The Human Perspective on Artificial Intelligence (AI): How do people perceive and relate to AI in the social sphere?

Tsourgianni, Afroditi

PhD Thesis

University College London (UCL), 2025

Experimental Psychology

I, Afroditi Tsourgianni confirm that the work presented in my thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

This thesis explores how people perceive and make decisions with Artificial Intelligence (AI), using methods from psychology to map the perceptions and investigate the behaviours. Through eight studies, it provides empirical insights into the human perspective on AI.

Chapter 2 (Studies 1 and 2) examines AI perception using three models: the Stereotype Content Model (SCM) (Fiske et al., 2002), the Mind Perception Dimensions (MPD) model (Gray et al., 2007), and the newly developed AI Stereotype Model (AISM). These models were evaluated across various AI agents, considering variations in design features, embodiment and intended purpose of use. The findings reveal that AI perception is not homogeneous; instead, distinct stereotypes emerge based on competence and experience—the two core dimensions of AISM. This model proved more effective than SCM and MPD in capturing AI perception.

Chapter 3 (Studies 3 and 4) explores trust in AI based on three key determinants: performance, process, and purpose (Lee & See, 2004). Trust in AI was found to depend on how these factors were weighted, particularly in moral versus non-moral decisions. In moral decisions, trust was shaped by the AI's moral stance ('why' it decides), whereas in non-moral decisions, trust was driven by the AI's decision-making process ('how' it decides), with detailed explanations fostering greater trust.

Chapter 4 (Studies 5–8) investigates how people respond to AI-generated versus human advice. Findings confirmed that people are more likely to trust AI than humans when decisions are perceived as objective (Studies 5 and 6). Studies 7 and 8 extended this research by examining AI's role in validating subjective preferences-based decisions. Results indicate that people value validation similarly, whether from AI or human advisors. Collectively, these findings provide a nuanced understanding of AI perception, trust, and decision-making

Impact statement

Artificial Intelligence (AI) is becoming an increasingly essential part of human life, with more sophisticated algorithms and more humanlike avatars and robots potentially becoming a common reality within the next 20 years—or even sooner. However, our understanding of how people perceive AI and make decisions with AI and how these perceptions and behaviours change over time remains relatively underdeveloped. The current work contributes to ongoing research which investigates and diligently chronicles the evolving human perspective on AI. Since most of the empirical data was collected in 2022 -2023, it provides a snapshot of what is coined ‘the human perspective on AI’ at that time. Additionally, it provides a framework for understanding and mapping human perception of AI over time and across the spectrum of AI agents, those currently available and those that are yet to be invented, and presents the first evidence of emerging AI stereotypes based on perceptions of competence and experience. It also underscores the need to continuously update our understanding of people’s perceptions and behaviours towards AI as both the underlying technology and people's experience with it evolve.

Acknowledgements

> ‘For me, becoming isn’t about arriving somewhere or achieving a certain aim. Becoming is a forward motion, a means of evolving, a way to reach continuously toward a better self. The journey doesn’t end.’
>
> — Michelle Obama, Becoming

I never imagined myself becoming a researcher. The journey I took with this PhD revealed to me the power of the scientific method in bringing us closer to finding answers. More important than finding answers, it helped me ask questions, think deeply and critically about findings, conduct research, and collaborate. It also taught me to be brave, patient and to never give up on the thing that I love doing in life. Just keep going.

I would like to first thank my supervisor, Lasana Harris, for his support, his patience and the many intellectually stimulating discussions we have had (and hopefully will continue to have). This journey would not have been possible without you, **Lasana**. Thank you for believing in me and offering me a ride. You also taught me, by your example, how to be a researcher who collaborates, shares their findings with confidence, stick to doing their work correctly, remains humble, and continuously seeks to learn. Sincere thanks also go to my co-supervisor **Essi Viding** for also believing in me and offering me her feedback whenever I turned to her for support.

Thank you to the examiners committee, **Emily Cross** and **Joe Devlin**, who will read this thesis and offer me their valuable feedback, guiding me on the journey of becoming.

Finally, I am blessed to have good friends and family in my life. Thank you to my mum and dad for their love and support, and my sister for being there for me, especially during the most challenging parts of this journey. Even though she was miles away physically, she was always close. Finally, a special thanks go to wonderful humans whose presence in my life during various parts of this journey made all the difference. **Gunn**, **Anna**, **Rachel**, **Emmanuel**, **Melissa**, **Ram**, **Adelyn**, **Deepak**, and **Tim** - thank you for your support.

Finally, there is one person whose friendship, love, and support mean the world to me—the best possible companion on this journey. **Agata**, thank you for being the strong, intelligent, kind, badass woman you are and for being in my life. You have been my anchor of wisdom and joy, and I only hope to be half the friend to you that you are to me. We did it—this thesis is complete, and it is dedicated to you.

Table of contents

Chapter 1 Introduction

People perceive other people and their behaviour on a daily basis. Thinking about other people's minds, their thoughts, and feelings, and making inferences about their intentions occurs spontaneously to humans (Frith & Frith, 2001; Schurz et al., 2014; Van Overwalle, 2009) and is a fundamental way for humans to navigate the social world. Yet, today's world is not populated only with humans but also non-human agents such as Artificial Intelligent (AI). AI such as virtual assistants, chatbots, avatars, smart devices and social robots are increasingly forming part of the human daily existence. In the future, such entities are expected to be further integrated and co-exist alongside humans in various sectors, including education and healthcare (Ayeni et al., 2024; Kasula, 2024). More recently, interactive disembodied AI in the form of Large Language Models (LLMs) like ChatGPT, have risen to prominence. Only within the first week of its release, back in November 2022, ChatGPT received 1 million users (Sier, 2022). These interactive disembodied AI use natural language—a form of communication currency typically associated with human interactions—and demonstrate the ability to engage in a dialogue on almost any topic.

AI is not only entering the social sphere but while doing so, it is also increasingly taking the role of an advisor in human decision making (Rahwan, Cebrian, Obradovich, Bongard, Bonnefon, Breazeal, Crandall, Christakis, Couzin, Jackson, et al., 2019). From AIs that prefilter online content (Adomavicius et al., 2013; Jesse & Jannach, 2021) to AIs that determine healthcare support eligibility (Ledford, 2019; Zack et al., 2024) and advise judges on the probability of reoffending (Angwin et al., 2022; Scantamburlo et al., 2018), AI plays a significant role in human decision making. People are expected—and will continue to be

expected—to navigate a world where human decision-making is both supported by and, at times, reshaped by AI (El Naqa et al., 2020; Kaggwa et al., 2024; Tewari & Pant, 2020).

Studying how people perceive and make decisions with AI is relevant because of two key reasons. First, AI in different forms and functions integrates into both public spaces and private lives, and as people's familiarity and experience with different AI agents increases, numerous questions can be asked regarding how humans perceive AI and respond to AI-generated outputs. Questions that, when addressed, can provide valuable insights into human psychology, inform the design and deployment of AI, and contribute empirical evidence to the ongoing discourse on the moral rules and legal canon surrounding the use and treatment of AI. Secondly, as people's familiarity and experience with AI grows, and as new AI agents are being developed and introduced to the public, perceptions and behaviours are expected to change. This dynamic nature—both of AI technology and human familiarity with it—dictates the need for ongoing research, making empirical evidence on human perception and behaviour inherently timestamped, tied to the specific point in time they are collected.

This thesis explores how people perceive AI and trust it with their decisions, using methods from psychology to map the perceptions and investigate the behaviours. It takes a snapshot of the human perspective on AI with the timestamp as of 2022, as most of the empirical work was conducted between May and August 2022, employing various measurements to tap into both perception and behaviour towards AI. It comprises eight studies (Chapter 2 through 4). It starts with the mapping of human perception across various AI agents using theoretical models from social psychology and introduces a new data-driven model (Chapter 2). It then examines behaviour towards AI, specifically how people form evaluations for the trustworthiness of AI models (Chapter 3) and how they respond to AI-generated advice, particularly advice that

validates their decisions (Chapter 4). To provide additional context for the current work, an overview of key findings from the literature on how people perceive and trust technology, automation, and AI, is presented first in the rest of this chapter, along with a summary of relevant concepts. This is followed by the research questions explored in this thesis.

Artificial Intelligence (AI): A Working Definition

There is an inherent problem with AI. Despite its widespread use among the general public and within the AI research community, there is no single, universally agreed-upon definition for it. Since its inception as a term by John McCarthy in 1955, when writing a research proposal requesting funds for a summer research project on Artificial Intelligence in Dartmouth (McCarthy et al., 2006), AI has embraced the idea of simulating (i.e., imitating with the use of models) human intelligence, including scientific knowledge, common sense, and self-improvement (Haenlein & Kaplan, 2019; McCarthy, 2022).

However, throughout the years, AI has become synonymous with different things, often reflecting the most popular technology of the time. For example, today's AI is most probably synonymous with Geoffrey Hinton's neural networks, represented by tools like ChatGPT, or with smart devices like Alexa and Siri. However, not too long ago, AI was all about big data. And before that, AI was mostly associated with robots. At other points in time, AI was synonymous with chess-playing, while during the 1980s, the thing that was most associated with AI was expert systems. If anything, AI nowadays sounds more like a marketing term than a technology or a field of research.

Here, AI refers to the various *AI agents,* e.g., agents that are built using the technology commonly known as artificial intelligence. As such, throughout the rest of this thesis, the terms AI and AI agents (or 'AIs' for simplicity) will be used interchangeably unless otherwise

specified. *AI agents* are non – biological entities that can initiate their own behaviour and exhibit varying degrees of *autonomy* (e.g., ability to operate without a constant oversight), *proactiveness* (e.g., ability to display goal-oriented behaviour), *reactivity* (e.g., ability to act on their environment in a timely fashion) and *social ability* (e.g., ability to interact with other agents, including humans) (Wooldridge & Jennings, 1995). For a comprehensive definition of agent from an engineer's perspective, see Franklin and Graesser (1996).

How Humans Perceive AI

The earliest recorded attempt to understand how people perceive AI dates back to the early 1970s when Japanese robotics professor Masahiro Mori introduced a theory on people's emotional responses to embodied AI, particularly robots. This theory, later translated into English as the Uncanny Valley (UV) theory, describes how human affinity towards robots fluctuates based on their resemblance to humans (Mori, 1970; Mori et al., 2012). From them, AI served different functions by taking a variety of forms, from digitally operated robotic arms used in the car industry in the 1950s to digital computers entering the workspace and households in the early 1980s (Broadbent, 2017).The multidisciplinary fields of human-computer interaction (HCI) and human-robot interaction (HRI) , spanning computer science, engineering, psychology and cognitive science, also emerged to understand the factors that affect the interaction of humans with computer and robotic systems respectively (Goodrich & Schultz, 2008; Preece et al., 1994). Research in these fields has primarily concentrated on designing and evaluating systems (computers or robots) for use by or in collaboration with humans, with the ultimate goal to improve the computer interface, robot design and the interaction techniques more generally (MacKenzie, 2024). In psychology and cognitive neuroscience, research on human perception of AI surged following the explosion of social robotics design in the early 2000s. This growth was

largely driven by advancements in robotics and a growing interest in integrating robots into everyday life beyond factories and research labs (Mahdi et al., 2022). The study of anthropomorphism (attributing human like qualities to AI) is another area of research within social psychology and neuroscience where human perception of AI has been studied, although not always explicitly labelled as such. Over the years, AI in the literature of anthropomorphism has been described in various ways, including as a machine, a computer, or a robot that humans perceive 'as if' it were human, primarily due to design features that resemble human appearance or behaviour in some way.

The section that follows starts with an overview of what has been learned about human perception of AI through psychological studies on social robots and research on anthropomorphism of AI as the review of these two literatures informed the research questions explored in this thesis. Also, since this thesis examined the existence of stereotypes in the perception of AI through perception frameworks from social psychology—specifically, the Stereotype Content Model (SCM) (Fiske et al., 2002) and the Mind Perception Dimensions (MPD) (Gray et al., 2007)—the section that follows also includes an overview of these two perceptual frameworks.

Psychological Studies involving Social Robots

While social robotics constitutes a significant focus within HRI research, there remains no clear consensus on what specific attributes or behaviours make a robot truly social (Henschel et al., 2021). A qualitative analysis of various definitions of social robots, spanning from 2009 to 2015, revealed that social robots are typically understood as physically embodied agents with varying levels of autonomy. These robots engage in social interactions with humans, including communication, cooperation, and decision-making, with their behaviours being interpreted by

human observers as social according to prevailing norms and conventions (Sarrica et al., 2020). In studies where people's perceptions of these robots were investigated (de Graaf et al., 2015; Dereshev et al., 2019), people highlighted the ability of a two-way reciprocal interaction as the main factor for a social robot to be perceived as social. These studies also revealed that people's perceptions of social robots are influenced by their perceptions of other social actors (e.g., their friends). For instance, in a longitudinal home study, participants repeatedly compared the social robot with their friends, dwelling on the fact that the robot's lack of social capabilities make it unlikely for it to become an actual 'friend'(de Graaf et al., 2015). Studies also emphasise the novelty effect as a common pattern in social robots perception, where engagement is initially high but gradually declines as the novelty wears off over time (Leite et al., 2013; Tanaka et al., 2015), whereas, on the contrary, the findings from other studies suggest that attitudes towards social robots tend to improve over time with repeated interactions and are influenced by pre-existing attitudes (Stafford et al., 2014).

Two relatively recent literature reviews (Baraka et al., 2020; Henschel et al., 2021) highlight the growing body of research in psychology that uses social robots as research tools. This is next to other well-known application areas for social robots, including industry (Shukla & Karki, 2016), healthcare and therapy (Cifuentes et al., 2020; Dawe et al., 2019; Pennisi et al., 2016), education (Belpaeme et al., 2018), entertainment (Bruce et al., 2000; Chen et al., 2011) home (Srinivasa et al., 2010) and workplace environments (Drexler & Lapré, 2019), to search and rescue applications in hazardous locations (Kas & Johnson, 2020).

