by an observer in the next project directly. I conducted an experiment and a quasi-experiment with participants having varied task expertise. The participants were asked to evaluate the expertise of agents with varying expertise based on their task performance. The results showed that participants with little expertise were less accurate than those with high levels of expertise. Within the strict experimental constraints of the experiment in the first study, they were highly accurate when the agents knew less than they did. If the agent knew more, the participants had to guess the degree to which the agent's expertise exceeded their own, and thus their ratings became less accurate. In this case, the participants tended to rate the agent's expertise close to theirs. The first experiment's results were not very ecologically valid, so I replicated the results in a more widely applicable scenario with general knowledge questions. Like in the first experiment, the results showed the benefits of expertise in evaluating and interpreting an agent's behavior.

Since the perceived expertise of an agent strongly influences the trust that an observer places in it, trust was the target of the third project (Hafizoğlu & Sen, 2018b; Lucassen & Schraagen, 2011; R. C. Mayer et al., 1995). I varied the agent's performance in the same way that I did in the first study of Chapter 2. In addition to the expert and novice performance of the agent, I manipulated the participants' expectations by giving them second-hand information about the agent's reputation. I used two different measures of trust. I employed pre- and post-experimental self-report questionnaires as is common in research on trust (Kohn et al., 2021). Additionally, to understand the development of trust throughout the experiment, I measured the delegations the participant made to the agent as a behavioral measure of trust. The agent's reputation had a large effect on the self-reported trust of the observer before the experiment. However, the effect of reputation on the behavioral measure of trust was undetectable during the experiment. Following the experiment, self-reported trust was influenced by the experience of the interaction with the agent, and the effect of the agent's reputation was no longer detectable. The results suggest that the reputation of an agent can manipulate trust. However, the use of the agent is not meaningfully

influenced by it. Rather, the performance of the agent as experienced first-hand through interaction influences its use and also trust in it going forward. The results concerning pre-experiment trust and delegation are somewhat inconsistent, since the effect on self-reported trust is not reflected in the participant's behavior. The inconsistency can be explained by the fact that the decision to delegate depends not only on the participant's trust but also on the unmeasured perceived risk of bad decisions and the curiosity of the participant.

The projects provide strong evidence that there is a limit to social perception that originates from an observer's task knowledge. A person's domain expertise constrains the accuracy of agent perception within that domain. Objectively, an agent's performance holds extensive information about its abilities if the agent does not try to hide the information. The information is usable by an observer with the same or more task expertise. However, the performance of an expert is uninformative for someone with low expertise. I have provided empirical support for this theoretical statement in three projects on the social perception of an agent's actions, expertise, and trustworthiness. In the following sections, I will discuss the generalizability of the effect, its relevance for the field of research, and the possible consequences for application. I will further raise research questions that are yet to be answered.

Theoretical Relevance

This dissertation addresses how people assess an agent's expertise, actions, and trustworthiness. Such interpersonal assessments are highly relevant for successful collaboration and communication between agents (Barsalou, 2008; Lucassen & Schraagen, 2011; R. C. Mayer et al., 1995). The agency of autonomous computer systems is increasing and will likely continue to increase. To look at how people assess these aspects of artificial agents is a relevant and future-guiding task (Gunning et al., 2019; Heyselaar, 2023).

This dissertation's main hypotheses and potentially most exciting contributions are derived from a simple truth: People cannot understand decisions based on unavailable information. The logical consequences of this statement are the hypotheses that novices

misestimate expertise and cannot anticipate an expert's decision. The statement itself could be criticized as a truism that does not need empirical evidence. However, my work gives insight into the limits of social perception beyond the simple truth they are based on.

I have shown in the projects completed for this dissertation that novices' lack of task knowledge limits the types of information that they can interpret. The ability to interpret semantic features of behavior is limited for novices, so they must rely on source and surface features that are potentially unavailable or misleading (Lucassen & Schraagen, 2011). Although semantic features of behavior can also be manipulated to mislead an observer, the agent has to perform below their capacity and accept a less beneficial state of the environment as a consequence. If they do so, they must accept worsening their position to misdirect an observer. The opposite, for the agent to perform better than their ability allows, is logically impossible since their ability is defined by their best possible performance.

Even if, to some readers, the results are apparent consequences of a trivial truth, I doubt that the conclusions I draw from the results would have been accepted without the empirical work. Other readers may be surprised by the results since considerable research has shown that novices, frequently children, can judge the abilities of others despite their lack of understanding of the task they are observing (Birch et al., 2010; Brosseau-Liard & Poulin-Dubois, 2014; Jaswal & Malone, 2007; Matsui et al., 2016; Sabbagh & Baldwin, 2001). For example, in one experiment, 4-year-old children learned novel words of objects better from a teacher who said that they were the creator of the object (expert) as opposed to saying they are a friend of the maker (novice) (Sabbagh & Baldwin, 2001). In comparison, 3-year-olds did not exhibit a preference for learning from experts. The authors cite theory of mind as the driving factor for the difference (Sabbagh & Baldwin, 2001). This experiment showed that even children can perceive the expertise of agents. However, they did not base their judgment on the semantic cues of agents' behavior but instead on the source cues (maker or friend) and perhaps surface cues such as the confidence and tone of the speaker. Therefore, the experiment does not contradict the claims of the present

dissertation but rather highlights the importance of distinguishing between the type of information that is interpretable by novices and the type of information they cannot use because of their limited knowledge.

In another experiment on the perception of expertise, laypersons could estimate the relevance of different disciplinary expertise on a specific topic. The participants were informed about an agent's area of expertise and could estimate how relevant the expertise was to the topic (Bromme & Thomm, 2016). Unlike in my research, the participants were told the agent's area of expertise (source information) rather than shown the agent's performance (semantic information). Therefore, the results are not fully comparable to those presented here. This thesis goes beyond the earlier findings by Bromme and Thomm (2016) by moving the focus from third-party source information to semantic information of observable behavior.

In conclusion, the arguments I present in this dissertation are not self-evident, given the results relating to social perception that show that novices can estimate an agent's expertise with considerable accuracy. Yet they do not conflict with the existing research because the information the observers could use for their judgment of expertise was qualitatively different than the information provided in comparable research. Given only the products of an agent's efforts and lacking source information about the agent and information about the task, an observer cannot evaluate the correctness of the agent's action and, by extension, cannot accurately judge the agent's expertise. Whether the effect I describe is trivially true or requires more empirical research is debatable. However, misjudging an agent's expertise and consequently miscalibrating one's trust could impact society severely, even if it only happens in niche circumstances. Work, education, and human-machine teams are likely affected by the effects presented in this dissertation, as I will explain in more detail in the next section.

Application

The conclusions drawn from the empirical work in this dissertation have societal consequences and applications in the fields of work, education, and human-computer interaction.

However, transferring the results is not possible without restrictions. Within the constraints of the studies, the expertise of the agents was known and fixed. In more complex situations, the true expertise of agents is usually unknown and changes over time. Crucial information can be learned or forgotten. The applications I discuss are thus somewhat speculative, given the added complexity of the real world that was not accounted for in the studies.

Societal Impact

The work I present deals with interactions between *two* agents. Societies are not the focus of this dissertation. However, societies emerge from the actions and interactions of the society's agents (Castellano et al., 2009). In this paragraph, I discuss how the limitations on social perception that derive from imbalances in expertise in single interactions can affect society as a whole. I have established that in certain situations with unavailable or unreliable source and surface cues, people must interpret semantic cues from texts and behavior to judge the agent's expertise and, ultimately, trustworthiness. Usually, all types of cues are available and sufficiently reliable in face-to-face interactions. But there are exceptions. For example, in computer-mediated interactions, the anonymity of other agents obscures the source and surface cues. In such situations, observers can only judge the expertise of an agent based on the semantic cues offered by the agent's observable behavior. As I have shown, the accuracy of expertise judgments is dependent on the expertise of the observer (Chapter 3). Furthermore, an observer's judgment of an agent's trustworthiness and their willingness to depend on it are predicted by their perceptions of the expertise of this agent (Chapter 4).

An example of such a situation would be watching footage of somebody playing a video game. Without knowing the player, a judgment of the player's skills could only be made based on the footage. The footage has surface information, such as the video quality, but it is questionable how informative it is about the player's skill. In the end, the observer has to decide if they should accept advice from the player based on the footage they see, which they can only evaluate using

their understanding of the game mechanics. Misjudging the expertise of the agent could lead them to allocate trust poorly, become misinformed, and adopt policies that would harm their game.

In video games, the results of misallocating trust can be frustrating for the individual but do not have severe consequences for society. Agents in other areas of expertise, such as science, medicine, engineering, and politics, guide policies for whole societies. In these areas, the consequences of misinformation through misjudgment of an agent's expertise could be severe. A single policymaker could make suboptimal decisions because they misjudge the expertise of a counselor. Whether any low-expertise observer in society trusts people of high or low expertise would become random. This would lead to a society in which all experts and random non-experts agree on a topic while other non-experts randomly hold other beliefs.

This is certainly not the only process that leads to divergent opinions in societies. In-group and out-group perception, motivated reasoning, and other social processes play a more prominent role. However, the dissertation results present a new angle through which societal divisions could be explained.