The psychological studies that have used social robots have done so primarily to address questions related to cognitive neuroscience, seeking to understand how the brain support cognitive activities including perception (Bossi et al., 2020; Thellman & Ziemke, 2020; Wiese et

al., 2017) , attention (Cao et al., 2019; Chevalier et al., 2020; Kajopoulos et al., 2021), theory of mind (Banks, 2020; Bianco & Ognibene, 2019) , and decision making (Hsieh et al., 2020; Marchesi et al., 2020). By doing so, however, they have also provided insights into how humans perceive and behave towards AI in the form of social robots. Perhaps influenced by the humanlike appearance of the social robots commonly used in these studies—such as the humanoid social robots Pepper, Nao, and Robovie (for an extensive list and classification of social robots see Baraka et al. (2020)) these insights largely focus on how humanlike appearance or behaviour of the social robot influence the cognitive processes being examined in each study. These studies suggest that both the robot's humanlike appearance and behaviour shape people's assumptions about its capabilities and influence their interactions (Abubshait & Wiese, 2017; Cross et al., 2012; Goetz et al., 2003).

For example, in a series of experiments, participants perceived more positively and collaborated more with a robot when its appearance matched the nature of the job (a more humanlike robot for a more social in nature job such as a dance instructor and a more machinelike robot in less social jobs such as a night security guard) (Goetz et al., 2003). Another study examining coordination through a joint-action task (playing a musical duet) with either a humanoid robot or an algorithm showed that perceived human-likeness in terms of both appearance and behaviour of the AI agent affected behavioural variability, which is a key sensorimotor signal of coordination. When the appearance of the AI agent matched its behaviour, coordination was increased. For example, participants exhibited lower variability in their performance when the humanoid robot made human like errors and when the algorithm made machine like errors (Ciardo et al., 2022). And, when manipulating both appearance of an agent (human vs robot) and behaviour (reliable vs random) during an eye gazing task, appearance was

found to impact attitudes towards the agent while behaviour had a stronger impact on performance (Abubshait & Wiese, 2017).

Interestingly, at the brain level, the agent's appearance (human vs Lego robot) was not found to influence the preferential engagement of the action observation network (AON) which was engaged more robustly to robot-like motion than natural human motion for both agents (Cross et al., 2012), while in another neuroimaging study, the right temporoparietal junction (rTPJ) was involved only when both an agent looked like and was believed to be human, suggesting that the rTPJ is biologically tuned to regulate the tendency to imitate an observed agent, but only when the agent both appears human and is believed to be human (Klapper et al., 2014). It should be mentioned that coupled with psychologists, social robotics engineers have also observed the impact of a robot's appearance and behaviour on people's perceptions, noting that more human-like robots can raise people's expectations and thus risk falling into the Uncanny Valley (Duffy & Joue, 2004; Pandey & Gelin, 2018). To mitigate these effects, some engineers have thus chosen alternative morphologies, such as animal-like robots (Collins et al., 2015).

The impact of an AI's humanlike appearance and behaviour on perception of AI is not the only area where cognitive neuroscience research using social robots provides valuable insights. Studies have explored how a robot's features beyond human-likeness influence perception and interaction, suggesting that other factors may be more effective in triggering human-like socio-cognitive processes (Cross et al., 2016; Jastrzab et al., 2024; Liepelt & Brass, 2010; Stanley et al., 2007; Stenzel et al., 2012; Tsai & Brass, 2007). For instance, belief about an agent's animate origins (e.g., believing that agent's motion originated from human motion capture vs computer animation) was shown to modulate engagement of person perception and mentalizing networks,

while the level of human-likeness in the appearance of the agent (human vs robot) had less impact on social brain networks (Cross et al., 2016). Likewise, using a moving dot stimulus, Stanley et al. (2007) demonstrated that the perception of animacy was primarily influenced by knowledge of the origin of the dot's movement (whether participants were told that it was generated by a human or a computer), rather than by the motion properties themselves. And in a study where participants played rock-paper-scissors (RPS) games against human and AI agents with varying degrees of human-like appearance (a humanoid robot, a mechanoid robot, and a computer algorithm), the engagement of the mentalizing network increased as the robot's appearance became more human-like. However, perceived socialness of the agent—evaluated based on traits such as fun, sympathy, competitiveness, success, strategy, intelligence, and overall competitiveness—explained the differences in mentalizing network engagement more effectively than the agents' physical appearance alone (Jastrzab et al., 2024). Overall, these studies suggest that human-likeness alone may not fully explain which robots are perceived as more desirable social partners or predict how socially engaging people find them. Other features of the AI agent, such as knowledge cues (e.g., understanding the origin of its behaviour) or its level of socialness, and possibly other factors—since this is still an emerging research area—may play a more significant role in triggering human-like social-cognitive processes.

Moreover, a number of psychological studies using social robots have explored mind attribution towards robots (Broadbent et al., 2013; Klapper et al., 2014; Laban et al., 2021; Özdem et al., 2017; Stafford et al., 2014; Wiese et al., 2012; Wykowska et al., 2014). Overall, these studies suggest that the degree of the AI's human-likeness in appearance affect people's mind attribution. For instance, people rated a robot with a face on its screen as more 'minded' compared with robots with no or a silver face (Broadbent et al., 2013). In another study

examining self-disclosure for psychological health to an embodied AI (a humanoid robot), a disembodied AI (Google Nest Mini), and a human, while participants reported no difference in perceived agency between the humanoid and disembodied AI, they attributed higher levels of experience to the humanoid robot (Laban et al., 2021). And politeness norms were triggered more often by a humanoid compared to a mechanical robot (Babel et al., 2022), suggesting that human-likeness triggers more intentionality ascription to AI. Research has also explored the influence of contextual factors on attributing mind to robots. Findings suggest that aspects such as a robot's function (e.g., whether it is framed as having social or economic value), its behaviour (e.g., speech and nonverbal cues), and the way it is introduced (e.g., framing) can significantly impact attributing mind to robots (Wallkötter et al., 2020; Wang & Krumhuber, 2018). Also, observing someone interacting socially with a robot can enhance the adoption of an intentional stance. For instance, people who collaborated with a humanoid were more likely to ascribe intentions towards it after the interaction than people who did not collaborate with it (Abubshait et al., 2021). The attribution of intentional traits towards a robot was also higher after social compared with non-social priming. In the social priming condition, before evaluating different types of robots, participants were told that the robots represent types of agents that they will interact with in the coming decades(Spatola et al., 2021).

Several neuroimaging studies using social robots have also examined the degree to which the social cognition brain regions - also known as the mentalizing network (Schurz et al., 2014)- which has evolved to interpret other people's thoughts, intentions and actions is also engaged when processing the thoughts, intentions and actions of non-human social partners such as robots. Through a range of tasks, including economic games (Chaminade et al., 2012; Krach et al., 2008; Takahashi et al., 2014), attention cueing tasks (Özdem et al., 2017; Wiese et al., 2018)

and empathy tasks (Cross et al., 2019; Rosenthal-Von Der Pütten et al., 2014), these studies demonstrate that the mentalizing brain network is activated during human-robot interactions, albeit to a lesser degree than during human-human interactions. Brain regions that have been reported as less or not activated by robots include the temporoparietal junction (TPJ) (Hmamouche et al., 2020; Kelley et al., 2021; Rauchbauer et al., 2019; Wang & Quadflieg, 2015), medial prefrontal cortex (MPFC) (Hmamouche et al., 2020), and dorsolateral prefrontal cortex (DPFC) (Rauchbauer et al., 2019).

In addition, the combined findings of two literature reviews (Lee & Harris, 2014; Vaitonytė et al., 2023) looking at studies reporting comparisons of brain processing of responses to human and to AI targets (both embodied; such as robots and disembodied; such as computer algorithms) suggest that, at a neural level, AI is not processed homogeneously. Embodied AI drives increased engagement in certain brain regions relative to humans, such as the precuneus and the ventromedial prefrontal cortex (VMPFC) while interactions with disembodied AI do not lead to increased activation of any social brain regions relative to human.

Overall, the above literature demonstrates that human perception of AI is not homogeneous, providing evidence that variations in perception exist, influenced by different design features (such as form, motion, or socialness) or contextual differences (such as variations in knowledge cues, intended purposes, or the way the robot is presented). And these variations are, in turn, supported by distinct brain mechanisms. As researchers have similarly concluded a 'one size fits all machines' type of cognition is unlikely (Cross & Ramsey, 2021). However, this raises the question: what differentiates AI agents in human perception, across design, embodiment, and contextual differences? This motivated the effort to map AI perception across

different AI agents and formed the basis of the first research question (*RQ1*) that this thesis aims to address.

Next, the following section focuses on the literature regarding the psychological phenomenon of anthropomorphism of AI. This body of research has extensively explored the factors that influence the human tendency to attribute human-like qualities to AI and the effects of these attributions on the quality of human-AI interaction (HAI). We examine this literature as it is a key area of research within the broader field of AI perception. Additionally, we drew from this literature to further investigate and address *RQ1*.

The Study of Anthropomorphism of AI

Anthropomorphism is the psychological term used to refer to the human tendency to think about the mind or mental states—such as thoughts, feelings, intentions, and motivations—of real or imagined non-human entities, including nonhuman animals, natural forces, religious deities, objects, and mechanical or electronic devices (Epley et al., 2007). This tendency is so strong that people even readily describe moving geometrical shapes as having intentions and feelings (Heider & Simmel, 1944) and is already present in young children (Manzi et al., 2020). This tendency makes sense as being human is what people know. In other words, when interacting with unfamiliar non-human entities, people may use their knowledge of themselves as a basis for understanding those entities (Epley et al., 2007). In addition, people are used to thinking about the mind of other people to explain and predict their actions (Fiske, 1991; Fiske & Taylor, 2020). As such, they may apply the same strategy to try to understand unexplainable actions of nonhuman entities. Also, anthropomorphism may provide a form of increased sense of belonging and imagined social connection when the perceiver is feeling socially isolated (Eyssel & Reich, 2013) .

The psychological phenomenon of anthropomorphism has attracted significant interest within AI research, particularly in examining how design features of the AI agent —such as humanlike appearance or behaviour—influence people's tendency to anthropomorphise AI with a focus on how this tendency, in turn, affects the overall HAI. Studies in this area have explored various aspects of the influence of anthropomorphism, including its effect on people's trust (Eyssel et al.; Goudey & Bonnin, 2016), attitudes ((Li & Sung, 2021; Liu et al., 2019; Wagner et al., 2019), acceptance (Yao et al., 2025), perceived threat (Yogeeswaran et al., 2016) and empathy (Riek et al., 2009) . The design features of an AI agent that have been manipulated across the literature looking at anthropomorphism depend heavily on the specific context and the research question under examination. For instance, in the case of autonomous vehicles, researchers have manipulated human likeness by giving an identity, gender, or a human voice to the vehicle (Waytz et al., 2014). In the study of chatbots, other methods to increase perceived human likeness have been employed, such as incorporating a human-like face or increasing message interactivity (Go & Sundar, 2019). In the case of embodied AI, such as robots, various design features have been manipulated to resemble humans, including appearance and behaviour, with a particular emphasis on the effect of motion (e.g., eye, head, or body movement), as well as identity and presence (e.g., physically present vs. telepresence (Thellman et al., 2022).

Most of the studies looking at anthropomorphism in robots report that the tendency to anthropomorphise increases with human like appearance (Abubshait & Wiese, 2017; Martini et al., 2016; Xu & Sar, 2018). For instance, when explicitly asked, participants self-reported anthropomorphising machines with human appearance more than other types of machines (Xu & Sar, 2018). And increased human like appearance of a robot has been found to be associated with increased activity in the mentalising network (Krach et al., 2008). Studies also report stronger

tendencies to anthropomorphise a robot when it exhibits human like behaviour such as i.e., gaze (Abubshait & Wiese, 2017) or exhibit unpredictable behaviour (Eyssel et al., 2011) and display emotion (Złotowski et al., 2014). For instance, when participants anticipated that they would interact with an unpredictable robot they made more anthropomorphic inferences about its behaviour (Eyssel et al., 2011). And when the humanoid robot NAO displayed emotion by making characteristic sounds, such as 'Yippee', and gestures, such as rising hands, his emotional display was found to make participants perceive the robot as more humanlike (Złotowski et al., 2014). Furthermore, assigning an identity to a robot, such as describing the robot as an ingroup, as indicated by its name and a country of production, has been shown to increase the tendency to anthropomorphise it (Eyssel & Kuchenbrandt, 2012). And participants were more likely to anthropomorphise a robot when it is physically present rather than tele present (Kiesler et al., 2008; Straub, 2016).

In addition to design features of the robot, human factors—such as age and culture—have been shown to play a role in the tendency to anthropomorphise AI within the specific context of the interaction (Pak et al., 2014; Takahashi et al., 2016; Tan et al., 2018; Thellman et al., 2022). For instance, stronger tendencies were reported among Japanese than Western participants in an online survey where participants were asked to evaluate explicit mental capacities (e.g., capacity to feel hunger) of robots (Takahashi et al., 2016). And the tendency to anthropomorphise robots has been shown to be stronger in children than adults (Okanda et al., 2021) and stronger to younger than older children(Manzi et al., 2020).

A methodological challenge highlighted throughout the robot literature of anthropomorphism is that findings tend to vary depending on the type of measure used, such as i) self-report (verbal), ii) behavioural (non-verbal), or iii) and neural measures. Self-report

measures have been shown to be more likely to report a weaker tendency to anthropomorphise compared to behavioural data (Thellman et al., 2022). For instance, while participants described robots using mental state terms during an interview (behavioural measure), they later denied that robots possess various mental capacities when asked explicitly in a post-interview questionnaire (Fussell et al., 2008). And differences have been also found between self-report and neural measures. For instance, participants rated a robot appearing to be electrocuted as experiencing various levels of pain, yet no corresponding activation in the participants' pain matrix was detected during the observation of the electrocution (Cross et al., 2019). We revisit this methodological challenge stemming from the differences in the types of measurements in the end of this chapter, as it represents one of the key challenges identified in the study of AI perception. Below, an overview of the most frequently used framework for measuring anthropomorphism and mind attribution in robots through self-report measures—the Mind Perception Dimensions (MPD) model—is provided. Our focus on self-report measures for capturing perception of different AI agents was motivated by the desire to address *RQ1* through the collection of self-reported data, at least as an initial step.

The Mind Perception Dimensions (MPD) Model

According to Gray et al. (2007), attributing mind consists of two dimensions: the capacity of agency (covering one or several of the following capacities: self-control, morality, memory, emotion recognition, planning, communication, and thought) and the capacity for experience (covering one or several of the following capacities: hunger, fear, pain, pleasure, rage, desire, personality, consciousness, pride, embarrassment, and joy). In the initial mind perception dimension study by Gray et al. (2007), researchers had a large sample of participants (N=2040) making pairwise comparisons on a 5-point scale of 13 characters on mental capacities (e.g.,

capacity to feel pain) and on six personal judgments (e.g., '*which character do you like more?*'). The 13 characters consisted of seven living human forms (7-week-old fetus, 5-month-old infant, 5-year-old girl, adult woman, adult man, man in a persistent vegetative state, and the respondent him- or herself), three nonhuman animals (frog, family dog, and wild chimpanzee), a dead woman, God, and a robot (the social robot Kismet). Their findings revealed that the entities examined could be arrayed into a two -dimensional space defined by the dimensions of agency and experience. The robot in this study was attributed some degree of agency but very low degree of experience.

The MPD model has since influenced the operationalisations of anthropomorphism and mind attribution in AI research, particularly HRI research (Kühne & Peter, 2023). More specifically, depending on how researchers define and operationalise the concepts of agency and experience, studies have utilised all, some, or variations of the mental capacities examined in the initial Gray et al. (2007) study (i.e., the capacity to think, the capacity to self-control, the capacity to feel pain, fear, pleasure etc.) as measurement items for assessing the extent to which participants tend to anthropomorphise AI. These studies examine how the extent to which participants attribute agency and experience to AI influences human behaviour in a large array of tasks, including collaborating with an AI agent - often humanoid or mechanical robots (Ferrari et al., 2016; Fraune, 2020; Fraune et al., 2020; Trovato & Eyssel, 2017), making decisions with assistance from a humanoid robot (Lefkeli et al., 2018), helping a robot (Tanibe et al., 2017), observing harming behaviour towards robots (Ward et al., 2013; Wieringa et al., 2024), and attributing blame or responsibility to an AI agent which is often a robot or, sometimes, a computer in competitive game settings (Kawai et al., 2023; Miyake et al., 2019).

Although MPD model's widespread use for measuring people's tendency to anthropomorphise and attribute mind to AI, the agency- experience distinction it introduces is far for ideal. First, there is considerable variation in how researchers separate and operationalise the concepts of agency and experience with no universal agreement on the definitions of agency or experience. The lack of consensus in defining the two fundamental concepts of this model inevitably leads to conceptual confusion and vagueness (Kühne & Peter, 2023). It also makes the findings on the effect of anthropomorphism of AI across studies somewhat unclear, as the inconsistency introduced by varying definitions complicates the consolidation and comparison of results between studies.

Secondly, the MPD model is not the only operationalisation of anthropomorphism that has been proposed and used in studies exploring anthropomorphism, particularly in HRI settings. Both unidimensional and multidimensional operationalisation have been proposed, yet with no clear consensus on this matter either. E.g., there is no clear consensus on whether anthropomorphism of AI is best understood through a unidimensional or multidimensional approach, and on which dimensions should be included in its conceptualisation (Kühne & Peter, 2023). For instance, indicators such as 'humanlike' 'lifelike,' and 'natural' have been used as unidimensional measure of anthropomorphism (Bartneck et al., 2009). Another frequently used operationalisation of anthropomorphism is rooted in the dual model of dehumanisation (Haslam & Loughnan, 2014; Haslam et al., 2008), which distinguishes two categories of mental capacities that may or may not be attributed to others: uniquely human characteristics, which set humans apart from other nonhuman animals (e.g., reason), and human nature characteristics, which are shared with other nonhuman animals (e.g., curiosity). Several studies use this distinction to measure anthropomorphism of robots (Eyssel et al., 2011; Ferrari et al., 2016; Salem et al.,

2013). For instance, in a study where researchers investigated the effect of unpredictable behaviour by a robot on the tendency to anthropomorphise it, anthropomorphism was measured by having participants evaluating the extent to which traits of human uniqueness (i.e.., rationality, refinement, civility) and human nature (i.e., openness, warmth, emotionality) were attributed to the robot. And more recently, new multidimensional conceptualisations of anthropomorphism have been proposed in an effort to distinguish perceptions related to shape and movement which are seen as precursors to anthropomorphism from perceptions that lead to attribution of personality and moral character to an AI agent which are seem as consequences of anthropomorphism(Kühne & Peter, 2023).

Thirdly, although MPD model is often used to operationalise the human tendency to anthropomorphise robots, it should be treated as only a proxy that implicitly measures anthropomorphism as, the exact relationship between mind perception (also referred to as mind attribution in the literature) and anthropomorphism is not yet clear. In the case of robots, it has been suggested that anthropomorphism is a two-step process in which an individual first engages in higher-level reasoning to adopt a mentalistic or intentional stance toward the robot before lower-level attribution processes are activated (Wiese et al., 2017; Wykowska et al., 2014). However, as Thellman et al. (2022) note in their meta-analysis, this two-step process appears contradictory to observations from other studies, where people spontaneously attribute mental states to robots while they self-report that they do not believe that these entities possess a mind (Banks, 2020; Fussell et al., 2008). While whether anthropomorphism is a two-step process, with mind attribution involved or not, remains an open question that requires further research for more conclusive insights- and despite the fact that researchers often use the terms anthropomorphism and mind perception (or mind attribution) interchangeably - the current

understanding of the relationship between anthropomorphism and mind perception calls for caution in treating the MPD's categorical process of attributing mind as a proxy measure rather than a direct indicator of the psychological phenomenon of anthropomorphism.

Despite MPD model's limitation in fully capturing anthropomorphism and the definition confusion surrounding agency and experience that abounds, it is a model that has been widely used in the literature. And setting aside for a moment the definitional confusion (which complicates the aggregation of results across studies), it has proven effective as a proxy for capturing anthropomorphism within the scope of each individual study. Also, it originates from a categorisation of a broad range of agents that are not all human or humanlike (from God to chimpanzee, baby, and a deceased person). In this regard, MPD model's ability to capture human perception of diverse perceptual targets— which do not necessarily resemble humans or are human— has been tested and validated. Such a 'non-human-centric' (in terms of the range of perceptual targets examined) model could be especially useful for mapping perception across the wide variety of AI agents which are not all necessary designed to resemble the human (i.e., GPS application, drones). This motivated us to consider it as a potential candidate for addressing RQ1 of this thesis. Next, we turned to social psychology in search of a person perception frameworks to address *RQ1*.

The Stereotype Content Model (SCM)

To map human perception of AI, it may be helpful to explore other attempts in the literature that have similarly attempted to map human perception. If such mapping exists and have been empirically validated, it could be worth assessing its applicability for AI perception. After all, in addressing RQ1, an approach would be to consider any framework that maps human perception of any perceptual target, as long as there is some theorising about that target's share

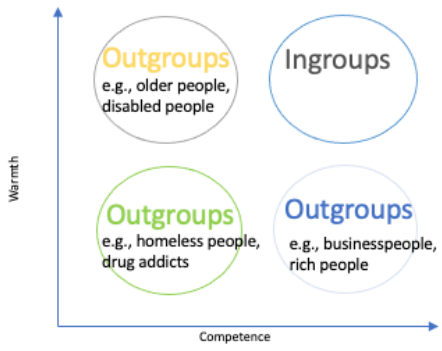
space with the AI perceptual target. This is because, although in such a framework the perceptual target might be different (e.g., humans, animals, objects, or AIs), the perceiver remains always the same (human). This reasoning led us to explore social psychology, which has long studied how people perceive other humans. It's worth noting here that we could have also turned to research on object or animal perception (e.g., how people perceive objects or non-human animals respectively) in search of such a framework. Object perception, in particular, has been proposed as a promising avenue to explore the shared space between social machines and objects(Cross & Ramsey, 2021). However, we chose to focus on person perception, which has received more extensive research attention than animal or object perception and has a well-validated person perception framework, the Stereotype Content Model (SCM).

Based on the SCM, there are two things that people care about when they encounter other people, one is their perceived warmth (how friendly or threatening this person is) and the other is their perceived competence which describes their ability to follow through on their intentions (Cuddy et al., 2008; Fiske et al., 2007). Person perception can be therefore mapped onto a two – dimensional space defined by these two-core dimension. Warmth encompasses traits such as friendliness, good-naturedness, sincerity, and warmth, while competence includes traits such as competence, confidence, and skill. These dimensions originate from classic person perception theories based on which certain traits tend to separate into clusters (Asch, 1946; Rosenberg et al., 1968) and were first validated using U.S. samples (Fiske et al., 2002). Subsequent research has then demonstrated their applicability across diverse cultural contexts, including European and East Asian countries such as Hong Kong, Japan, and South Korea, reinforcing their universality despite cultural differences (Cuddy et al., 2009).

SCM predicts how people form differentiated prejudices towards different societal groups based on the appraisal of their intentions for help or harm (warmth) and their capacity to enact those intentions (competence). It differentiates ingroups from outgroups into four clusters on the competence x warmth space (Figure 1), with each combination eliciting four distinct emotional response toward social groups: pride, envy, pity, and disgust. Ingroups and allies are perceived as high on both dimensions and receive pride and admiration, whereas outgroups fall into the other three quadrants where they are stereotyped as either competent or warmth, but not both, and elicit ambivalent emotions. For instance, groups stereotyped as competent but not warm (e.g., rich people) elicit envy and jealousy, groups stereotyped as warm but not competent (e.g., elderly people) elicit pity and sympathy while the most extreme outgroups, that are perceived as low on both competence and warmth (e.g., homeless people, drug addicts), elicit disgust and contempt (Fiske et al., 2007).

Figure 1

The Stereotype Content Model (SCM)



Moreover, perceptions of warmth and competence have also be found to predict behavioural responses to groups (A. J. Cuddy et al., 2007). Groups that are perceived as high on both SCM dimensions receive both active and passive facilitation. In contrast, groups in the

mixed quadrants experience a combination of positive and negative behaviours: those eliciting envy face passive facilitation and active harm, while those evoking pity receive active facilitation and passive harm. Groups perceived as low in warmth and competence, and eliciting disgust, are subjected to both active and passive harm.

It should be noted here that, similar to MDP, the SCM conceptual framework—like any other conceptual framework—is not without its limitations. When applied to the study of AI perception, it will therefore bring along inherent challenges. A widely debated issue in the literature is whether SCM adequately captures the perception of morality under the warmth dimension or if morality should be considered an independent dimension that plays a distinct role in person perception (Leach et al., 2015). The morality dimension encompasses traits such as honest and trustworthy whereas warmth includes characteristics indicative of human benevolence such as good-natured, friendly, warm, well-intended, helpful, and trustworthy. Thus, for some researchers, the warmth dimension is akin to the conceptualisation of morality, and is therefore referred to as the 'moral' dimension (Wojciszke, 2005). However, for other researchers, such broad conceptualisation of morality may obscure the distinct role of trustworthiness in judgments about others' morality. Indeed, out-group members can be seen as warm, friendly, and likeable without being seen as moral people who can be trusted (Leach et al., 2007, 2008). Overall, there is a debate within the SCM literature regarding the importance of distinguishing perceptions of trustworthiness from perceptions of less obviously moral aspects of out-group benevolence, such as sociability. Additionally, the *competence* x *morality* framework has also been used to map person perception. For instance, Phalet and Poppe (1997) examined the role of competence and morality in stereotypes across- six eastern-European countries and found that stereotypes could be mapped on the *competence* x *morality* space revealing four

quadrants (e.g., moral and incompetent, moral and competent, immoral and incompetent, moral and competent).

Despite the lack of agreement on the definition of the warmth dimension, the SCM is a well-validated framework for mapping person perception. As such, it could also be tested for its applicability in mapping AI perception. It is by definition a 'human centric' model since the perceptual targets that it maps are all humans. In this regard, it would be interesting to explore the extent to which this framework applies to AI as a perceptual target, shedding light on whether and to what degree people rely on 'person dimensions' to differentiate between various AI agents. This motivated us to consider SCM as another potential candidate for addressing RQ1 of this thesis.

Next, we shift our focus to the literature on the determinants of trust in AI, as the second part of this thesis examines how people make decisions with AI. The following section also introduces the second and third research questions explored in this thesis, RQ2 and RQ3 respectively.

How Humans Trust Technology, Automation, and AI

How humans trust technology has been widely discussed in the literatures of computer science, human–computer interaction, robotics, management, decision sciences, marketing, and psychology. Synthesising knowledge dispersed across the literature of various disciplines presents a significant challenge. This challenge is further exacerbated by two key issues: firstly, the varied ways in which *acceptance* and *trust* are operationalised in the literature, and secondly, the lack of consensus on what is understood by the term AI, when studies are looking specifically at the factors that drive trust in AI (Glikson & Woolley, 2020; Kelly et al., 2023). As highlighted by Kelly et al. (2023) in their systematic review of factors influencing trust in AI, this ambiguity around the term AI prevents us from '*fully understanding if people accept real AI*

or the idea of AI.' To make things even more challenging in synthesising findings across literature, studies do not always provide their participants with a clear definition of AI, leaving room for different interpretations (Kelly et al., 2023). In an effort to provide additional context for the work presented in this thesis in terms of how people trust and make decisions with AI (Chapters 3 and 4), an overview of the most commonly used theoretical framework for studying trust in technology, along with an examination of trust trajectories in AI and the key factors that are frequently reported in studies as influencing trust in AI.