Task Delegation in Work and Education

Expertise perception plays a vital role in work and education contexts. The difference in expertise between teacher and student is necessary in formal education. In informal learning scenarios, expertise perception can allow for quick allocation of teacher and student roles. However, in some work scenarios, the dynamics of expertise can create challenges. Imagine you are hiring a new programmer for your company. You are a successful businessperson, but you are not a programmer. The only information about the applicants you get is the code they have written in the past. How would you make a judgment about the expertise of the applicants?

For a company, hiring new people is a critical decision (Lievens et al., 2021). The person making hiring decisions is rarely an expert in the field for which the company is hiring. Nevertheless, the person has to decide based on their perception of the applicant's expertise. In the work presented here, I have shown that a non-expert hiring person should not be able to make

accurate hiring decisions based only on the semantic features of the applicant. However, our society has found ways to alleviate the risk for the company. Field experts who trained the applicant certified the latter's ability, and standardized exams created by experts make applicants comparable. The information from these sources allows the hiring person to make accurate hiring decisions despite being unable to evaluate the performance of an applicant independently. Hiring decisions are more complex than just evaluating the applicants' expertise. Other factors, such as personality and cost, play a role. However, this example illustrates how society is implicitly aware of the problem of expertise perception.

The second application in the work context of the results presented here is in the distribution of tasks as part of a leadership role. Ideally, the person distributing tasks is a task expert. Their expertise allows them to assess the expertise of their coworkers accurately. Ironically, they could do the task better than their coworkers since task distributors should be of higher expertise. This paradox exists superficially, as the delegation of tasks can challenge and strain less experienced coworkers or clear the backlog of tasks for the task expert, even if the task is fulfilled less efficiently. The task distributor then has to weigh the value of their own time against their trust in the task performance of their coworker.

Having a task layperson supervise an expert at that task can lead to greater difficulties. In Chapter 2, I reported novices' inability to accurately assess a more competent person's expertise. A supervisor misjudging a worker's abilities is enough to lead to imperfectly calibrated task delegation. Additionally, the supervisor cannot assess the difficulty of the delegated tasks and the quality of the solutions produced. These problems can be solved through constant communication and the building of mutual trust.

In formal teaching settings, the teacher is a task expert, and the student strives to become one. The teachers' greater expertise is necessary to assess the student's current level of task knowledge accurately. However, it is not sufficient to take the student's perspective; the teacher needs skills beyond mere task expertise. In situations in which the roles of teacher and

student are not institutionally defined, the agents might need to evaluate each other's knowledge dynamically. If one agent makes a decision, the other can evaluate whether and in which way that decision was optimal or less then optimal and, based on this perception, adjust their assessment of the other's expertise at the task. If the observer finds an error in the decision and is confident in their abilities, they might be able to teach the other agent something.

The roles remain unclear if the observer does not feel confident and cannot evaluate the agent's decision. The agent may be better at the task, but it is also possible that they too are guessing. This misunderstanding can happen, for example, when two academics visit a conference in an unknown city and do not communicate about who navigates. Both may assume the other has researched the way to the hotel and so they follow each other, reacting to meaningless movements as if they are directions. Since both trust the other to navigate, they ignore where they are, and it takes a long time to notice their error. A difference in expertise would have helped to clarify the roles. Even if the expert initially follows the novice, they would be likely to notice the errors quickly.

Human-Machine Interaction

A core application of the work presented here is human interactions with machines (including, more specifically, computers). The empirical results show that trust calibration depends on the perceived expertise of the agent. The accuracy of expertise perception, however, depends on the relative expertise of the observer. These results have implications for the interaction with super-human agent technology.

Based on the results I have presented, human experts can still evaluate the sub-human performance of artificial agents. Even if their decision processes are different, the human can grasp the agent's expertise and calibrate their trust accordingly if the two have the same information. As soon as artificial agents outperform even human experts, it becomes more difficult to accurately calibrate trust in the machine since nobody can evaluate the agent's expertise by observing the agent's performance. In areas in which trust is critical, and one cannot just look at

the outcome, such as medicine, economy, and politics, such uncertainty can have severe consequences. Scientists in the field of explainable AI research ways to make artificial agents understandable to humans. The research presented here underlines the importance of these endeavors and highlights that humans need explainable AI, especially in areas in which their expertise is limited. Therefore, it would be helpful to adapt the explanations given for AI output to the level of the observer's expertise.

In chess, for example, computers already far outperform even the best grandmasters of the game. The outcomes of chess games have no severe societal consequences, making chess a perfect model for researching human interactions with the super-human performance of artificial agents (McIlroy-Young et al., 2020). Human players can use so-called chess engines, artificial agents for playing chess, to evaluate their game, using the agent's choice as the indicator of the "correct" choice, even though the game is not solved in the sense that optimal decisions are not mathematically proven. Before the rise of chess engines, the correctness of a choice was up for discussion among players of similar levels and a matter of authority in interactions between players of higher and lower ratings. Today, such arguments can be cut short by referencing the highest authority, the chess engine. Implementing artificial agents in chess has led to a deeper understanding of the game and a shift in how individuals approach the game. Another consequence is that in computer chess—that is, chess games between chess engines—some of the decisions of the engines are inexplicable to the highest human experts of the game.

My research can gives a framework to the dynamics of trust and expertise effects in areas in which artificial agents are at the level of expert humans or above. The chess example illustrates how trust shifted from human to artificial authorities and how some decisions of superhuman artificial agents are inexplicable to humans at any level of expertise.

Scope and Limitations

In the following paragraphs, I will discuss the generalizability of the results. In all of the experiments detailed in this dissertation, the participants interacted with or perceived the actions

of an artificial agent. This fact leads to a question: Do the effects I report here translate to interactions between humans and humans or even to humans and any agent? After discussing the question, I present the three major limitations I identified in the present work. The experiments happened on the computer, with information about the agent limited by design and a specific cooperative scenario. These decisions limit the interpretability of the results to similarly computer-mediated, cooperative situations with information limited to semantic cues from the agent's behavior.

Generalizability to Human-Human Interaction

I have discussed the transferability of social cognitive theory to machines in the introductory subsection, "Theory of Machine Mind". There, I reviewed the transferability of results in theory of mind research to interactions with artificial agents. The two main arguments also apply to the transferability of the results of this dissertation to interactions between humans.

First, the CASA paradigm states that machines are regularly treated as humans would be (Gambino et al., 2020; Nass et al., 1994; Nass & Moon, 2000). It suggests the possibility that machines, especially novel agents, and humans are treated similarly in a human mental representation. However, the paradigm does not mention the inverse case, which is that human interaction can be researched using artificial agents. Furthermore, it has been demonstrated that machines are treated differently from humans when compared directly to them. For example, an aversion to relying on algorithmic decisions after seeing them result in errors has been shown (Burton et al., 2020; Dietvorst et al., 2015). Also, people feel less guilty when exploiting machines than humans (Melo et al., 2016). Therefore, artificial agents might be unsuitable for research on human interaction.

Second, stimuli such as avatars, puppets, videos, and stories of people that do not, strictly speaking, involve direct human-human interaction are regularly used in social cognition research (i.e., Low & Watts, 2013; Newton & de Villiers, 2007; Van der Wel et al., 2014). Results from such studies are generally accepted as representative of human-human interaction.

Compared to the use of stories or videos, my use of artificial agents should generalize even better to human-human interaction due to the interactivity of the agent and the observer. The observer can only passively use the information given to them when presented with stories and videos; to some degree, they are communications from the experimenter to the participant. In contrast, a participant who interacts with agents can search for information about the agent by challenging and testing it. For these reasons, I argue that using agents in experiments leads to more ecologically valid results than using fixed materials such as texts, pictures, or videos.

Whether research on interactions with artificial agents generalizes well to human-human interaction cannot be fully answered; it might even depend on the parameters of an experiment. If it does not generalize, the interpretation of the results from these projects would be limited to artificial agents only and maybe even to a type of artificial agent close to the ones I used. However, as I mentioned frequently throughout this work, artificial agents are rapidly entering human environments. Even if one cannot accept the generalization of the results to interactions between humans, the findings have vital implications for interactions with artificial agents now and in the future (see the Section “Application”).

Limitation to Computer-Mediated Interaction

Another limitation is the reliance on computer-mediated interactions in all three studies; all experiments were conducted on the Internet using a computer. According to media richness theory, the quality of interactions changes with the mode of communication, with face-to-face interactions containing the highest richness and written communications, such as e-mails or letters, the lowest (Daft & Lengel, 1986; Hiltz et al., 1986). Computer-mediated interactions like those used in the experiments in this thesis fall into the lower third of media richness. Previous empirical results on media richness show that team decision accuracy, for example, is higher in face-to-face than in computer-mediated communication (Hedlund et al., 1998). Face-to-face communication changes the quality of the interaction and makes more information available to the observer. A change to richer communication makes information available that is unavailable in

most computer-mediated interactions, such as the physical appearance of the agent and the time it takes to make decisions.

Limitation to Semantic Information

Throughout the dissertation, I used the 3S model to categorize the information about an agent's expertise into source, surface, and semantic cues (Lucassen & Schraagen, 2011). I focused the research on semantic information while controlling the available source and surface information as I believe the semantic cues to be the most reliable source of information. However, the source and surface cues likely influence an observer's judgment of expertise. Specifically, the interaction effect of observer and agent expertise is probably influenced by the type of information given to the observer. Depending on the information's type and quality, the effects of observer expertise on the accuracy of agent expertise estimation could either strengthen or be overshadowed by other effects. If, for example, the observers get reliable information about the agent's reputation, the expertise ratings by novice observers might become more accurate. However, the effect of unreliable prior information on the agent is difficult to predict. The novice observers might get even worse if they rely on unreliable information, and expert observers could stay accurate if they managed to ignore the reputation information and base their ratings on their experience of the interaction or concrete observation (although this might be more effortful for the observer). However, if experts cannot ignore the more readily available source information, they might be influenced by unreliable but accessible information. Reasons to rely on less reliable information could be time pressure or cognitive load as the inference of expertise from semantic information is effortful.