The Technology Acceptance Model (TAM)

Acceptance denotes a personal decision. However, when it comes to AI, the level of human agency in accepting AI can vary, as AI may operate subtly without people's knowledge or awareness. For example, purchasing an Alexa device involves an understanding that some sort of AI technology is part of the product, which the user knowingly accepts. In contrast, during an online customer service interaction, an AI chatbot may present itself as a human agent or equally, it may leave it unclear to the user whether they are interacting with a human or a chatbot, leading the customer to believe they are conversing with a person rather than an AI. In the latter example, the presence of AI is subtle or opaque to the individual, making the acceptance of AI more of an involuntary action, or at least not as conscious and deliberate a choice as it is when buying an AI-powered voice assistant.

The Technology Acceptance Model (TAM) or variations of it has been extensively used across disciplines as a framework for assessing whether and how people accept AI in their lives across different domains (Kelly et al., 2023). Although initially proposed to describe the factors that predict the intention to use any new technology (Davis, 1986; Davis, 1989), rather than AI specifically, it has since been applied to examine acceptance of AI in domains such as i.e.,

healthcare, education and customer service. It is important to notice here that TAM predicts the intention of use rather than actual behaviour, although it postulates that intention ultimately drives actual behaviour (Davis, 1986; Davis, 1989). Based on TAM, *perceived usefulness* (PU)- e.g., the degree to which people perceive a technology to be useful in their everyday lives- and *perceived ease of use* (PEOU) - e.g., the degree to which people perceive that the technology will be effortless to use- are the two fundamental predictors of people's intentions or otherwise, willingness to use a new technology. *Familiarity* with the technology has also been found to influence PEOU, diminishing PEOU predictive power in technology adoption (Liu et al., 2016; Lunney et al., 2016). This may be because frequent use of a technology reduces the significance of ease of use, as familiarity compensates for any initial difficulty of use.

To gain deeper insights into and predict more accurately the integration of AI into various domains of people's lives, researchers have frequently built upon the TAM and extend it by incorporating additional variables. These include variables that attempt to operationalise individual factors such as trust, attitudes, knowledge, performance expectancy, and effort expectancy, as well as external factors like social norms and social influence. Especially, social norms and social influence has been found to positively predict intentions to use AI across various industries such as customer service (Gursoy et al., 2019) and healthcare (Lin et al., 2021) while their role in studying trust in AI among cohorts highly susceptible to the influence by their peers such as adolescents and young adults (Knoll et al., 2015) has been recommended (Kelly et al., 2023).

Another often-used extension of TAM is the Unified Theory of Acceptance and Use of Technology (UTAUT) which on top of the variables mentioned above, also accounts for the moderating effects of factors such as gender, age, voluntariness of use, and prior experience

(Venkatesh et al., 2003). This model has been found to explain approximately 60–70 % of the variance in behavioural intentions to trust technology across cultures (Thomas et al., 2013; Venkatesh et al., 2003).

Finally, to study the acceptance of AI- enabled products and services, whose advancements occur at a much rapid pace compared to other emerging technologies (Sohn & Kwon, 2020), a distinct framework called the AI Device Use Acceptance Model (AIDUA) has been introduced and applied to the study of AI in service delivery (Gursoy et al., 2019). Building upon the TAM framework and tested using data from potential consumers, the AIDUA model proposes that individuals undergo a three-stage process: primary appraisal, secondary appraisal, and the outcome stage. During the primary appraisal stage, consumers evaluate the AI based on factors such as social influence, hedonic motivation (e.g., the anticipated enjoyment or satisfaction derived from using the AI), and anthropomorphism. Following this, in the secondary appraisal stage, they consider the AI's perceived performance expectancy and effort expectancy. The outcome of this deliberation shapes an emotional response towards the AI, which subsequently determines the outcome stage—where the consumer either expresses a willingness to adopt the AI or rejects its use.

A common critique of studies employing one of the aforementioned TAM models to examine AI adoption is their reliance on measuring intentions rather than actual behaviours. Most studies use dependent variables such as willingness to use, intention, or acceptance, instead of observed behaviour. For instance, in a systematic literature review of 60 studies on AI acceptance across various domains, Kelly et al. (2023) found only seven measuring actual behaviour. This is not an inherent issue of the TAM model or the TAM- based themselves, as they measure intentions by design. However, problems may occur when intentions fail to

translate into behaviours, and conclusions about actual behaviour are drawn based solely on measured intentions. This reliance can be problematic as attitudes do not always match behaviours (Festinger, 1957; Gollwitzer & Sheeran, 2006). This critique of TAM in the study of acceptance and trust of AI motivated the experimental paradigms used in studies in Chapters 3 and 4. In these studies, we chose to measure trust using both self-report and behavioural measures, knowing that findings across different types of measurement may vary as self-report intentions do not always translate to behaviours and vice versa.

Trust Trajectories of AI

There is a trajectory in the formation of interpersonal trust (Rempel et al., 1985). For example, people generally hesitate to trust a total stranger. Building trust between individuals often takes time and effort, and yet a single event can shatter that trust, requiring significant effort to restore it. Similarly, trust in AI can have a dynamic nature. It can start from very positive due to people exhibiting a positivity bias in trust of a novel AI system and then suddenly dissolve due to errors or unpredictability of the system (Dzindolet et al., 2003; Madhavan & Phillips, 2010). Similarly, the reverse order might also be true. Initial scepticism can shift to a positive perception following the first interaction, with this positivity growing over time, after repeated interactions and as the system maintains reliability and predictability.

The AI trust trajectories proposed by Glikson and Woolley (2020) are based on an extensive meta-analysis of over 150 peer-reviewed empirical studies on human trust in AI across multiple disciplines, including computer science, human-computer interaction, human factors, information systems, robotics, management, marketing, and psychology. The findings indicate that human trust in AI follows three distinct trajectories which are determined by the form of the AI representation (e.g., robot, virtual, or embedded) and its level of machine intelligence (i.e., its

capabilities). For embodied AI such as robots, trust increases following direct interaction (Haring et al., 2016; Ullman & Malle, 2017). For example, participants who drove a partially autonomous car expressed greater trust in its capabilities compared to those without such experience (Waytz et al., 2014). And even a short interaction with a robotic pet was found to significantly improve attitude towards it (Bartneck et al., 2007). On the other hand, for virtual AI such as chatbots or avatars, e.g., representations in which the AI may or may not have a physical representation but has a distinguished identity, the trajectory suggested in most of the studies reviewed is the reverse to that of robotic AI. It stars with a demonstrated high trust which then declines usually due to reasons of declining reliability (De Visser et al., 2016) or mismatch between the AI agents' human-like representation and their actual capabilities (Mimoun et al., 2017).

For example, in a field study, Mimoun et al. (2017) analysed data from virtual agents on commercial websites and saw a decline in use over time. They suggested this was due to a mismatch between the agents' human-like appearance and their actual capabilities, leading to user frustration and abandonment. The human-like appearance of AI can raise user expectations for high-level intelligence, which often doesn't align with the technology's true capabilities. Embedded AI, e.g., AI without physical representation as is the case with algorithms embedded in different applications (such as i.e., a GPS or a search engine), follows the same trajectory as virtual AI. Similar to virtual AI, laboratory-based studies have shown that people generally display high initial trust in embedded AI when it functions as an algorithmic decision-making tool (Dietvorst et al., 2015; Manzey et al., 2012). However, this trust tends to diminish when the AI makes errors and then trust restoration is required. Lack of transparency of an embedded AI

(an Uber algorithm, in particular) has also been shown not only to reduce trust but also encourage Uber drivers to engage in attempts to game the system (Lee et al., 2015).

Finally, for all three trajectories, there exists a minimum level of machine intelligence required for the AI to effectively perform the desired task within a specific context. This level establishes the lowest threshold of trust necessary for use and influences the trust trajectories. For instance, as the level of machine intelligence increases, the minimum required trust also rises.

An inherent limitation of the AI trust trajectories is that it is skewed towards the findings of the majority of studies reviewed. As the authors point out and illustrate with examples in their review, there have also been studies that observed reverse patterns in all three representations (e.g., going from high to low trust in robotic AI and starting from low trust that then is increased in virtual and embedded AI). Another limitation, again inherent to the trust trajectories, is that it is based on a snapshot taken in 2020 by looking at studies conducted during the period from 1999 to 2019. However, since then, different, and more sophisticated AI has emerged (e.g., ChatGPT was released in 2022). People's experience and familiarity with AI has also gradually increased since 1999. This means that if we were to take a most updated snapshot now (year 2025) we might have observed different trajectories depending on the findings of more recent studies. The above limitation underscores the dynamic nature of findings on trust in AI and emphasises the need to monitor trust trajectories over time. We further elaborate on this inherent challenge in the study of how people perceive and trust AI in the final section of this chapter.

Factors that affect Trust in AI

Trust is a subjective attitude that involves an individual's willingness to accept vulnerability based on positive expectations of another's behaviour (Chang et al., 2017; Glikson

& Woolley, 2020; Mayer, 1995). Trust, as such, inherently encompasses characteristics such as vulnerability, uncertainty, and risk, alongside the belief in a high probability of a favourable outcome. The latter is often rooted in positive perceptions about the other party's competence, benevolence, and reliability (Zerilli et al., 2022). When applied to AI, trust represents an individual's willingness to rely on the technology, grounded on the belief that it possesses attributes capable of effectively addressing their concerns (Chang et al., 2017). For example, trusting 'a drive finder application to find a driver for getting with safety to one's destination.

Empirical research on trust in AI spans across different disciplines, in a time period of over 20 years, with AI having different forms and purposes of use. Consequently, synthesising evidence is a challenging task. Systematic meta- analyses (Burton et al., 2020; Glikson & Woolley, 2020; Jussupow et al., 2020; Kelly et al., 2023) identify factors such as *transparency* (the level to which the 'inner logic' or the operating rules of AI are apparent to the user), *reliability* (the level to which AI exhibits the same and expected behaviour over time), *task characteristics* (the level to which the task is perceived as technical and requiring data analysis), *tangibility* (the level of physical representation of the AI), and *immediacy of behaviour* (i.e., personalisation, responsiveness and adaptiveness of the AI) as crucial in developing cognitive trust, while anthropomorphism of AI has been shown to be key to developing emotional trust (Glikson & Woolley, 2020). Cognitive trust refers to the evaluation of characteristics of the trustee such as i.e., competence or reliance coupled with the evaluation of situational features whereas, emotional trust involves any affect elicited by the AI that can affect assessments of its trustworthiness (Komiak & Benbasat, 2006; Schoorman et al., 2007). And in the case of advanced, sophisticated AI trust often necessitates a '*leap of faith'* as it required trust in

processes that are neither directly observable nor easily comprehensible (Hoff & Bashir, 2015; Lee & See, 2004).

To ease the above '*leap of faith'* or eliminate the need for it altogether, much of the AI research community's focus in recent years has been on making AI transparent, as well as explainable and interpretable— terms that are often used interchangeably (Angelov et al., 2021). Indeed, from the identified factors influencing trust in AI, transparency can be challenging to define and even more challenging to achieve. Transparency also becomes problematic when AI such as e.g., Deep Neural Networks (DNNs) e.g., algorithms that are inspired by the way neurons process information in the brains, are 'black boxes' to the people affected by them and sometimes, even to the people developing and deploying them. This opacity is due to the complexity of the function or the model that these algorithms implement and the fact that these algorithms are trained using machine learning techniques, rather than being explicitly programmed (LeCun et al., 2015).

DNNs are behind recent advances in AI such as image classification, language production, translation and navigation (LeCun et al., 2015; Russell & Norvig, 2021) while efforts to alleviate their opacity are typically discussed in terms of transparency, interpretability, and explainability. Although there is little agreement about what these key concepts mean, a key aspect in all the efforts to alleviate the opacity of 'black box' AI is the creation of various types of explanations about how the AI works or why a specific decision was made such that are understandable to human users, even when they have little technical knowledge (Fleisher, 2022) . Some of these efforts are post hoc explainable (XAI) methods. XAI methods are attempts to explain black box AI, such as e.g., DNNs, by building a second 'explanation' model that approximates the behaviour of initial black box model while being as simpler type of model, e.g.,

one that is more understandable to humans than the original black box model. These methods are 'post hoc' because the explanation model is used to explain the black box system after its use, without altering the original system. For instance, the *Linear Interpretable Model-agnostic Explanations* method, known as LIME, is a feature importance-based XAI method (Ribeiro et al., 2016).

Based on Hoff and Bashir (2015) systematic review, a more holistic view of the factors that influence trust in automation reveals three layers of trust: dispositional, situational, and learned trust. Dispositional trust represents an individual's tendency to trust automation, independent of context and system used. Factors that have been identified to influence dispositional trust include culture, age, gender, and personality. For instance, Merritt and Ilgen (2008) , using the Big Five personality traits, showed that extroverts exhibit a greater propensity to trust automation than introverts do. Situational trust depends on aspects of the external environment (i.e.., nature or framing of the task, the norms about the use of the system in the specific context), aspects of the automation (i.e., the system's complexity) and internal, context-dependent characteristics of the human (i.e., subject matter expertise, self-confidence, mood, and attentional capacity). For instance, the environment helps the evaluation of risks and benefits associated with using the automation as shown in a study where participants trusted route-planning advice from a GPS less when the situational risk increased with the addition of more driving hazards in the experimental driving setting (Perkins et al., 2010). Finally, learned trust can be trust stemming from prior experience with a system or trust developed during the current interaction (e.g., dynamic learned trust).