To conclude, limiting my investigation to semantic information was a conscious decision to be able to research its effects in isolation. Nevertheless, the decision means that the results have limited applicability to natural situations in which additional information is available. Further research on the interaction of more sources of information with varying reliability would give more insight. However, the added variables would make controlling the experiments more difficult.

Limitation to Cooperative Scenarios

A final limitation of the research I presented is that I assume that the agent wants to be predictable or at least does not actively want to deceive the observer about their ability. In the experiments that were presented, the agents performed the tasks alone, with observation of the results (Chapter 3), or cooperatively with the observer (Chapters 2 and 4). Alone, the agent does not have to worry about someone else's gains; in cooperative scenarios, the gains of any agent are the gains of the team. As a consequence, the actions they choose represent their best efforts, which makes them predictable. However, predictability is a disadvantage in competitive situations. For example, in zero-sum games, the gains of one agent equal the losses of another. Optimal strategies in these scenarios do not only involve maximizing one's gains but also minimizing the possible gains of the opponent (Fan, 1953; Van Neumann, 1928). This includes sacrificing one's gains to become less predictable and give the opponent less opportunity to prepare. Competitive scenarios add a layer of deception that I have not accounted for in the relationship between agent and observer expertise. Future experiments could investigate the trade-off between predictability and optimal decision-making in competitive scenarios.

Verdict on Scope

In conclusion, I would argue that the effects I presented in this dissertation can likely be replicated in computer-mediated human cooperation. However, the effects could be moderated by the stereotypes and prior expectations that people have of artificial agents instead of humans. Furthermore, the conclusions drawn from the projects conducted for this dissertation are limited to information from the semantic features of observed actions, with no intent to deceive by the agent. These limitations can be adressed in future research projects on the significance of expertise differences for social perception.

Outlook

I have already briefly suggested ways one could build on the work presented in this dissertation to get more generalizable results and how the results could be applied in other fields.

In this section, I will discuss how future research could build upon the findings of this dissertation. I suggest a direct investigation of the effect of expertise differences on trust, adding moderators, such as different expertise cues, to get further applicable results, and simulating the effect the limitation of social perception has on a societal level.

Trust

I have shown how trusting an artificial agent is dependent on how the expertise of the agent is perceived. The observer's (trustor's) predisposition to trust and the agent's (trustee's) benevolence, ability, and integrity are well-known determinants of trust (R. C. Mayer et al., 1995). In this dissertation, I established that the perception of expertise (ability) depends on the observer's expertise (Chapter 3). Together, these results suggest that the formation of trust in an agent is dependent on the expertise of the observer. However, this hypothesis was not explicitly tested through dedicated manipulation of observer expertise in a trust experiment. The agent's expertise was manipulated in the experiment using trust as an independent variable, but the observer's expertise was not (Chapter 4). The observer's expertise varied naturally through different experiences and the effectivity of the training, but it was not manipulated by different trainings or measured elaborately. Indirect measures of the observer's expertise suggested an effect of observer expertise on trust formation. The hypothesis could be explicitly tested by combining the designs of the experiments presented in this dissertation. One could manipulate the participants' expertise, as in Study 2 of Chapter 2 or Study 1 in Chapter 3, and measure the self-reported trust and possibly the delegations, as in Chapter 4. Using such a design, one could directly test and see the effect of observer expertise on self-reported trust as well as objectively measure delegation behavior as an additional assessment of trust in the agent.

Moderators

In the experiments of this dissertation in which the observer has to judge the agent's expertise, the information the observers get about the agent is limited to semantic cues. Information regarding the agent (source cues) and style (surface cues) was kept uninformative to

isolate the effect of semantic cues. However, these information sources likely influence expertise judgments, too (Lucassen & Schraagen, 2011). I hypothesize that novices especially rely on information other than semantic cues to make expertise judgments since they cannot evaluate semantic cues, as I have shown in Chapter 3. Conversely, experts can use all information but might make different choices under different circumstances. Semantic cues are the most reliable information about an agent's expertise. Nevertheless, evaluating the agent's decision is likely more effortful than evaluating source or surface cues. Therefore, experts might depend on less reliable information to make expertise judgments when under time pressure or experiencing a heavy cognitive load. Also, in experiments where source and surface cues prove reliable and the risk of misjudging an agent's expertise is low, experts might refrain from evaluating the semantic cues to save time and mental effort.

In Chapter 4, I added reputation, a source cue, as a moderator. While it did have a large effect on self-reported trust, it did not meaningfully influence the behavioral measure of trust in the agent. In this experiment, all observers had similar expertise, and the risk that would result from failure was low, so the observers were equally able to test the agent and form judgments based on the agent's performance rather than its reputation. Adding relevant consequences to bad decisions for the observer might change the results. An observer that fears a bad decision of the agent might be less willing to test its abilities out of curiosity.

To summarize, source and surface cues, in combination with circumstances of experiments such as time, risk, and the reliability of information would likely moderate the effects I have presented in this dissertation. It would require large projects with many experiments to test all my suggested hypotheses, but they will likely reward the experimenters with exciting results.

Societal Impact

I have already laid out how I believe that misjudgments based on expertise differences influence dynamics on a societal level (see "Application – Societal Impact"). Hypotheses regarding society are difficult to test experimentally, and surveys are susceptible to sampling effects and

depend on the introspection of the participants. Multi-agent modeling allows testing the effect of limitations on social perception in individual interactions on a broader scale. A major advantage is that the perception effect can be manipulated in isolation, so the observed consequences stem purely from the theorized effect. However, this purity comes with the challenge that the researchers have to make many assumptions about the nature of the variables and their dependency. Disagreement about these assumptions leads to invalidation of the observation and questions about its applicability to real societies.

To model the social perception limitation, I suggest a multi-agent model in which agents can present true and false information to other agents. In th model, the agents should vary in expertise and should not be able to accurately evaluate information that is too complex for their expertise. In turn, they will only present information to other agents if they evaluate it as true. In this model society, the experts could function as a filter to block false information from moving on. However, some agents will believe in false information by chance and repeat it to others. This model could simulate how some beliefs take hold in parts of society despite the overwhelming evidence presented by experts, although it considers only the limitations on social perception discussed in this dissertation. In reality, other factors, such as group mentality and motivated reasoning, might have a similar or more powerful effect on this phenomenon.

Conclusion

In this dissertation, I presented three projects on the social perception of expertise and trust. This work fills a research gap in social perception as the task knowledge of an observer is rarely considered. I have found that agents performing a task to the best of their abilities give information about their expertise to observers. However, the observer needs similar or greater expertise to interpret the cues from the agent's performance accurately. With an accurate assessment of the agent's expertise and expert understanding of the agent's task, the observer can predict the agent's actions and calibrate their trust in the agent accurately.

I discussed the limitations of the results to semantic information about the agent's actions and computer-mediated situations in which deception is not in the agent's interest. These limitations might suggest that the research only applies to very niche scenarios. However, its applications are very diverse; communication between humans and other humans, or artificial agents is often mediated by computers, and mistrust is widespread in online communication. Furthermore, hierarchies in schools and workplaces are designed around agents' expertise. Misunderstanding of the environment can lead to mistrust in authority figures. Teachers and managers must choose tasks for their subordinates carefully and challenge them without discouraging them.

Finally, I proposed experiments for extending the expertise inequality effect that I have described to trust calibration and adding moderators such as additional agent information. Investigating the consequences of the expertise effects in pairs of agents in a multi-agent model could lead to interesting results about the polarisation of society. Together with replications of the effects of my projects, these experiments could secure and extend the claims of this work.

With my dissertation, I have contributed to understanding the impact of domain expertise on social perception. I focused on the perception of artificial agents, which will increasingly influence human environments. To understand the impact of new agents—and eventually super-human agents—it is necessary to understand how a human's limited understanding of their environment affects their perception of other agents.

References

Abele, A. E., Hauke, N., Peters, K., Louvet, E., Szymkow, A., & Duan, Y. (2016). Facets of the Fundamental Content Dimensions: Agency with Competence and Assertiveness—Communion with Warmth and Morality. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01810

Aboody, R., Davis, I., Dunham, Y., & Jara-Ettinger, J. (2021). *I can tell you know a lot, although I'm not sure what: Modeling broad epistemic inference from minimal action*. https://doi.org/10.31234/osf.io/uymtz

Al Shamsi, J. H., Al-Emran, M., & Shaalan, K. (2022). Understanding key drivers affecting students' use of artificial intelligence-based voice assistants. *Education and Information Technologies*, *27*(6), 8071–8091. https://doi.org/10.1007/s10639-022-10947-3

Alarcon, G. M., Gibson, A. M., Walter, C., Gamble, R. F., Ryan, T. J., Jessup, S. A., Boyd, B. E., & Capiola, A. (2020). Trust perceptions of metadata in open-source software: The role of performance and reputation. *Systems*, *8*(3), 1–14. https://doi.org/10.3390/systems8030028

Andrews, K. (2008). It's in your nature: A pluralistic folk psychology. *Synthese*, *165*(1), 13–29. https://doi.org/10.1007/s11229-007-9230-5

Apperly, I. A. (2008). Beyond simulation–theory and theory–theory: Why social cognitive neuroscience should use its own concepts to study "theory of mind." *Cognition*, *107*(1), 266–283. https://doi.org/10.1016/j.cognition.2007.07.019

Apperly, I. A. (2011). *Mindreaders: The cognitive basis of "theory of mind."* Psychology Press.

Bakan, D. (1966). *The duality of human existence: An essay on psychology and religion* (p. 242). Rand Mcnally.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064. https://doi.org/10.1038/s41562-017-0064

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349. https://doi.org/10.1016/j.cognition.2009.07.005

Baltieri, M., Iizuka, H., Witkowski, O., Sinapayen, L., & Suzuki, K. (2023). Hybrid life: Integrating biological, artificial, and cognitive systems. *WIREs Cognitive Science*, e1662. https://doi.org/10.1002/wcs.1662

Barandiaran, X. E., Di Paolo, E., & Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behavior*, *17*(5), 367–386. https://doi.org/10.1177/1059712309343819

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a “theory of mind” ? *Cognition*, *21*(1), 37–46. https://doi.org/10.1016/0010-0277(85)90022-8

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology*, *59*(1), 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Beer, R. D. (1995). A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence*, *72*(1–2), 173–215.

Berke, M., & Jara-Ettinger, J. (2021). *Thinking about thinking through inverse reasoning*. PsyArXiv. https://doi.org/10.31234/osf.io/r25qn

Berke, M., & Jara-Ettinger, J. (2022). Integrating experience into bayesian theory of mind. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44). https://escholarship.org/uc/item/8397z6nx

Berry, D. C., & Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. *The Quarterly Journal of Experimental Psychology Section A*, *36*(2), 209–231. https://doi.org/10.1080/14640748408402156

Birch, S. A. J. (2005). When knowledge is a curse: Children's and adults' reasoning about mental states. *Current Directions in Psychological Science*, *14*(1), 25–29. https://doi.org/10.1111/j.0963-7214.2005.00328.x

Birch, S. A. J., Akmal, N., & Frampton, K. L. (2010). Two-year-olds are vigilant of others' non-verbal cues to credibility. *Developmental Science*, *13*(2), 363–369. https://doi.org/10.1111/j.1467-7687.2009.00906.x

Bishop, D., & Adams, C. (1991). What do referential tasks measure? A study of children with specific language impairment. *Applied Psycholinguistics*, *12*, 199–215. https://doi.org/10.1017/S0142716400009140

Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, *77*(1), B25–B31. https://doi.org/10.1016/S0010-0277(00)00096-2

Bonezzi, A., Ostinelli, M., & Melzner, J. (2022). The human black-box: The illusion of understanding human better than algorithmic decision-making. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001181

Bower, A. H., Han, N., Soni, A., Eckstein, M. P., & Steyvers, M. (2024). How experts and novices judge other people's knowledgeability from language use. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-023-02433-9

Bromme, R., Rambow, R., & Nückles, M. (2001). Expertise and estimating what other people know: The influence of professional experience and type of knowledge. *Journal of Experimental Psychology: Applied*, *7*(4), 317–330. https://doi.org/10.1037/1076-898X.7.4.317

Bromme, R., & Thomm, E. (2016). Knowing Who Knows: Laypersons' Capabilities to Judge Experts' Pertinence for Science Topics. *Cognitive Science*, *40*(1), 241–252. https://doi.org/10.1111/cogs.12252

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American Psychologist*, *32*(7), 513–531. https://doi.org/10.1037/0003-066X.32.7.513

Brosseau-Liard, P. E., & Poulin-Dubois, D. (2014). Sensitivity to confidence cues increases during the second year of life. *Infancy*, *19*(5), 461–475. https://doi.org/10.1111/infa.12056

Bubic, A., Von Cramon, D. Y., & Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience*, *4*. https://doi.org/10.3389/fnhum.2010.00025

Buchholz, V., Kulms, P., & Kopp, S. (2017). It's (not) your fault! Blame and trust repair in human-agent cooperation. *Kognitive Systeme Workshop*. https://doi.org/10.17185/duepublico/44538

Buder, J., Becker, F., Bareiß, J., & Huff, M. (2024). Beyond Mere Algorithm Aversion: Are Judgments About Computer Agents More Variable? *Communication Research*. https://doi.org/10.1177/00936502241303588

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239. https://doi.org/10.1002/bdm.2155

Cancro, G. J., Pan, S., & Foulds, J. (2022). *Tell me something that will help me trust you: A survey of trust calibration in human-agent interaction* (arXiv:2205.02987). arXiv. http://arxiv.org/abs/2205.02987

Candrian, C., & Scherer, A. (2022). Rise of the machines: Delegating decisions to autonomous AI. *Computers in Human Behavior*, *134*. https://doi.org/10.1016/j.chb.2022.107308

Castellano, C., Fortunato, S., & Loreto, V. (2009). Statistical physics of social dynamics. *Reviews of Modern Physics*, *81*(2), 591–646. https://doi.org/10.1103/RevModPhys.81.591

Chiou, E. K., & Lee, J. D. (2016). Cooperation in human-agent systems to support resilience. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *58*(6), 846–863. https://doi.org/10.1177/0018720816649094

Chiou, E. K., & Lee, J. D. (2023). Trusting automation: Designing for responsivity and resilience. *Human Factors*, *65*(1), 137–165. https://doi.org/10.1177/00187208211009995

Churchland, P. M. (1989). Folk psychology and the explanation of human behavior. *Philosophical Perspectives*, *3*.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Coane, J. H., & Umanath, S. (2021). A database of general knowledge question performance in older adults. *Behavior Research Methods*, *53*(1), 415–429. https://doi.org/10.3758/s13428-020-01493-2

Collins, H., & Evans, R. (2009). *Rethinking expertise*. University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/R/bo5485769.html

Conway, C. M., Bauernschmidt, A., Huang, S. S., & Pisoni, D. B. (2010). Implicit statistical learning in language processing: Word predictability is the key. *Cognition*, *114*(3), 356–371. https://doi.org/10.1016/j.cognition.2009.10.009

Correia, F., Guerra, C., Mascarenhas, S., Melo, F. S., & Paiva, A. (2018). Exploring the Impact of Fault Justification in Human-Robot Trust. *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, 507–513.

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and Competence as Universal Dimensions of Social Perception: The Stereotype Content Model and the BIAS Map. In *Advances in Experimental Social Psychology* (Vol. 40, pp. 61–149). Elsevier. https://doi.org/10.1016/S0065-2601(07)00002-0

Cuddy, A. J. C., Fiske, S. T., Kwan, V. S. Y., Glick, P., Demoulin, S., Leyens, J.-P., Bond, M. H., Croizet, J.-C., Ellemers, N., Sleebos, E., Htun, T. T., Kim, H.-J., Maio, G., Perry, J., Petkova, K., Todorov, V., Rodríguez-Bailón, R., Morales, E., Moya, M., … Ziegler, R. (2009). Stereotype content model across cultures: Towards universal similarities and

some differences. *British Journal of Social Psychology*, *48*(1), 1–33. https://doi.org/10.1348/014466608X314935

Curran, T., & Keele, S. W. (1993). Attentional and nonattentional forms of sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(1), 189–202. https://doi.org/10.1037/0278-7393.19.1.189

Daft, R. L., & Lengel, R. H. (1986). Organizational Information Requirements, Media Richness and Structural Design. *Management Science*, *32*(5), 554–571.

Daley, B. J. (1999). Novice to expert: An exploration of how professionals learn. *Adult Education Quarterly*, *49*(4), 133–147. https://doi.org/10.1177/074171369904900401

Daronnat, S., Azzopardi, L., & Halvey, M. (2020). Impact of Agents' Errors on Performance, Reliance and Trust in Human-Agent Collaboration. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *64*(1), 405–409. https://doi.org/10.1177/1071181320641091

Daronnat, S., Azzopardi, L., & Halvey, M. (2022). Comparing Levels and Types of Situational-Awareness based Agent Transparency in Human-Agent Collaboration. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *66*(1), 1169–1173. https://doi.org/10.1177/1071181322661498

Daronnat, S., Azzopardi, L., Halvey, M., & Dubiel, M. (2021). Inferring Trust From Users' Behaviours; Agents' Predictability Positively Affects Trust, Task Performance and Cognitive Load in Human-Agent Real-Time Collaboration. *Frontiers in Robotics and AI*, *8*. https://doi.org/10.3389/frobt.2021.642201

Daronnat, S., Halvey, M., & Azzopardi, L. (2019). Human-Agent Collaborations: Trust in Negotiating Control. *Workshop Proceedings Everyday Automation Experience'19 In Conjunction with CHI'19, May 5th, 2019, Glasgow, UK*.

De Visser, E. J., Peeters, M. M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams.