Overall, the existing literature on the key factors influencing trust in technology, automation, and trust in AI, identifies a wide range of individual, contextual, and AI (or

automation)-specific factors. When it comes to AI (or automation)-specific factors, literature examining the features of the automation that influence trust that spans from the early nineties until recently, consistently identifies *performance*, *purpose*, and *process* as key determinants (Chiou & Lee, 2023; Hoff & Bashir, 2015; Lee & Moray, 1994; Lee & See, 2004; Perkins et al., 2010; Schaefer et al., 2016). This motivated the second research question (RQ2) of this thesis, which explores how people evaluate AI, specifically disembodied AI in the form of AI models, based on three key trust determinants of trust: performance, process, and purpose, collectively referred to as the '3Ps'.

Trust in AI - generated Advice

A significant part of the existing literature on trust in AI focuses on trust in decision making environments where AI provides advice (Burton et al., 2020; Castelo et al., 2019; Dietvorst et al., 2015; Jussupow et al., 2020; Leib et al., 2024; Logg et al., 2019). This body of literature highlights two key insights. First, AI advice uptake is highly context dependent. For instance, the perceived objectivity of a decision (Castelo et al., 2019), the level of human control on the advice output (Dietvorst et al., 2018) or the extent to which individuals consider themselves experts (Logg et al., 2019), are some of the things that have been shown to influence AI advice uptake. Second, literature on AI advice uptake points to the need for more behavioural studies, as most findings to date rely on self-report measurements of willingness to rely to AI advice (Burton et al., 2020). It also underscores the need for field studies in specific decision contexts (medical, legal, financial) given the high contextuality of AI advice uptake and its prevalence in real-life decision-making settings.

One aspect that has been overlooked in this literature is the validating effect of AI. Specifically, how people react to AI-generated advice that reinforces their decisions, especially

those related to their preferences. This kind of validating advice is very common in online spaces, where AI-driven recommendations are tailored based on a user's previous choices. Similarly, social media platforms are widely recognised for reinforcing user preferences by filtering content based on browsing history, often contributing to polarisation (Evans & Kasirzadeh, 2021). The ability of AI to fulfil social functions during decision-making, such as provide validation, motivated the third research question (RQ3) addressed in this thesis which is explored in the studies under Chapter 4.

Challenges in the Study of How Humans Perceive AI

Assessing individuals' perceptions and behaviours, whether in controlled environments or real-world contexts, has always been a challenging task, as every measurement approach has inherent limitations. In the study of how people perceive and trust AI, surveys have been used extensively, especially in the form of questionnaires that assess attitudes and willingness to trust AI in real or hypothetical scenarios. Beyond academic research, large-scale longitudinal surveys also capture public perception of AI through self-report attitudes and opinion surveys, like for example as The Nationally Representative Survey of Public Attitudes to Artificial Intelligence[1] in the UK and Stanford University's One Hundred Year Study on Artificial Intelligence (AI100) [2] in the US.

In some cases, instead of using self-report questionnaires, researchers employ behavioural measures (Castelo et al., 2019; Dietvorst et al., 2018; Kulms & Kopp, 2019; Leib et al., 2024; Logg et al., 2019). During these experiments, participants are required to complete specific tasks, while researchers measure and analyse data from their behaviour throughout the

[1] https://www.adalovelaceinstitute.org/our-work/library/

[2] https://ai100.stanford.edu/

tasks. Using solely behavioural measures is not ideal either, as internal states do not always manifest in behaviours (Barrett et al., 2019). In addition, especially when behavioural measures are taken from people's behaviour in lab setting, depending on the external validity of the findings, these findings do not always translate into real behaviours (Maner, 2016). To account for the limitation of self-report and behavioural measures, researchers rely on a combination of self-report questionnaires and behavioural tasks in the same experiment. It is not rare for these studies, however, to report a mismatch between subjective (e.g., self-report) and behavioural measures. For instance, in a study where the effect of anthropomorphism of trust in an AI agent was studied with both subjective and behavioural measures, anthropomorphism did not affect people's behavioural trust, however, anthropomorphism increased self-reported trust in the AI (Kulms & Kopp, 2019).

Measuring physiological responses to AI has also been used in the study of decision making with AI (Oertel et al., 2020; Subramanian et al., 2016) . For instance, Choi et al. (2012) measured heart rate and electrodermal activity during decision making in a prisoner dilemma with an AI interaction partner. Physiological measurements provide the benefit of overcoming some of the limitations associated with self-reports and behavioural measurements. However, since they correspond to cognitive processes, they are proxy measures of the underlying cognitive mechanisms. For instance, pupillary dilation is commonly used as a proxy for attention (Duchowski et al., 2018) or arousal (Williams et al., 2019). Neuroscientific measures help overcome some of the above limitations as they look at the underlying neural mechanisms. Although more costly to conduct than surveys or behavioural studies, they give insights into the brain regions associated with the processing of AI. (Greulich & Brendel, 2022; Tolgay et al., 2019).

Recognising the challenge stemming from the limitations of one type of measurement employed and integrating multiple types of measurement to complement and contrast measurements of the same construct (e.g., by gathering behavioural and neural data) is a methodological approach that many studies already adopt, especially in the study of perception and behaviour towards robots (Blut et al., 2021; Thellman et al., 2022) . This approach will be crucial for future research in other areas of the psychological study of AI too. For example, combining behavioural and self-report data or behavioural and neural data is an approach that could also benefit the study of human behaviour in interaction with algorithmic AI which is currently mostly based on self-report measures and to a lesser degree behavioural measures (Glikson & Woolley, 2020).

Current Work

This thesis explores how people perceive and make decisions with AI. Since most of the empirical data was collected in 2022 -2023, it provides a snapshot of what is coined ‘the human perspective on AI’ at that time. It also makes an argument for the need to continuously update our understanding of these perceptions and behaviours as both the underlying technology and people's experience with it evolve. Below is an outline of the research questions, along with the corresponding chapter in which each is addressed.

*RQ1 (*addressed in Chapter 2): What differentiates AI agents in human perception? Can human perception of AI be mapped across a wide range of AI agents, accounting for variations in design, embodiment, and contextual differences?

*RQ2 (*addressed in Chapter 3): How do people evaluate AI models based on the three determinants of trust in automation (e.g., performance, process, and purpose)?

*RQ3 (*addressed in Chapter 4): How do people respond to AI advice that validates them? Are they more, less, or equally likely to listen to AI validating them as to other people?

Chapter 2 Mapping AI Perception

Introduction

In the words of Stephen William Hawking, '*The rise of AI will be either the best or the worst thing ever to happen to humanity. We do not know which*.' It is quite likely that some of today's stereotypes surrounding Artificial Intelligence (AI) resonate with feelings of uncertainty or even apprehension, particularly as science fiction and popular culture have extensively depicted AI overtaking humanity. To map human perception of AI, we used two well-validated theoretical models from social psychology: i) the Stereotype Content Model (SCM), which offers a theoretical account of how people perceive other people (Fiske et al., 2002) and ii) the Mind Perception Dimensions (MPD) model, which maps how people attribute mind to entities (Gray et al., 2007).

Building on classic person perception (Asch, 1946; Rosenberg et al., 1968), the SCM identifies two dimensions that play a fundamental role in how people perceive other people: warmth and competence (Fiske et al., 2007; Fiske et al., 2002). Warmth pertains to the appraisal of another person's intentions, ranging from benevolent/helpful to malevolent/harmful. Competence, on the other hand, pertains to the appraisal of their ability to enact these intentions. Based on these two dimensions, the SCM identifies four stereotype categories. High competence/High warmth refers to people who are seen as highly competent and warm, such as ingroups or allies. High competence/Low warmth includes individuals who are perceived as competent but not particularly warm, such as rich people or businesspeople. Low competence/High warmth describes people who are seen as warm but lacking in competence, like the elderly or disabled. Finally, Low competence/Low warmth applies to individuals who

are elicit low appraisals of both competence and warmth, often being perceived as having minimal societal value, such as homeless people, welfare recipients, and drug addicts.

Drawing from research in person perception, studies have investigated whether appraisals of warmth and competence extend to AI agents, finding that these traits are indeed attributed to AI agents as well (Di Dio et al., 2023; Kim & Im, 2023; Kulms & Kopp, 2018; Lee & Harris, 2014; Pozharliev et al., 2023; Scheunemann et al., 2020; Xue et al., 2023). For instance, people characterise AI targets, such as computers, in terms of competence and warmth (Lee & Harris, 2014), and they select digital avatar characters based on the perceived warmth of the avatar and how well it aligns with their personal need for warmth (Fong et al., 2023). Similarly, people perceived a smart voice assistant (SVA) as both competent and warm, with perceived warmth influenced by factors such as personalisation, responsiveness, humanness, and effective communication (Xue et al., 2023). People also assess the moral behaviour of AI based on the perception of low warmth, and as such often see AI as more inclined toward utilitarian decisions in scenarios like the trolley dilemma (Zhang et al., 2022).And a comparative evaluation between autonomous vehicles (AVs) and human drivers revealed that AVs were perceived as competent albeit less competent than human drivers. This perception of competence, in turn, led to AVs being ascribed less blame in instances of negative service outcomes (Pozharliev et al., 2023).

On the other hand, based on the MPD theoretical model, attributing mind consists of two dimensions: the capacity of agency (covering one or several of the following capacities: self-control, morality, memory, emotion recognition, planning, communication, and thought) and the capacity for experience (covering one or several of the following capacities: hunger, fear, pain, pleasure, rage, desire, personality, consciousness, pride, embarrassment, and joy). Mind perception was shown to map onto the two-dimensional space of agency and experience, with

different entities, both human and non-human, occupying distinct areas. For instance, a baby and an animal were perceived as having low agentic but high experiential minds, god was attributed low experiential but high agentic mind, and a robot was (the social robot Kismet) was perceived as possessing moderate agentic but low experiential mind (Gray et al., 2007).

Building on the role of agency and experience attributions in mind perception, studies investigating the factors influencing anthropomorphism of AI (i.e., the tendency to attribute humanlike qualities to AI agents such as a human like mind) provide evidence of varying attributions of agency and experience to AI agents (Gray & Wegner, 2012; Laban et al., 2021; Tanibe et al., 2017; Van Der Woerdt & Haselager, 2019; Ward et al., 2013). For instance, a humanlike embodied AI (e.g., a humanlike robot) was attributed more experience than a non-humanlike embodied AI (e.g., a mechanical robot). However, both were rated similarly in terms of agency (Gray & Wegner, 2012). In another study examining self-disclosure for psychological health to an embodied AI (a humanoid robot), a disembodied AI (Google Nest Mini), and a human, while participants reported no difference in perceived agency between the humanoid and disembodied AI, they attributed higher levels of experience to the humanoid robot (Laban et al., 2021). And when a robot's failure seemed to result from a lack of effort rather than a lack of ability, human observers attributed significantly more agency to the robot (Van Der Woerdt & Haselager, 2019).

Leveraging the evidence in the literature that people perceive AI agents in terms of competence, warmth, and varying levels of agency and experience, the current set of studies assesses the effectiveness of the SCM and MPD theoretical models in mapping AI perception. This analysis was conducted across a spectrum of AI agents, including AI agents with varying design features, intended uses or purposes and embodiment (including both embodied AI; with

physical form such as robots and disembodied AIs; without physical form such as computer algorithms). To the best of our knowledge, no previous attempts have been made to map AI perception across a wide range of AI agents using the aforementioned models.

Overview of studies

A preliminary, comprehensive list of AI agents was compiled, encompassing both commercially available AI agents and those used in research, with the latter sourced from academic articles. As the goal was to map perceptions of real-world AI - those that are either publicly available or in an experimental stage but recognisable to the public - 'fictional movie/literature AIs,' such as the iconic droids R2D2 and C-3PO, were excluded. Beyond this exclusion, the list covered a wide range of AI agents, including pet robots, androids, self-driving cars, drones, and avatars in online forums, as well as AI used in medical decision-making, news filtering on social media, recommendation systems like Spotify and Netflix, and voice-activated assistants such as Alexa and Siri (for the full list, see Appendix 1, Table 1). After curating this list, we distributed it to professionals in AI and data science, as well as master's students in the Department of Engineering, for their review and feedback. Study 1 was then conducted to identify the most recognisable AI agents of the list e.g., the AI agents that most people were more familiar with. This allowed us to reduce the original list of 67 AI agents to 23, eliminating the potential confounding factor of familiarity while still preserving a diverse set of AI agents. Following Study 1, we proceeded with the main study, where a different group of participants evaluated the 23 most popular AIs across the dimensions of warmth, competence, agency, and experience. Cluster analyses were applied to identify groups within each of the two-dimensional spaces defined by the SCM and MPD theoretical models, respectively.

Study 1

Participants

50 UK participants (N= 50, age M= 37.6 years; 78% female) were recruited from Prolific subject pool and received standard compensation. Participants were paid based on an hourly rate (£7.50/hour) for the time spent in the study. Since the purpose of the study was to generate information rather than test a hypothesis using inferential statistics, a rule of thumb (Baumol & Quandt, 1964) was used for deciding on the sample size of 50 participants. The exact materials and data for Study 1 are available in the Open Science Framework at https://osf.io/ex8us.

Materials and Procedure

Participants were tasked with completing an online questionnaire that sought their familiarity ratings for 67 distinct AI targets. To compile this comprehensive list, we initially curated various types of AI agents, encompassing a broad spectrum of technologies. The list consisted of i) embodied AIs, incorporating entities such as social robots, domestic robots, robots utilised in manufacturing, autonomous cleaning robots for household tasks, humanoids, androids, and pet robots (totaling 27), ii) AIs presented in digital form featuring human-like characteristics, including facial features, human-shaped structures, voice characteristics, or language similarities. Examples include avatars, chatbots, non-player characters (NPCs) in video games, and voice-command-responsive personal assistants like Alexa or Siri (totaling 6), iii) AI manifested purely in algorithmic form, covering an array of functions such as movie recommendations, music composition, painting creation, poetry writing, essay writing and stock market pricing setting (totaling 34). Furthermore, the list included commercially available AI such as drones, sex robots as well as non-commercially available AI with extensive media attention including self-driving cars.