International Journal of Social Robotics, *12*(2), 459–478. https://doi.org/10.1007/s12369-019-00596-x

Deschrijver, E., & Palmer, C. (2020). Reframing social cognition: Relational versus representational mentalizing. *Psychological Bulletin*, *146*(11), 941–969. https://doi.org/10.1037/bul0000302

Dienes, Z., & Berry, D. (1997). Implicit learning: Below the subjective threshold. *Psychonomic Bulletin & Review*, *4*(1), 3–23. https://doi.org/10.3758/bf03210769

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

Dunning, D. (2011). The dunning–kruger effect. In *Advances in Experimental Social Psychology* (Vol. 44, pp. 247–296). Elsevier. https://doi.org/10.1016/B978-0-12-385522-0.00005-6

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why People Fail to Recognize Their Own Incompetence. *Current Directions in Psychological Science*, *12*(3), 83–87. https://doi.org/10.1111/1467-8721.01235

Duradoni, M., Collodi, S., Perfumi, S. C., & Guazzini, A. (2021). Reviewing stranger on the internet: The role of identifiability through “Reputation” in online decision making. *Future Internet*, *13*(5), Article 5. https://doi.org/10.3390/fi13050110

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the Unskilled Are Unaware: Further Explorations of (Absent) Self-Insight Among the Incompetent. *Organizational Behavior and Human Decision Processes*, *105*(1), 98–121. https://doi.org/10.1016/j.obhdp.2007.05.002

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, *87*(3), 327–339. https://doi.org/10.1037/0022-3514.87.3.327

Fan, K. (1953). Minimax Theorems. *Proceedings of the National Academy of Sciences*, *39*(1), 42–47. https://doi.org/10.1073/pnas.39.1.42

Fiske, S. T. (2018). Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science*, *27*(2), 67–73. https://doi.org/10.1177/0963721417738825

Flavell, J. H., & Miller, P. H. (1998). Social cognition. In *Handbook of child psychology: Volume 2: Cognition, perception, and language* (pp. 851–898). John Wiley & Sons, Inc.

Forgas, J. P., & Laham, S. M. (2016). Halo effects. In R. F. Pohl (Ed.), *Cognitive Illusions: Intriguing Phenomena in Judgement, Thinking and Memory* (2nd ed.). Psychology Press. https://doi.org/10.4324/9781315696935

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. Springer New York. https://doi.org/10.1007/978-1-4419-0742-4

Franklin, S., & Graesser, A. (1997). Is it an agent, or just a program?: A taxonomy for autonomous agents. In J. P. Müller, M. J. Wooldridge, & N. R. Jennings (Eds.), *Intelligent Agents III Agent Theories, Architectures, and Languages* (Vol. 1193, pp. 21–35). Springer Berlin Heidelberg. https://doi.org/10.1007/BFb0013570

Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, *19*(4), 1273–1302. https://doi.org/10.1016/S1053-8119(03)00202-7

Frith, C. D., & Frith, U. (2012). Mechanisms of social cognition. *Annual Review of Psychology*, *63*(1), 287–313. https://doi.org/10.1146/annurev-psych-120710-100449

Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology*, *78*(4), 708–724. https://doi.org/10.1037/0022-3514.78.4.708

Gambino, A., Fox, J., & Ratan, R. (2020). Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication*, *1*(1). https://doi.org/10.30658/hmc.1.5

Gibson, F. P. (1996). The sugar production factory?A dynamic decision task. *Computational and Mathematical Organization Theory*, *2*(1), 49–60. https://doi.org/10.1007/bf00125763

Goldewijk, K. K. (2005). Three centuries of global population growth: A spatial referenced population (density) database for 1700-2000. *Population and Environment*, *26*(4), 343–367.

Gonzalez, C., Fakhari, P., & Busemeyer, J. (2017). Dynamic decision making: Learning processes and new research directions. *Human Factors*, *59*(5), 713–721. https://doi.org/10.1177/0018720817710347

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—explainable artificial intelligence. *Science Robotics*, *4*(37), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

Hafizoğlu, F. M., & Sen, S. (2018a). Reputation based trust in human-agent teamwork without explicit coordination. *Proceedings of the 6th International Conference on Human-Agent Interaction*, 238–245. https://doi.org/10.1145/3284432.3284454

Hafizoğlu, F. M., & Sen, S. (2018b). The effects of past experience on trust in repeated human-agent teamwork. *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 514–522.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527. https://doi.org/10.1177/0018720811417254

Hedlund, J., Ilgen, D. R., & Hollenbeck, J. R. (1998). Decision Accuracy in Computer-Mediated versus Face-to-Face Decision-Making Teams. *Organizational Behavior and Human Decision Processes*, *76*(1), 30–47. https://doi.org/10.1006/obhd.1998.2796

Hellström, T., & Bensch, S. (2018). Understandable robots—What, Why, and How. *Paladyn, Journal of Behavioral Robotics*, *9*(1), 110–123. https://doi.org/10.1515/pjbr-2018-0009

Herling, R. W. (2000). Operational definitions of expertise and competence. *Advances in Developing Human Resources*, *2*(1), 8–21. https://doi.org/10.1177/152342230000200103

Heyselaar, E. (2023). The CASA theory no longer applies to desktop computers. *Scientific Reports*, *13*(1), 19693. https://doi.org/10.1038/s41598-023-46527-9

Higgins, E. T., & Bargh, J. A. (1987). Social cognition and social perception. *Annual Review of Psychology*, *38*(1), 369–425.

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., & Couzin, I. D. (2015). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, *19*(1), 46–54. https://doi.org/10.1016/j.tics.2014.10.004

Hiltz, S. R., Johnson, K., & Turoff, M. (1986). Experiments in Group Decision Making Communication Process and Outcome in Face-to-Face Versus Computerized Conferences. *Human Communication Research*, *13*(2), 225–252. https://doi.org/10.1111/j.1468-2958.1986.tb00104.x

Ho, A., Hancock, J., & Miner, A. S. (2018). Psychological, Relational, and Emotional Effects of Self-Disclosure After Conversations With a Chatbot. *Journal of Communication*, *68*(4), 712–733. https://doi.org/10.1093/joc/jqy026

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, *59*(1), 91–117. https://doi.org/10.1016/0010-0277(96)81418-1

Huemer, M., Schröder, L. M., Leikard, S. J., Gruber, S., Mangstl, A., & Perner, J. (2023). The knowledge (“true belief”) error in 4- to 6-year-old children: When are agents aware of

what they have in view? *Cognition*, *230*, 105255. https://doi.org/10.1016/j.cognition.2022.105255

Husemann, S., Pöppel, J., & Kopp, S. (2022). Differences and Biases in Mentalizing About Humans and Robots. *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 490–497. https://doi.org/10.1109/RO-MAN53752.2022.9900849

Imhoff, R., & Koch, A. (2017). How Orthogonal Are the Big Two of Social Perception? On the Curvilinear Relation Between Agency and Communion. *Perspectives on Psychological Science*, *12*(1), 122–137. https://doi.org/10.1177/1745691616657334

Jaswal, V. K., & Malone, L. S. (2007). Turning believers into skeptics: 3-year-olds' sensitivity to cues to speaker credibility. *Journal of Cognition and Development*, *8*(3), 263–283. https://doi.org/10.1080/15248370701446392

Jian, J., Bisantz, A., & Drury, C. (2000). International Journal of Cognitive Foundations for an Empirically Determined Scale of Trust in Automated Systems Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

Joiner, J., Piva, M., Turrin, C., & Chang, S. W. C. (2017). Social learning through prediction error in the brain. *Npj Science of Learning*, *2*(1), 1–9. https://doi.org/10.1038/s41539-017-0009-2

Jung, C. G. (1996). *Mysterium coniunctionis* (6. Aufl). Walter.

Kassambara, A. (2020). *rstatix: Pipe-friendly framework for basic statistical tests (version 0.6.0*. https://CRAN.R-project.org/package=rstatix

Kaufman, S. B., DeYoung, C. G., Gray, J. R., Jiménez, L., Brown, J., & Mackintosh, N. (2010). Implicit learning as an ability. *Cognition*, *116*(3), 321–340. https://doi.org/10.1016/j.cognition.2010.05.011

Kaup, B., Ulrich, R., Bausenhart, K. M., Bryce, D., Butz, M. V., Dignath, D., Dudschig, C., Franz, V. H., Friedrich, C., Gawrilow, C., Heller, J., Huff, M., Hütter, M., Janczyk, M., Leuthold, H., Mallot, H., Nürk, H.-C., Ramscar, M., Said, N., … Wong, H. Y. (2024). Modal and amodal cognition: An overarching principle in various domains of psychology. *Psychological Research*, *88*(2), 307–337. https://doi.org/10.1007/s00426-023-01878-w

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2023). Trustworthy artificial intelligence: A review. *ACM Computing Surveys*, *55*(2), 1–38. https://doi.org/10.1145/3491209

Kelly, M., Kumar, A., Smyth, P., & Steyvers, M. (2023). Capturing humans' mental models of AI: An item response theory approach. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 1723–1734. https://doi.org/10.1145/3593013.3594111

Kohn, S. C., De Visser, E. J., Wiese, E., Lee, Y.-C., & Shaw, T. H. (2021). Measurement of trust in automation: A narrative review and reference guide. *Frontiers in Psychology*, *12*, 604977. https://doi.org/10.3389/fpsyg.2021.604977

Koriat, A. (2008). Subjective confidence in one's answers: The consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 945–959. https://doi.org/10.1037/0278-7393.34.4.945

Kosinski, M. (2024). *Evaluating Large Language Models in Theory of Mind Tasks* (arXiv:2302.02083). arXiv. http://arxiv.org/abs/2302.02083

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can Machines Think? Interaction and Perspective Taking with Robots Investigated via fMRI. *PLOS ONE*, *3*(7), e2597. https://doi.org/10.1371/journal.pone.0002597

Krauss, R. M., & Glucksberg, S. (1977). Social and Nonsocial Speech. *Scientific American*, *236*(2), 100–105.

Kruger, J., & Dunning, D. (1999). *Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments*. https://doi.org/10.1037/0022-3514.77.6.1121

Kulke, L., von Duhn, B., Schneider, D., & Rakoczy, H. (2018). Is implicit theory of mind a real and robust phenomenon? Results from a systematic replication study. *Psychological Science*, *29*(6), 888–900. https://doi.org/10.1177/0956797617747090

Kulms, P., & Kopp, S. (2016). The effect of embodiment and competence on trust and cooperation in human–agent interaction. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10011 LNAI*, 75–84. https://doi.org/10.1007/978-3-319-47665-0_7

Kulms, P., & Kopp, S. (2019). More human-likeness, more trust? The effect of anthropomorphism on self-reported and behavioral trust in continued and interdependent human–agent cooperation. *ACM International Conference Proceeding Series*, 31–42. https://doi.org/10.1145/3340764.3340793

Kulms, P., Mattar, N., & Kopp, S. (2015). An interaction game framework for the investigation of human–agent cooperation. In W.-P. Brinkman, J. Broekens, & D. Heylen (Eds.), *Intelligent Virtual Agents* (pp. 399–402). Springer International Publishing. https://doi.org/10.1007/978-3-319-21996-7_43

Kulms, P., Mattar, N., & Kopp, S. (2016). Can't do or won't do? Social attributions in human–agent cooperation. *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, 1341–1342.

Kumar, A., Smyth, P., & Steyvers, M. (2023). *Differentiating mental models of self and others: A hierarchical framework for knowledge assessment*. PsyArXiv. https://doi.org/10.31234/osf.io/jvkp5

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). **lmerTest** package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Larson, L., & DeChurch, L. A. (2020). Leading teams in the digital age: Four perspectives on technology and what they mean for leading teams. *Leadership Quarterly*, *31*(1), 101377. https://doi.org/10.1016/j.leaqua.2019.101377

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lee, S., Ratan, R., & Park, T. (2019). The Voice Makes the Car: Enhancing Autonomous Vehicle Perceptions and Adoption Intention through Voice Agent Gender and Style. *Multimodal Technologies and Interaction*, *3*(1), Article 1. https://doi.org/10.3390/mti3010020

Leiner, D. J. (2023). *SoSci survey (version 2.5.00-i1142*. http://www.soscisurvey.com

Lewis, S. L., & Maslin, M. A. (2015). Defining the anthropocene. *Nature*, *519*(7542), Article 7542. https://doi.org/10.1038/nature14258

Lievens, F., Sackett, P. R., & Zhang, C. (2021). Personnel selection: A longstanding story of impact at the individual, firm, and societal level. *European Journal of Work and Organizational Psychology*, *30*(3), 444–455. https://doi.org/10.1080/1359432X.2020.1849386

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

Low, J., & Watts, J. (2013). Attributing false beliefs about object identity reveals a signature blind spot in humans' efficient mind-reading system. *Psychological Science*, *24*(3), 305–311. https://doi.org/10.1177/0956797612451469

Lucassen, T., & Schraagen, J. M. (2011). Factual accuracy and trust in information: The role of expertise. *Journal of the American Society for Information Science and Technology*, *62*(7), 1232–1242. https://doi.org/10.1002/asi.21545

Maes, P. (1993). Modeling Adaptive Autonomous Agents. *Artificial Life*, *1*(1_2), 135–162. https://doi.org/10.1162/artl.1993.1.1_2.135

Matsui, T., Yamamoto, T., Miura, Y., & McCagg, P. (2016). Young children's early sensitivity to linguistic indications of speaker certainty in their selective word learning. *Lingua*, *175–176*, 83–96. https://doi.org/10.1016/j.lingua.2015.10.007

Mayer, M., Broß, M., & Heck, D. W. (2023). Expertise determines frequency and accuracy of contributions in sequential collaboration. *Judgment and Decision Making*, *18*, e2. https://doi.org/10.1017/jdm.2023.3

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review*, *20*(3), 709. https://doi.org/10.2307/258792

McCroskey, J. C., & Teven, J. J. (1999). Goodwill: A reexamination of the construct and its measurement. *Communication Monographs*, *66*(1), 90–103. https://doi.org/10.1080/03637759909376464

McIlroy-Young, R., Sen, S., Kleinberg, J., & Anderson, A. (2020). Aligning superhuman AI with human behavior: Chess as a model system. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1677–1687. https://doi.org/10.1145/3394486.3403219

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

McIntosh, R. D., Moore, A. B., Liu, Y., & Della Sala, S. (2022). Skill and self-knowledge: Empirical refutation of the dual-burden account of the dunning–kruger effect. *Royal Society Open Science*, *9*(12), 191727. https://doi.org/10.1098/rsos.191727

McNeese, N. J., Demir, M., Chiou, E. K., & Cooke, N. J. (2021). Trust and Team Performance in Human–Autonomy Teaming. *International Journal of Electronic Commerce*, *25*(1), 51–72. https://doi.org/10.1080/10864415.2021.1846854

McNeese, N. J., Demir, M., Cooke, N. J., & Myers, C. (2018). Teaming With a Synthetic Teammate: Insights into Human-Autonomy Teaming. *Human Factors*, *60*(2), 262–273. https://doi.org/10.1177/0018720817743223

Melo, C. D., Marsella, S., & Gratch, J. (2016). People Do Not Feel Guilty About Exploiting Machines. *ACM Transactions on Computer-Human Interaction*, *23*(2), 1–17. https://doi.org/10.1145/2890495

Morgan, E. J., Carroll, D. J., Chow, C. K. C., & Freeth, M. (2022). The effect of social presence on mentalizing behavior. *Cognitive Science*, *46*(4), e13126. https://doi.org/10.1111/cogs.13126

Moscovici, S. (1988). Notes towards a description of Social Representations. *European Journal of Social Psychology*, *18*(3), 211–250. https://doi.org/10.1002/ejsp.2420180303

Moskowitz, G. B. (2005). *Social cognition: Understanding self and others*. Guilford Press.

Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, *56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Celebrating Interdependence - CHI '94*, 72–78. https://doi.org/10.1145/191666.191703

Navarro, D. J. (2015). *Learning statistics with R: a tutorial for psychology students and other beginners (version 0.5*. University of Adelaide. Adelaide. http://ua.edu.au/ccs/teaching/lsr

Newton, A. M., & de Villiers, J. G. (2007). Thinking while talking: Adults fail nonverbal false-belief reasoning. *Psychological Science*, *18*(7), 574–579. https://doi.org/10.1111/j.1467-9280.2007.01942.x

Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological Bulletin*, *125*(6), 737–759. https://doi.org/10.1037/0033-2909.125.6.737

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, *35*(4), 250–256. https://doi.org/10.1037/0022-3514.35.4.250

Nohara-LeClair, M. (2001). A Direct Assessment of the Relation Between Shared Knowledge and Communication in a Referential Communication Task. *Language and Speech*, *44*(2), 217–236. https://doi.org/10.1177/00238309010440020501

Nuhfer, E., Cogan, C., Fleisher, S., Gaze, E., & Wirth, K. (2016). Random number simulations reveal how random noise affects the measurements and graphical portrayals of self-assessed competency. *Numeracy*, *9*(1). http://dx.doi.org/10.5038/1936-4660.9.1.4

Nuhfer, E., Fleisher, S., Cogan, C., Wirth, K., & Gaze, E. (2017). How random noise and a graphical convention subverted behavioral scientists' explanations of self-assessment data: Numeracy underlies better alternatives. *Numeracy*, *10*(1). http://dx.doi.org/10.5038/1936-4660.10.1.4

O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2022). Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature. *Human Factors*, *64*(5), 904–938. https://doi.org/10.1177/0018720820960865

Ono, T., Imai, M., & Nakatsu, R. (2000). Reading a robot's mind: A model of utterance understanding based on the theory of mind mechanism. *Advanced Robotics*, *14*(4), 311–326. https://doi.org/10.1163/156855300741609

Pajitnov, A. (1984). *Tetris* [Elektronika 60].