The study's objective was to identify the AI targets participants were most familiar with, aiming to mitigate the influence of familiarity that could act as a potential confounding factor in the main study, and to ensure participants were familiar with the AI they were rating. Participants were presented with descriptions of 67 different AI targets appearing on their screens sequentially in random order. They were then asked to indicate their level of familiarity after each target (‘*How familiar are you with this type of AI*?’) using a continuous scale ranging from 1 (Not at all familiar) to 7 (Extremely familiar). This approach allowed for assessing familiarity across the 67 AI targets. Three supplementary questions were also included to validate participants' awareness of these AIs and cross-check their familiarity ratings. These questions were: i)*'Are you aware of AI like the above?*' (Answer: Yes/No), ii) *'Have you ever encountered or used an AI like the above?*' (Answer: Yes/No) and iii) *'Name as many AIs as you know that fit under this type of AI. (If you don't know of any, write 'N/A').*' (Answer: Open- ended text).

Finally, throughout the study, we deliberately refrained from delineating various types of AIs solely through commercially available examples. For instance, rather than inquiring about familiarity with a specific AI agent like ChatGPT or Alexa, we phrased the question in a broader context. For example, the question ‘*How familiar are you with AI that generates substantial passages of text in various styles when given a few initial words or lines?*’ was used to describe AI models like ChatGPT. Similarly, the description ‘*AI that acts as a personal assistant, taking voice commands (e.g., searching the web, ordering products online, triggering events, or playing movies and music on request)*’ was used for AI systems like Alexa. This approach aimed to prevent biasing participants' recognition to only the most prevalent or commercially accessible AIs. This method also enabled us to verify participants’ comprehension by requesting them to

list as many AIs as they knew that fell within the described type of AI rather than us referring to only a few examples.

Results

The results of the Study 1 are included in the Appendix 1 (Table 1) where the 67 AI targets appear in descending order based on their average familiarity rating. Following the Study 1, we curated a new list comprising the 23 most familiar AI targets, e.g., the ones that participants in Study 1 gave higher familiarity ratings. This selection included AI agents with average familiarity ratings falling within the range of [2.50, 5.54], as evaluated on a continuous scale from 1 to 7, with percentile values q1= 2.94, q2=3.8, q3=4.57, and q4=5.54.

Participants consistently answered affirmatively to both the first and second complementary questions for the 23 AI agents that demonstrated high familiarity scores [2.50, 5.54], thereby validating the high familiarity attributed to each of these AIs. Furthermore, for these 23 AI agents, respondents commonly supplied multiple examples in response to the third complementary question, something that further allowed cross verifying familiarity ratings. The 23 most popular AI agents were then used in the Study 2 (the 23 most popular AI agents used in Study 2 are included in Appendix1, Table 2).

Study 2

A new sample of participants evaluated the 23 AI targets using the SCM's competence and warmth scales, along with MPD's agency and experience scales. This evaluation sought to gauge how this new sample of participants perceived the 23 AI targets in terms of competence, warmth, agency, and experience related traits.

Participants

We did not conduct a power analysis to determine the appropriate sample size since we did not run inferential statistics. Instead, we relied on prior research to inform our decision regarding the sample size for the main study. Following the approach of Sevillano and Fiske (2016), who examined the SCM model in animals, we chose to recruit 138 participants. This represents an increase in sample size compared to the initial study by Fiske et al. in 2002 (N=124 in Study 1, long survey). Our rationale for this decision was to adopt a sample at least as large as the largest sample size among the two aforementioned SCM studies in the existing literature. As such, UK participants (N= 138, age M= 37.94 years, 63% female, 36% male and 1 participant preferred not to say) were recruited through the Prolific experimental subject pool for compensation (£9.00 per hour). From the total number of participants recruited (N=137), 12 were excluded for finishing the questionnaire in less than 10 mins, whereas 3 participants were timed out. That left us with a total sample size of N=123. Ethnicities were: 90% White, 2% Black, 2% Asian, 2% Mixed and 1% Unspecified. The exact materials and data for Study 1 are available in the Open Science Framework at https://osf.io/ex8us.

Materials and Procedure

In this study, the questionnaire featured the 23 AI targets obtained from the Study 1. Participants were tasked with rating these AIs using randomly presented questions related to SCM traits (Fiske et al., 2002) and questions linked to mind perception - traits (Gray et al., 2007).

More specifically, the adjectives used in the questions asking about perceived *warmth* were warm, well-intentioned, friendly, trustworthy, good-natured, helpful. For *competence*: competent, intelligent, efficient, ingenious, skilful, knowledgeable. For *agency*: capacity to exercise self-control, planning, thinking, communicating with humans, having moral

character, capable of remembering and recognising human emotions and for *experience*: ability to feel basic psychological states such as hunger, thirst, joy, fear, and pain, being self-aware of things and having the ability to self-reflect and ability to experience emotional states. In all the questions, participants were asked to rate each one of the 23 AI targets on a scale from 0 (not at all) to 100 (extremely) according to how most people view them, similar to previous application of SCM on person and animal perception (Fiske et al., 2002; Sevillano and Fiske (2016)). A 'Does not apply' option was also provided to accommodate situations where participants held the opinion that these traits might not be suitable to describe AI targets.

Furthermore, participants were asked to assess the perceived human likeness of the evaluated AI targets ('*How much humanlike do you think most people find this AI?'*) on a continuous scale from 0 (Not at all humanlike) to 100 (Extremely humanlike). The study protocol involved providing written instructions at the outset and debriefing participants upon completion of the study. To prevent participant fatigue and similar to previous SCM studies (Fiske et al., 2018; Fiske et al., 2002), the sample was divided, with participants randomly being allocated to one of two groups, rating half of the 23 AIs (11 and 12, respectively).

Results

Every participant had at least one instance where they selected a 'Does not apply' answer for one of the adjectives under each dimension (Table 1). No participant selected 'Does not apply' for all the adjectives under the same dimension, indicating that the dimensions were applicable for describing AI, although different adjectives were seen as not applicable for describing AI by different participants (e.g., not all participants considered the same adjectives as not applicable). Participants most frequently chose 'Does not apply' for adjectives under the

experience dimension, with 57% of the overall ratings falling into this classification. 'Does not apply' ratings were not included in our data analyses.

The high reliability scores indicated by Cronbach's alphas for warmth (α = .86, across 6 relevant questions) and competence (α = .91, across 6 relevant questions), as well as for agency (α = .84, across 7 relevant questions) and experience (α = .80, across 5 relevant questions) suggest high levels of consistency in the ratings across relevant questions (see Table 1 for the list of questions used per dimension). We thus aggregated over the questions to compute the dimension composite scores by computing the mean ratings for competence, warmth, agency, and experience based on participant responses to each of the 23 AI targets. Using these means, the 23 AI targets were then plotted on a two-dimensional competence x warmth (SCM) space (see Figure 2) and on a two-dimensional agency x experience (MPD) space (see Figure 3).

Table 1

Scales used. For the Competence and Warmth, Agency and Experience Scales, the points of ellipsis were replaced by the words in brackets for each question.

Construct	Items
Competence	In terms of how most people view this AI: - How … do most people find this AI? [competent, intelligent, efficient, ingenious, knowledgeable, skilful]
Warmth	In terms of how most people view this AI: - How … do most people find this AI? [warm, well-intentioned, good-natured, trustworthy, friendly, helpful]
Agency	In terms of how most people view this AI: - How … do most people find this AI? [capable of exercising self-control, capable of planning, capable of thinking, capable of communicating, capable of moral character, capable of remembering, capable of recognising human emotions]
Experience	In terms of how most people view this AI: - How … do most people find this AI? [has a personality, capable of being aware of things, ability to self-reflect, ability to experience emotional states, capable of feeling hungry/joy/fear/pain/desire]

Two separate cluster analyses were conducted to explore the structure of each one of the two-dimensional spaces (SCM, MPD). Specifically, following the approach outlined by Hair et al. (1998), we conducted hierarchical cluster analyses (Ward's method, which minimising within-cluster variance) (Ward Jr, 1963) to identify the optimal number of clusters. Subsequently, we employed agglomeration statistics using typical decision rules (per Blashfield and Aldenderfer (1988)) to determine where the last large change occurred. In both SCM and MPD, this change occurred at the transition between clusters 2 and 3, leading to the adoption of a three-cluster solution for both (Figure 2 and Figure 3 for SCM and MPD respectively).Next, we conducted k-means cluster analyses, utilising the parallel threshold method, to further delineated which type of AI fell into each cluster. Notably, the different AI targets formed cohesive clusters that remained consistent across less informative two or four-solution clustering attempts for both frameworks (Table 2 and Table 3 for SCM and MPD respectively).

Table 2

SCM Group Clusters in Four-, Three- and Two-Cluster Solutions.

	Clusters		
AI targets	**4-solution**	**3- solution**	**2- solution**
Self-driving cars	1	1	1
AI that recommends products or services to buy	1	1	1
AI that calculates credit scores for granting credit cards, loan or mortgages	1	1	1
AI that categorises emails in your inbox and offers quick reply options	2	2	2
Facial recognition AI used to identify potential suspects and conducts mass surveillance	1	1	1
AI that filters and organises the content and the news feeds of social media sites	1	1	1
Drones	1	1	1
AI that connects you to potential friends/people you might know of	2	2	2
AI that recommends movies, shows or series	2	2	2
Robots used in manufacturing (e.g. digitally operated robotic arms)	1	1	1
AI that recommends people to go on a date with	3	3	1
Chatbots used in customer service to answer questions & provide information	3	3	1
Non-player characters (NPC) in video games	3	3	1
Avatars used in Internet forums, social media and other online communities	3	3	1
Avatars used in video games to represent different players	2	2	2
Domestic robots	2	2	2
AI that recommends music to listen to	2	2	2
Facial recognition AI used to open digital devices	2	2	2
AI that acts as a typing assistant that reviews spelling, grammar, clarity, and corrects mistakes	4	2	2
AI that matches people searching for a ride with potential drivers and also offers ridesharing services	2	2	2
Autonomous robots that do vacuum cleaning in houses	2	2	2
AI that acts as a personal assistant that takes voice commands	4	2	2
AI that provides navigation services (e.g. how to go from point A to B, suggesting alternative itineraries etc.)	4	2	2

Table 3

MPD Group Clusters in Four-, Three- and Two- Cluster Solutions.

	Clusters		
AI targets	**4-solution**	**3- solution**	**2- solution**
AI that calculates credit scores for granting credit cards, loans, and mortgages	1	1	1
AI that categorises emails in your inbox and offers quick reply options	1	1	1
AI that acts as a typing assistant that reviews spelling, grammar, clarity, and corrects mistakes	1	1	1
AI that filters and organises the content and the news feeds of social media sites	1	1	1
AI that recommends movies, shows or series	1	1	1
Chatbots used in customer service to answer questions & provide information	1	1	1
AI that recommends products or services to buy	1	1	1
Facial recognition AI used to identify potential suspects and conducts mass surveillance	1	1	1
AI that recommends people to go on a date with	1	1	1
AI that matches people searching for a ride with potential drivers and also offers ridesharing services	1	1	1
AI that connects you to potential friends/people you might know of	1	1	1
Facial recognition AI used to open digital devices	1	1	1
Avatars used in Internet forums, social media and other online communities	2	2	1
Autonomous robots that do vacuum cleaning in houses	2	2	1
Drones	2	2	1
Robots used in manufacturing (e.g. digitally operated robotic arms)	2	2	1
AI that recommends music to listen to	3	3	2
Self-driving cars	3	3	2
	3	3	2
AI that provides navigation services (e.g. how to go from point A to B, suggesting alternative itineraries etc.)			
AI that acts as a personal assistant	3	3	2
Domestic robots	4	3	2
Non-player characters (NPC) in video games	4	3	2
Avatars used in video games to represent different players	4	3	2

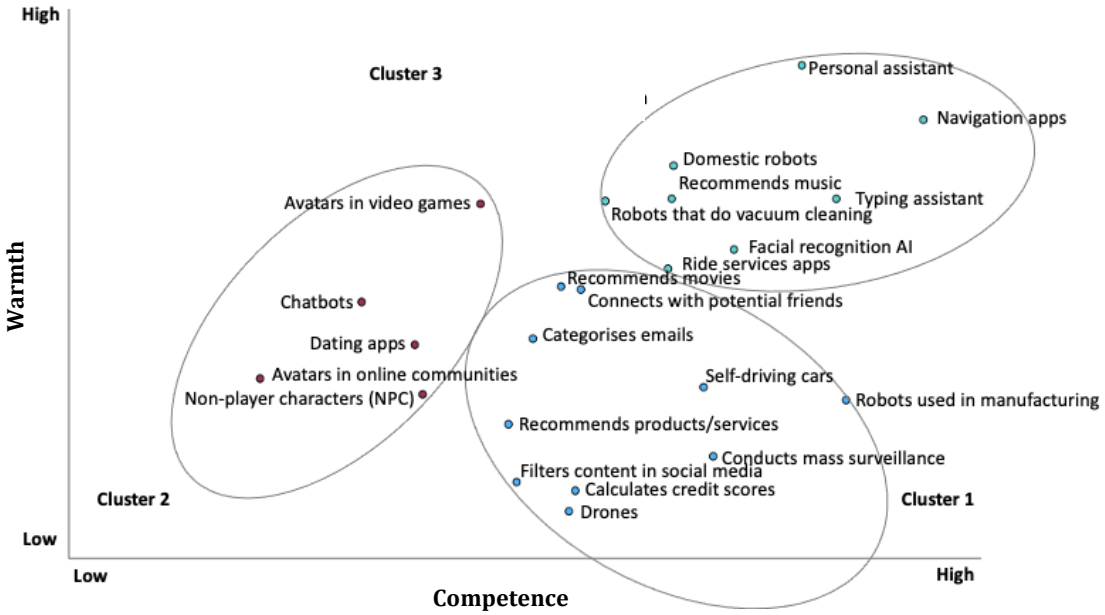
Mapping AI Perception using the SCM

SCM characterises mixed stereotypes towards social groups as displaying low ratings on one dimension and high ratings on the other. Three distinct analyses were conducted to examine whether there were any clusters of the AI targets under examination demonstrating high scores on either competence or warmth while displaying low scores on the other dimension. Firstly, an Analysis of Variance (ANOVA) and post-hoc tests were conducted to compare the means across the three identified clusters. This analysis aimed to discern whether significant differences existed in the mean scores across these clusters. Secondly, within each cluster, matched pair t-tests were conducted to directly compare competence and warmth scores. This approach allowed to investigate potential significant differences between these dimensions within the individual clusters. Thirdly, matched-pair t-tests were conducted at the level of individual AI targets within clusters. These tests provided a detailed examination of the competence and warmth scores for each AI target within the cluster, enabling to ascertain whether specific AIs exhibited notable differences between their mean competence and warmth scores.

First, a comparison of the three cluster-means confirmed that clusters differed on each dimension (competence $F_{2,20}$ = 17.851, p< .001, warmth $F_{2, 20}$=19.571, p<.001). Subsequent post-hoc analyses revealed that Cluster 3 exhibited the highest scores in competence (M = 65.38) relative to the other two clusters, with statistically significant differences in terms of competence (p< .05, 95% C.I = [3.88, 12.26] for the difference between Cluster 3 and Cluster 1, p< .001, 95% C.I = [9.84, 24.29] for the difference between Cluster 3 and Cluster 2). Cluster 3 also exhibited the highest scores in warmth (M= 56.65) relative to the other two clusters (p<.001, 95% C.I = [8.10, 19.14] for the difference between Cluster 3 and Cluster 1, p<.05, 95% C.I = [1.98, 15.25] for the difference between Cluster 3 and Cluster 2). Cluster 1, like Cluster 2, was also a cluster with the lowest warmth scores (M= 43.03), since the warmth scores of these two clusters (Cluster 2, M= 48.03 and Cluster1, M=43.03) were not found statistically significantly different (p= .142). However, when it comes to mean competence score, Cluster 1 scored moderately higher in competence (M= 59.14) than Cluster 2 (M= 48.32, p =.002) and lower compared to Cluster 3 (M=65.38, p= .041). As such, we refer to this cluster as 'Moderate Competence – Low Warmth' cluster.

Figure 2

SCM Three-Cluster Solution



Secondly, the matched-pair t-tests conducted within each cluster revealed noteworthy findings. In Cluster 1, the mean competence (M = 59.14) was significantly higher than the mean warmth (M = 43.03), t (9) = 7.02, $p < .001$. Similarly, in Cluster 3, the mean competence (M = 65.38) was significantly higher than the mean warmth (M = 56.65), t (7) = 5.42, $p < .001$. Conversely, the matched-pair t-test indicated no statistically significant difference between the mean scores of warmth and competence in Cluster 2, t (4) = 0.16, p = .88, as shown in Table 4.

Table 4

Mean Competence and Mean Warmth for each AI cluster. Within each row, means differ (p<.05) if > or < is indicated. Within each column, means that do not share a subscript letter differ (p<0.001). Standard deviations appear in parenthesis.

Cluster	Members	Competence		Warmth
1	10	59.14 (5.28) a	>	43.03 (4.97) a
2	5	48.32 (4.03) b	=	48.03 (4.65) a
3	8	65.38 (5.16) c	>	56.65 (4.04) b

Cluster 1 contained ten members which, mentioned in order of distance from the cluster centre, were the following: '*Self-driving cars*' (cluster's centre), '*AI that recommends products or services to buy*', '*AI that calculates credit scores for granting credit cards, loans, or mortgages*', '*AI that categorises emails in the inbox and offers quick reply to options*', '*Facial recognition AI used to identify potential suspects and conduct mass surveillance*', '*AI that filters and organises the content and the news feeds of social media sites*', '*Drones*', '*AI that connects you to potential friends/ people you might know of*', '*AI that recommends movies, shows and series*', '*Robots used in manufacturing (e.g., digitally operated robotic arms).*' Cluster 1 had a statistically significantly higher mean score of competence (M= 59.14) than warmth (M= 43.03)), t (9) =7.019 p<0.001 suggesting a prevailing perception of higher competence compared to warmth among its members.

Cluster 2, comprising five members, demonstrated the lowest scores in both warmth (M=48.32) and competence (M=48.03), akin to the group in the original person perception SCM study, known as the 'antipathy group.' (Fiske et al., 2018). The members within this cluster, listed in order of their distance from the cluster centre, were as follows: '*AI that suggests potential dates for individuals*' (cluster's centre), '*Customer service chatbots that handle queries and provide information*', '*Non-player characters (NPC) in video games*', '*Avatars used across Internet forums, social media, and online communities*', '*Avatars representing players in video games*.'

Cluster 3, the cluster with the highest competence (M=65.38) and warmth (M= 56.65) contained eight members. These members, arranged by their proximity to the cluster centre from the closest to the farthest, were as follows: '*Domestic robots'* (cluster's centre), '*AI that recommends music choices*', '*Facial recognition AI for digital device access'*, '*AI that assists in*

typing, reviewing spelling, grammar, and clarity', '*AI that facilitates ride matches and offers ridesharing*', '*Autonomous vacuum cleaning robots*', '*AI that serves as a voice-commanded personal assistant*', '*AI that provides navigation services*.' Within this cluster, matched-pair t-test revealed the mean score of warmth to be significantly lower than that of competence (M=8.739, t (7) =5.42, p<.001), suggesting a prevailing perception of higher competence compared to warmth among its members.

Finally, at the level of individual groups, e.g., within clusters, matched-pair t-tests were conducted to compare the mean competence and warmth scores for each AI target. Notably, mean competence and warmth scores differed statistically significantly for 20 AI targets out of the total 23 (see Table 5). The three AI targets for which the difference between the mean competence and warmth scores was not found to be statistically significant (at p<0.05) came from Clusters 1 and 2.

Table 5

Mean Paired Differences (Competence - Warmth)

	Competence	Warmth	t	p
Facial recognition AI used to identify potential suspects and conducts mass surveillance	60.74 (22.72)	39.02 (20.74)	12.741	<.001
Drones	57.01 (21.09)	37.99 (20.04)	11.763	<.001
AI that filters and organises the content and the news feeds of social media sites	56.01 (22.89)	38.25 (19.53)	11.131	<.001
AI that provides navigation services (e.g. how to go from point A to B, suggesting alternative itineraries etc.)	69.76 (17.96)	56.61 (18.33)	10.225	<.001
Robots used in manufacturing (e.g. digitally operated robotic arms)	64.19 (19.72)	46.99 (20.73)	10.191	<.001
Self-driving cars	63.65 (21.50)	44.22 (17.88)	9.865	<.001
AI that matches people searching for a ride with potential drivers and also offers ridesharing services	63.47 (19.08)	51.90 (18.42)	9.073	<.001
AI that acts as a typing assistant that reviews spelling, grammar, clarity, and corrects	67.63 (19.70)	54.17 (19.32)	8.852	<.001
AI that recommends people to go on a date with	60.19 (22.59)	45.32 (21.46)	8.586	p=0.06
AI that categorises emails in your inbox and offers quick reply options	56.41 (22.17)	42.54 (21.15)	8.529	<.001
AI that connects you to potential friends/people you might know of	57.56 (22.03)	44.07 (21.09)	8.523	<.001
AI that calculates credit scores for granting credit cards, loans or mortgages	57.42 (23.59)	38.10 (19.97)	8.100	<.001
Non-player characters (NPC) in video games	48.71 (23.90)	46.53 (21.53)	1.356	<.001
Avatars used in Internet forums, social media and other online communities	44.31 (23.62)	46.75 (22.07)	-1.432	<.001
Avatars used in video games to represent different players	46.58 (22.85)	49.27 (21.75)	-1.530	p=0.3
Domestic robots	51.83 (23.32)	50.87 (22.07)	0.531	<.001
AI that acts as a personal assistant that takes voice commands	66.79 (18.43)	58.28 (18.78)	7.108	<.001
Facial recognition AI used to open digital devices	63.60 (19.22)	54.18 (19.67)	6.958	<.001
Chatbots used in customer service to answer questions & provide information	58.90 (23.24)	46.62 (21.11)	6.373	<.001
AI that recommends products or services to buy	54.18 (24.33)	41.95(19.91	6.065	<.001
AI that recommends movies, shows or series	57.21 (20.32)	50.27 (20.49)	5.784	<.001
Autonomous robots that do vacuum cleaning in houses	61.98 (19.60)	54.11 (20.87)	4.860	p=0.2
AI that recommends music to listen to	62.15 (18.05)	56.71 (20.02)	3.946	<.001

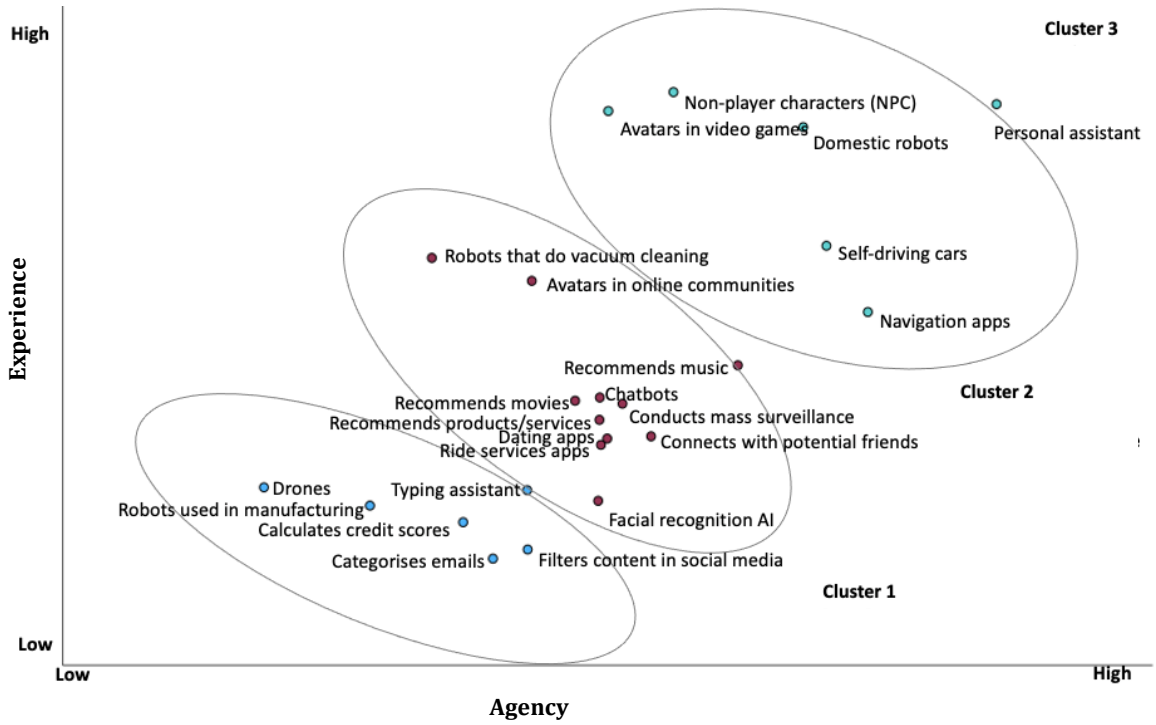
In summary, the analysis conducted on clusters (using ANOVA and matched-pair t-tests) and individual groups within clusters (employing matched-pair t-tests) indicated that a three-cluster solution was deemed the most appropriate. The analysis did not reveal any mixed stereotypes (e.g., AI targets scoring high on one dimension while low on the other), apart from the existence of Cluster 1, which included AIs that scored moderately in competence and high in warmth and, as such, it could be 'marginally' considered as a mixed cluster. The results indicate that the AI targets under review were not perceived homogeneously in terms of competence and warmth. Instead, they formed different clusters the two-dimensional space of SCM.

Mapping AI Perception using the MPD Model

Similar analyses to the ones described above when exploring the SCM's two-dimensional space defined by competence x warmth were also conducted to investigate the MPD two-dimensional space defined by agency x experience.

Figure 3

Mind Perception Dimensions Three-Cluster Solution.



First, comparison of the three cluster-means confirmed that clusters differed on each dimension (agency, $F_{2,20} = 18.45$, $p < .001$ and experience, $F_{2,20} = 40.42$, $p < .001$). Post-hoc analyses further demonstrated that Cluster 3, which exhibited the highest scores in both agency (M= 43.40) and experience (M= 30.41) compared to the other clusters, significantly differed from the other two clusters. Cluster 3 showed significantly higher agency compared to Cluster 1 ($p < .001$, 95% C.I = [7.04, 17.17]) and Cluster 2 ($p = 0.003$, 95% C.I = [2.37, 11.27]), as well as significantly higher experience compared to Cluster 1 ($p < .001$, 95% C.I = [9.30, 16.80])) and Cluster 2 ($p < .001$, 95% C.I = [5.18, 11.77])). Cluster 2 showed significantly higher agency ($p = .018$, 95% C.I = [0.84, 9.74]) and experience ($p = .006$, 95% C.I = [1.27, 7.83]) compared to Cluster 1.