Penn, D. C., & Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind.' *Philosophical Transactions of the*

Royal Society B: Biological Sciences, *362*(1480), 731–744. https://doi.org/10.1098/rstb.2006.2023

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*(5), 233–238. https://doi.org/10.1016/j.tics.2006.03.006

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526. https://doi.org/10.1017/S0140525X00076512

R Core Team. (2023). *R: a language and environment for statistical computing* [Manual]. R Foundation for Statistical Computing. https://www.R-project.org/

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*(6), 855–863. https://doi.org/10.1016/s0022-5371(67)80149-x

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, *118*(3), 219–235. https://doi.org/10.1037/0096-3445.118.3.219

Reynolds, R. I. (1992). Recognition of expertise in chess players. *The American Journal of Psychology*, *105*(3), 409–415. https://doi.org/10.2307/1423195

Rieh, S. Y., & Danielson, D. R. (2007). Credibility: A multidisciplinary framework. *Annual Review of Information Science and Technology*, *41*(1), 307–364. https://doi.org/10.1002/aris.2007.1440410114

Robinette, P., Howard, A. M., & Wagner, A. R. (2017). Effect of Robot Performance on Human-Robot Trust in Time-Critical Situations. *IEEE Transactions on Human-Machine Systems*, *47*(4), 425–436. https://doi.org/10.1109/THMS.2017.2648849

Ruby, P., & Decety, J. (2001). Effect of subjective perspective taking during simulation of action: A PET investigation of agency. *Nature Neuroscience*, *4*(5), 546–550. https://doi.org/10.1038/87510

Sabbagh, M. A., & Baldwin, D. A. (2001). Learning words from knowledgeable versus ignorant speakers: Links between preschoolers' theory of mind and semantic development. *Child Development*, *72*(4), 1054–1070. https://doi.org/10.1111/1467-8624.00334

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences*, *19*(2), 65–72. https://doi.org/10.1016/j.tics.2014.11.007

Schlinger, H. D. (2009). Theory of mind: An overview and behavioral perspective. *The Psychological Record*, *59*(3), 435–448. https://doi.org/10.1007/BF03395673

Schneider, D., Slaughter, V. P., & Dux, P. E. (2017). Current evidence for automatic theory of mind processing in adults. *Cognition*, *162*, 27–31. https://doi.org/10.1016/j.cognition.2017.01.018

Schwesig, R., Brich, I., Buder, J., Huff, M., & Said, N. (2023). Using artificial intelligence (AI)? Risk and opportunity perception of AI predict people's willingness to use AI. *Journal of Risk Research*. https://www.tandfonline.com/doi/full/10.1080/13669877.2023.2249927

Scott, R. M., & Baillargeon, R. (2017). Early false-belief understanding. *Trends in Cognitive Sciences*, *21*(4), 237–249. https://doi.org/10.1016/j.tics.2017.01.012

Seger, C. A. (1994). Implicit learning. *Psychological Bulletin*, *115*(2), 163–196. https://doi.org/10.1037/0033-2909.115.2.163

Shanks, D., & St. John, M. (1994). Characteristics of dissociable human learning-systems. *Behavioral and Brain Sciences*, *17*(3), Article 3.

Sheridan, T. B. (2019). Individual differences in attributes of trust in automation: Measurement and application to system design. In *Frontiers in Psychology* (Vol. 10, Issue MAY). Frontiers Media S.A. https://doi.org/10.3389/fpsyg.2019.01117

Srinivasan, V., & Takayama, L. (2016). Help Me Please: Robot Politeness Strategies for Soliciting Help From Humans. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4945–4955. https://doi.org/10.1145/2858036.2858217

Sun, R., Slusarz, P., & Terry, C. (2005). The interaction of the explicit and the implicit in skill learning: A dual-process approach. *Psychological Review*, *112*(1), 159–192. https://doi.org/10.1037/0033-295X.112.1.159

Sundar, S. S. (2020). Rise of Machine Agency: A Framework for Studying the Psychology of Human–AI Interaction (HAII). *Journal of Computer-Mediated Communication*, *25*(1), 74–88. https://doi.org/10.1093/jcmc/zmz026

Thornton, M. A., Weaverdyck, M. E., & Tamir, D. I. (2019). The social brain automatically predicts others' future mental states. *Journal of Neuroscience*, *39*(1), 140–148. https://doi.org/10.1523/JNEUROSCI.1431-18.2018

Tullis, J. G. (2018). Predicting others' knowledge: Knowledge estimation as cue utilization. *Memory & Cognition*, *46*(8), 1360–1375. https://doi.org/10.3758/s13421-018-0842-4

Tullis, J. G., & Feder, B. (2022). The "curse of knowledge" when predicting others' knowledge. *Memory & Cognition*. https://doi.org/10.3758/s13421-022-01382-3

Tullis, J. G., & Fraundorf, S. H. (2017). Predicting others' memory performance: The accuracy and bases of social metacognition. *Journal of Memory and Language*, *95*, 124–137. https://doi.org/10.1016/j.jml.2017.03.003

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Ullman, D., & Malle, B. F. (2020). *MDMT: Multi-Dimensional Measure of Trust*.

Ullrich, D., Butz, A., & Diefenbach, S. (2021). The Development of Overtrust: An Empirical Simulation and Psychological Analysis in the Context of Human–Robot Interaction. *Frontiers in Robotics and AI*, *8*(April), 1–15. https://doi.org/10.3389/frobt.2021.554578

Van Der Maas, H. L. J., & Wagenmakers, E.-J. (2005). A psychometric analysis of chess expertise. *The American Journal of Psychology*, *118*(1), 29–60.

Van der Velde, F. (2015). Communication, concepts and grounding. *Neural Networks*, *62*, 112–117. https://doi.org/10.1016/j.neunet.2014.07.003

Van der Wel, R. P. R. D., Sebanz, N., & Knoblich, G. (2014). Do people automatically track others' beliefs? Evidence from a continuous measure. *Cognition*, *130*(1), 128–133. https://doi.org/10.1016/j.cognition.2013.10.004

Van Neumann, J. (1928). Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, *100*(1), 295–320. https://doi.org/10.1007/BF01448847

Van Overwalle, F., & Vandekerckhove, M. (2013). Implicit and explicit social mentalizing: Dual processes driven by a shared neural network. *Frontiers in Human Neuroscience*, *7*. https://www.frontiersin.org/articles/10.3389/fnhum.2013.00560

Vinanzi, S., Cangelosi, A., & Goerick, C. (2021). The collaborative mind: Intention reading and trust in human-robot interaction. *iScience*, *24*(2), 102130. https://doi.org/10.1016/j.isci.2021.102130

Walther, J. B., Van Der Heide, B., Hamel, L. M., & Shulman, H. C. (2009). Self-generated versus other-generated statements and impressions in computer-mediated communication: A test of warranting theory using facebook. *Communication Research*, *36*(2), 229–253. https://doi.org/10.1177/0093650208330251

Warren, W. H. (2018). Collective Motion in Human Crowds. *Current Directions in Psychological Science*, *27*(4), 232–240. https://doi.org/10.1177/0963721417746743

Weinberger, A. B., & Green, A. E. (2022). Dynamic development of intuitions and explicit knowledge during implicit learning. *Cognition*, *222*, 105008. https://doi.org/10.1016/j.cognition.2021.105008

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*(3), 655–684. https://doi.org/10.1111/1467-8624.00304

Wojton, H. M., Porter, D., T. Lane, S., Bieber, C., & Madhavan, P. (2020). Initial validation of the trust of automated systems test (TOAST). *Journal of Social Psychology*, *160*(6), 735–750. https://doi.org/10.1080/00224545.2020.1749020

Xu, Q. (2014). Should I trust him? The effects of reviewer profile characteristics on eWOM credibility. *Computers in Human Behavior*, *33*, 136–144. https://doi.org/10.1016/j.chb.2014.01.027

Appendix

Appendix A

Appendix A1

Explicit Knowledge Questionnaire

Item	Random	Smart
The bot keeps changing its strategy.	F	F
The bot has a hidden ranking of all possible pieces. Every round, the highest-ranked, available piece is chosen.	F	F
The bot takes the current configuration of the playing field into account.	F	T
The bot considers all options.	F	T
The bot plans several steps ahead.	F	F
The bot always selects the longest element.	F	F
The bot acts according to a predefined sequence.	F	F
The bot picks a random element.	T	F
The bot tries to keep the number of holes small.	F	T
If two pieces fit equally well, the bot picks the left one.	F	T
The bot greedily tries to finish rows.	*F*	*F*
The bot minimizes the negative impact of the mistakes you can make.	*F*	*F*
The bot tries to keep the differences between columns small.	*F*	*T*
The bot uses information that is not visible to the player.	*F*	*F*
All possible rotations of each piece are taken into account by the bot when deciding on a piece.	*F*	*T*

Note: In the first study „The bot“ was called „The algorithm“. Items in italics are used only in the second Study. „F“ and „T“ mark False and True statements for the respective bot.

Appendix A2

Mutual Understanding Questions

- It was easy to predict the next element
- I have the impression that I can understand the algorithms decision
- I don't know what the algorithm intended
- I can explain how the algorithm comes to its decisions
- I was able to put myself in the algorithms shoes
- The actions of the algorithm were confusing to me
- I understood how the algorithm makes decisions
- The algorithm played different than me

Scale: strongly disagree, somewhat disagree, neither agree nor disagree, somewhat agree, strongly agree

Appendix A3

Assumptions

Q: How strongly do you believe these statments about the bot are true?

- The bot acts goal-directed.
- The bot acts optimally.
- The bot tries to get a good score.
- The bot interferes with my game.
- The bot acts intelligently.
- The bot is unhelpful.
- The bot's goals align with mine.
- The bot tries to prevent my success.
- The bot acts rationally.

Scale: certainly false - likely false - undecided - likely true - certainly true

Appendix B

Appendix B1

Agent Outputs

0 Rules	2 Rules	4 Rules	6 Rules
TE	MTTRTRMRMX	MMTTR	MTXR
IEYICCWP	MRXRTRXTXM	MXTTMXR	MXTXTXR
SN	MXMMXTTTR	MMTMTTR	MXTXTXTR
CHQXJEBR	MRXXMXRMRT	MMXTXTTR	MXTR

Appendix B2

Trivia Questions

Category	Question	Answer	Distractor 1	Distractor 2	Distractor 3
Science & Tech	Who was the most famous Greek doctor?	Hippocrates	Democritus	Galen	Parmenides
	What device is used to measure levels of radioactivity?	Geiger Counter	Hertz Counter	Ohm Receiver	Radiometer
	What mammal has armor-like bony plates as its most distinguishing feature?	Armadillo	Aardvark	Anteater	Platypus
	What do we call the bone which composes the lower jaw?	Mandible	Femur	Goman Arch	Maxilla
	What is a goat's offspring called?	Kid	Colt	Cub	Doe
	What does VoIP stand for?	Voice over Internet Protocol	Vision of Interest Piece	Vivendi operation Internet Protocol	Vocal or Internet Program
	What is the long process by which a dead organism turns to stone?	Petrification	Decomposition	Ossification	Rigor Mortis
	What type of paper is commonly used as an acid-base indicator in chemistry classes?	Litmus	Barium	Chromatin Red	Chromium
	What is a network designed to allow communication within an organization?	An intranet	Portal	The internet	Yahoo
	What is a community of ants called?	Colony	Farm	Hill	Hive
	What is the last name of the doctor who first developed a vaccine against polio?	Salk	Fleming	Pauling	Sabin
	What is the mixture of two or more metallic elements?	Alloy	Alchemy	Amalgam	Compound
	What is the colloquial term for patella?	Kneecap	Elbow	Funny Bone	Shin
	What is the orbiting particle of an atom?	Electron	Neutron	Nucleus	Proton
	In what geological period did birds evolve?	Jurassic	Permian	Pleistocene	Triassic

Category	Question	Answer	Distractor 1	Distractor 2	Distractor 3
	What is caterpillar waste called?	Frass	Clods	Coprolites	Guano
	Who discovered the law of electrolysis?	Michael Faraday	Alessandro Volta	James Watt	Samuel Morse
	Who invented the aerosol spray can?	Erik Rotheim	Fritz Haber	Percy Lavon Julian	Wallace Carothers
	Who is known as "The Father of Geometry"?	Euclid	Archimedes	Ptolemy	Pythagoras
	Which animal has the widest hearing range?	Dolphin	Bat	Dog	Human being
Art & Culture	What is the hard, white material sourced from elephant tusks?	Ivory	Enamel	Baleen	Porcelain
	What is the name of the art of Japanese paper folding?	Origami	Amigurumi	Bonsai	Temari
	Who is the artist that painted everyday objects such as soup cans?	Warhol	Dali	Monet	Picasso
	What is the name of the song traditionally sung at the stroke of midnight on New Year's Eve?	Auld Lang Syne	Ave Maria	Happy Days Are Here Again	Memories
	What is the traditional daytime sleep in Spain?	Siesta	Despierto	Dormir	Fiesta
	Whose autobiography is titled "Lady Sings the Blues"?	Billie Holiday	Nina Simone	Peggy Lee	Rosemary Clooney
	What is the last name of the actor who starred in both the Broadway production and the film, "A Streetcar Named Desire"?	Brando	Dean	McQueen	Newman
	Which Aerosmith song was re-made by Run D.M.C.?	Walk this Way	Dream On	Dude	Mama Kin
	What is the last name of the artist who painted "The Persistence of Memory"?	Dali	de Chirico	Duchamp	Miro
	What mythical creature did Perseus kill?	Medusa	Andromeda	Hydra	Kraken
	What is the last name of the actor who played the scarecrow in the movie, "The Wizard of Oz"?	Bolger	Ebsen	Haley	Lahr
	The theme tune for 'Monty Python's Flying Circus' was written by which composer?	John Philip Sousa	Albert Austin Harding	Charles Benter	Edwin Franko Goldman
	What is the last name of the singer who made a hit recording of the song "Who's Sorry Now?"	Francis	Darrin	Day	Sedaka

Category	Question	Answer	Distractor 1	Distractor 2	Distractor 3
	The Mayan and Aztec peoples used cocoa beans not only to make a delicious beverage but also as...?	Currency	Dye	Fertilizer	Medicine
	What is a chord consisting of three tones of a diatonic scale called?	Triad	Diminished fifth	Treble	Seventh
	Who wrote the "Threepenny Opera"?	Bertolt Brecht	Carl Zuckmayer	Ernst Toller	Frank Wedekind
	Who is the Hindu god associated with rain?	Indra	Brahma	Shiva	Vishnu
	Which Egyptian god was often represented as a falcon?	Horus	Osiris	Seth	Thoth
	In which British national daily newspaper does Rupert The Bear appear?	The Daily Express	The Guardian	The Sun	The Sunday Telegraph
	From what musical is the song “Baubles Bangles And Beads”?	Kismet	Carousel	Gigi	Kiss Me, Kate
History	What was the name of the zeppelin, blimp, that exploded in Lake Hurst, New Jersey in 1937?	Hindenburg	Graph	Ludendorf	Mannheim
	What is the last name of the first female pilot to cross the Atlantic?	Earhart	Johnson	Lindbergh	Rickenbacker
	What is the famous prehistoric structure situated on Salisbury Plain England?	Stonehenge	Blarney Stone	Druid Ruins	Easter Statues
	What was the name of the infamous American traitor in the Revolutionary War?	Arnold	Brown	Jackson	Johnson
	What was the last name of the explorer who tramped through what is now Florida looking for the fountain of youth?	Ponce De Leon	De La Vega	De Soto	Diego
	What was the location of George Washington's encampment where his men suffered every conceivable hardship from 1777-1778?	Valley Forge	Antietam	Bull Run	Gettysburg
	What is the name of the short sword fastened to the end of a musket or rifle?	Bayonet	Cutlass	Dagger	Dirk
	What peace treaty ended World War I?	Versailles	Berlin	Hague	Paris
	What is the name of the island on which Napoleon was born?	Corsica	Cyprus	Malta	Palma
	What is the name of the Chinese religion founded by Lao Tse?	Taoism	Confucianism	Shamanism	Zen
	What is the last name of the man who said, "I only regret that I have but one life to lose for my country"?	Hale	Arnold	Greene	Howe

Category	Question	Answer	Distractor 1	Distractor 2	Distractor 3
	What was the name of the Union ironclad ship that fought the Confederate ironclad Merrimack?	Monitor	Cumberland	Manassas	Virginia
	Where was the 1939 World's Fair held?	New York City	Berlin	London	Paris
	Halloween traces its origins to which pagan festival?	Samhain	Imbolc	Lughnassadh	Ostara
	What was the name of the U.S. surveillance ship that was attacked by North Vietnam in the Gulf of Tonkin?	USS Maddox	HMAS Brisbane	USS Canberra	USS Wichita
	What philosopher was considered the ugliest man in Athens?	Socrates	Democritus	Empedocles	Plato
	Which Pulitzer Prize winner has won more than twice?	Robert Frost	Barbara W. Tuchman	Margaret Leech	William Faulkner
	Which Cleveland mayor was the first African American mayor of a major U.S. city?	Carl Stokes	Harold Washington	Maynard Jackson	Sly James
	Who replaced Betty Boothroyd as "Speaker of the House of Commons"?	Michael Martin	Bernard Weatherill	George Thomas	John Bercow
	Who was the first person to admit to practicing witchcraft in Salem?	Tituba	Mary Bradbury	Rebecca Nurse	Sarah Good

Note: Questions and answers from Coane and Umanath (2021). The original questions were given categories using „gpt-3.5-turbo“. Using the automated categorization, the author manually created the databeases for the three categories „Science & Technology“, „Arts & Culture“, and „History“. From these databases the 25 questions were drawn randomly. Of these 25 questions unfitting or duplicate draws were dismissed, the rmainig questions were randomly dismissed to result in these 20 questions per category.

Appendix C

Appendix C1

Reputation Texts

Good reputation:

In the following rounds you will play the Tetris game with a bot. This bot has a high accuracy. The bot will always choose the block, which will fill out the row in the best possible way.
Try to fill out as many rows in the following 100 rounds. You can decide every round whether you want the bot to choose a block (by button press) or whether you choose it yourself. You can also place the tick at “trust the bot in the following rounds” to keep letting the bot choose in the following round by default. The tick can be removed every round, if you decide to.

Bad reputation:

In the following rounds you will play the Tetris game with a bot. This bot has a low accuracy. The bot may not always choose the block, which will fill out the row in the best possible way.
Try to fill out as many rows in the following 100 rounds. You can decide every round whether you want the bot to choose a block (by button press) or whether you choose it yourself. You can also place the tick at “trust the bot in the following rounds” to keep letting the bot choose in the following round by default. The tick can be removed every round, if you decide to.

Appendix C2

Trust Questionnaire

Scale	Subscale	Item
Multi-Dimensional Measures of Trust (MDMT) Performance trust	reliable	reliable
		predictable
		dependable
		consistent
	competence	competent
		skilled
		better than me (changed from original *capable*)
		meticulous
Source Credibility Measures	expertise	trained
		expert
		informed
		intelligent
MDMT Moral Trust	ethical	ethical
		principled
		moral
		has integrity
	benevolent	benevolent
		kind
		considerate
		has goodwill
Trust in Automated Systems (TOAST)	system processes	I understand how the system executes the tasks
		I wish I could had more control over the system
		I am surprised about executions (altered)
		I have all the info about the system I need (altered)
		The system is transparent (altered)
	System purpose	I understand what the tasks of the system are (altered)
		I understand what the expected outcomes are (altered)
		I understand the capabilities of the system
		I understand the limitations of the system
Self-generated		The system is trustworthy
		I trust the system

Note: Scale from -3 („Not at all“) to 3 („Very“), including non-informative option „Does not fit“.

Some items appeared in multiple sources