Secondly, the matched-pair t-tests conducted within each cluster revealed noteworthy findings. Specifically, within Cluster 1, the mean agency (M=31.30) was significantly higher than the mean experience (M=17.36), t (5) =7.970, $p < .001$. Similarly, Cluster 2 also demonstrated a significant difference between mean agency (M=36.59) and experience (M=

21.93), t (10) = 10.692, p<.001, and in Cluster 3, the mean agency (M=43.40) was higher than the mean experience (M=30.41), t (5) =4.820, p=.005 (Table 6).

Table 6

Agency and Experience means for each AI cluster. Within each row, means differ (p<.05) if > or < is indicated. Within each column, means that do not share a subscript letter differ (p<.05). Standard deviations appear in parenthesis.

Cluster	Members	Agency		Experience
1	6	31.30 (3.56) a	>	17.36 (1.11) a
2	11	36.59 (2.54) b	>	21.93 (2.63) b
3	6	43.40 (4.74) c	>	30.41 (3.36) c

Cluster 1, characterised by the lowest scores in both agency (M= 31.30) and experience (M= 17.36), consisted of six members ranked by their distance from the cluster centre: **'***AI that calculates credit scores for credit cards, loans, or mortgages***'** (cluster's centre), '*AI that categorises emails in the inbox and suggests quick reply to options*', '*Robots utilised in manufacturing processes (e.g., digitally operated robotic arms)*', '*AI which functions as a typing assistant for spell checks, grammar reviews, and error corrections*', '*AI responsible for filtering and organising content on social media platforms*', '*Drones*.'

Cluster 2 consisted of 11 members and displayed moderate scores in both agency (M=36.59) and experience (M=21.93) compared to the other clusters. The AI targets within Cluster 2, ranked in terms of their distance from the cluster centre, included***: 'AI that recommends movies, shows, or series'*** (cluster's centre), '*Chatbots used in customer service to answer queries and provide information*', '*AI that recommends products or services for purchase*', '*Facial recognition AI used in identifying potential suspects and conducting mass surveillance*', '*AI that suggests people for potential dating'*, '*AI matching individuals searching for rides with potential drivers and offering ridesharing services'*, '*AI that connects users to potential friends or people*

they might know', '*Facial recognition AI used to unlock digital devices*', '*Avatars used in Internet forums, social media, and online communities'*, '*AI that suggests music for listening*', '*Autonomous robots performing vacuum cleaning in houses.'*

Finally, at the level of individual groups within clusters, matched-pair t-tests were conducted to compare agency and experience scores for each AI target. The analysis revealed significant differences between agency and experience scores for all 23 types of AI (Table 7). Across these 23 AI categories, all were consistently perceived as significantly more agentic than experienced.

Cluster 3, which demonstrated the highest mean agency (M= 43.40) and experience (M=30.41) scores, comprised six members, listed in ascending order of their proximity to the cluster centre: '*Domestic robots*' (cluster's centre), '*Self-driving cars*', '*Non-player characters in video games*', '*AI that provides navigation services'*, '*Avatars used in video games to represent different players*', '*AI that serves as a voice-commanded personal assistant*.'

Table 7

Mean Paired Differences (Agency - Experience)

	Agency	Experience	t	p
AI that filters and organises the content and the news feeds of social media sites	33.08 (17.63)	16.23 (17.10)	9.967	<.001
Facial recognition AI used to open digital devices	34.11 (18.66)	18.03 (16.20)	9.483	<.001
AI that connects you to potential friends/people you might know of	35.91 (19.96)	20.41 (20.71)	8.920	<.001
AI that categorises emails in your inbox and offers quick reply options	31.84 (18.50)	15.90 (17.82)	8.738	<.001
AI that provides navigation services (e.g. how to go from point A to B, suggesting alternative itineraries etc.)	44.44 (19.16)	24.10 (21.13)	8.651	<.001
AI that recommends music to listen to	39.76 (21.87)	23.03 (22.86)	8.481	<.001
Chatbots used in customer service to answer questions & provide information	35.89 (19.73)	21.84 (17.64)	8.407	<.001
Self-driving cars	43.63 (16.51)	27.44 (19.10)	7.738	<.001
AI that acts as a typing assistant that reviews spelling, grammar, clarity, and corrects mistakes	33.95 (23.49)	18.42 (20.29)	7.728	<.001
AI that acts as a personal assistant that takes voice commands	33.95 (23.49)	18.42 (20.29)	7.728	<.001
AI that recommends people to go on a date with	35.02 (19.32)	20.33 (19.56)	7.569	<.001
Domestic robots	43.49 (19.89)	31.00 (20.89)	7.538	<.001
AI that matches people searching for a ride with potential drivers and also offers ridesharing services	36.42 (19.62)	20.09 (23.07)	7.430	<.001
AI that recommends movies, shows or series	35.12 (22.08)	21.72 (23.38)	6.695	<.001
Facial recognition AI used to identify potential suspects and conducts mass surveillance	35.61 (20.21)	21.61 (18.19)	6.559	<.001
AI that recommends products or services to buy	35.81 (20.97)	21.02 (22.22)	6.165	<.001
AI that calculates credit scores for granting credit cards, loans or mortgages	31.30 (20.64)	16.43 (19.87)	5.937	<.001
Robots used in manufacturing (e.g. digitally operated robotic arms)	28.27 (21.99)	16.82 (18.53)	5.534	<.001
Non-player characters (NPC) in video games	39.23 (22.75)	33.11 (21.12)	4.393	<.001
Drones	25.13 (20.98)	18.14 (17.42)	4.284	<.001
Avatars used in Internet forums, social media and other online communities	34.40 (21.72)	26.14 (21.04)	4.204	<.001
Avatars used in video games to represent different players	35.57 (19.43)	31.62 (20.66	2.804	p=.007
Autonomous robots that do vacuum cleaning in houses	30.99 (20.01)	26.99 (18.10)	1.792	p=.078

In summary, the analyses performed on clusters, both through ANOVA and matched-pair t-tests, as well as individual groups within clusters using matched-pair t-tests, supported selecting a three-cluster solution as the most suitable. The analysis did not reveal any mixed stereotypes (e.g., AI targets scoring high on one dimension while low on the other). Notably, in the case of the MPD theoretical model, the solution aligned along a diagonal within our sample, which essentially shows that agency and experience are correlated. This indicates that, according to the MPD model, AI perception is unidirectional: if people perceive an agent as high (or low) in agency, it is also perceived as high (or low) in experience, and vice versa.

Competence and Experience: The Two Predictors of Human likeness

A linear regression was run on the human likeness output variable. Upon confirming that the assumptions for linear regression were met, the analysis was run, and a significant model was obtained: $F_{4,18} = 21.052$, $p<.001$. Notably, after examining the Coefficients (see Table 8), it was observed that competence negatively and experience positively predicted human likeness.

Table 8

Coefficients of Regression Model with Dependent variable: Human likeness

	Unstandardised Coefficients beta	Coefficient Standard Error	t	p
(Constant)	-13.639	11.286	-1.208	.243
Competence	-.489	.189	-2.595	.018
Warmth	.357	.233	1.543	.140
Agency	.526	.327	1.607	.125
Experience	1.355	.318	4.258	<.001

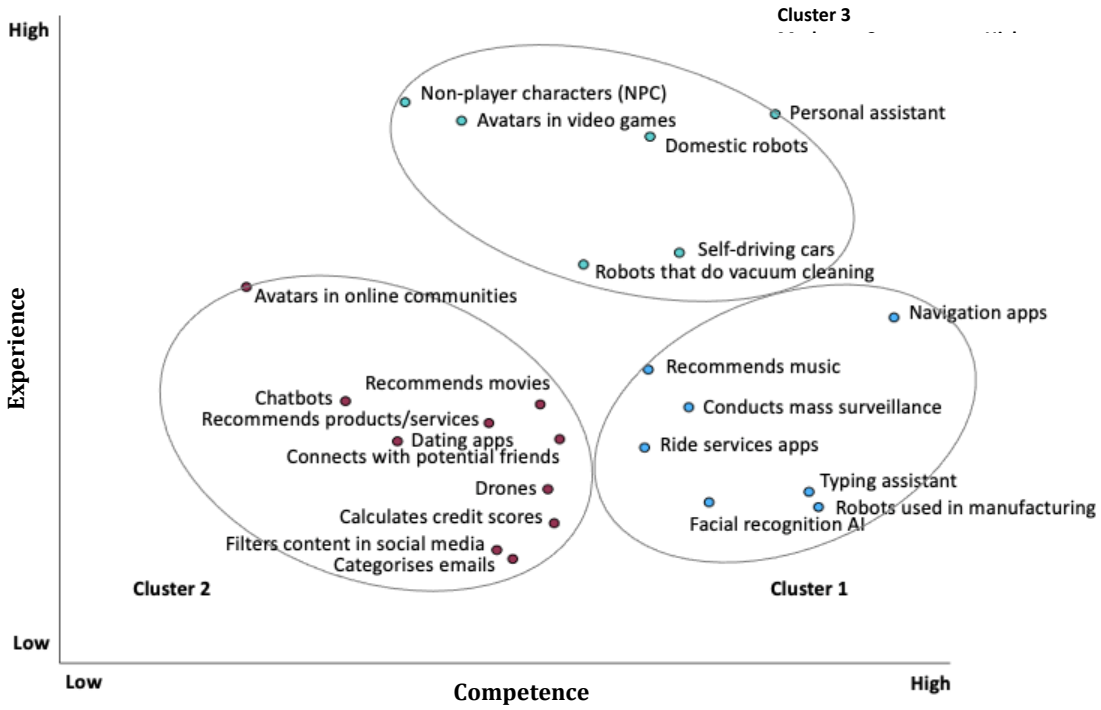
Building on the finding that competence and experience significantly predicted the perceived human likeness of AI, we examined the two-dimensional space defined by these dimensions. We refer to this model as the AI Stereotype Model (AISM). Determination of the optimal number of clusters was established through agglomeration statistics. Notably, the most

substantial change occurred between clusters 2 and 3, leading to the adoption of a three-cluster solution as shown in Figure 4.

Next, to explore whether different types of AIs will prove high on either competence or experience but low on the other, three analyses were performed. These three analyses were similar to the ones performed for investigating both SCM and MPD two-dimensional spaces.

Figure 4

AI Stereotype Model Three-Cluster Solution.



First, there were differences in the means of the three clusters across each dimension (competence: $F_{2,20} = 12.43$, $p < .001$, experience: $F_{2,\ 20} = 29.106$, $p<.001$). Subsequent post-hoc analyses revealed that the cluster exhibiting the highest mean competence compared to the others, Cluster 1 (M= 66.90), significantly differ from Cluster 2 ($p<.001$, 95% C.I = [6.75, 20.67]) but moderately from Cluster 3 with no statistical significance ($p=.060$, 95% C.I = [-2.93, 15.43]) whereas mean competence scores of Cluster 2 and 3 didn't differ statistically

significantly (p=.109, 95% C.I = [-13.44, 1.15]). Regarding experience, the post-hoc analyses demonstrated that the cluster with the highest experience relative to the other clusters, Cluster 3 (M= 30.74), statistically significantly differed from both Cluster 1 (p< .001, 95% C.I = [6.06, 14.27]) and Cluster 2 (p<.001, 95% C.I = [6.99,14.61]). Mean experience scores Cluster 1 and 2 didn't differ statistically significantly (p=.896, 95% C.I = [-2.99, 4.28]).

Secondly, matched-pair t- tests revealed that within Cluster 1, competence (M=66.90) was significantly higher than experience (M=20.58), t (6) =23.26, p< .001. Similarly, within Cluster 2, competence (M= 53.18) was significantly higher than experience (M=19.94), t (9) = 13.58, p< .001, as well as within Cluster 3, competence (M=59.33) was significantly higher than experience (M= 30.74), t (5) = 8.66, p <.001. (see Table 9).

Table 9

Mean Competence and Mean Experience means for each AI cluster. Within each row, means differ (p<.05) if > or < is indicated. Within each column, means that do not share a subscript letter differ (p<.05). Standard deviations appear in parenthesis.

Cluster	Members	Competence		Experience
1	7	66.90 (4.79) a	>	20.58 (2.75) a
2	10	53.18 (5.24) b	>	19.94 (3.10) a
3	6	59.33 (6.92) a, b	>	30.74 (2.77) b

Cluster 1 is a mixed cluster, boasting high competence score (M= 66.90) and low experience score (M= 20.58) and comprising seven members which are the following, ordered by their proximity to the cluster centre: '*Facial recognition AI used to identify potential suspects and conduct mass surveillance*'(cluster's centre), '*Facial recognition AI used to open digital devices*', '*AI that acts as a typing assistant that reviews spelling, grammar, clarity, and corrects mistakes'*, '*AI that matches people searching for a ride with potential drivers and also offers ridesharing services'*, '*Robots used in manufacturing (e.g., digitally operated robotic arms)*', '*AI*

that recommends music to listen to', 'AI that provides navigation services (e.g., how to go from point A to B, suggesting alternative itineraries etc.).'

Cluster 2, encompassed ten members arranged in order of distance from the cluster centre: '*AI that recommends products or services to buy*' (cluster's centre), '*AI that recommends people to go on a date with*', '*AI that recommends movies, shows or series*', '*Drones'*, '*AI that connects you to potential friends/people you might know of'*, '*AI that filters and organises the content and the news feeds of social media sites*', '*AI that calculates credit scores for granting credit cards, loans or mortgages', 'AI that categorises emails in your inbox and offers quick reply options*', '*Chatbots used in customer service to answer questions & provide information*', '*Avatars used in Internet forums, social media and other online communities.'* This cluster includes AI targets that scored both low in competence (M=53.18) and experience (M= 19.94).

Cluster 3, comprised six members arranged in order of proximity to the cluster centre, ranging from the closest to the farthest distance: '*Domestic robots'* (cluster's centre), '*Autonomous robots that do vacuum cleaning in houses*', '*Self-driving cars'*, '*Avatars used in video games to represent different players*', '*AI that serves as a voice-commanded personal assistant', 'Non-player characters (NPC) in video games.'* This cluster is another mixed cluster that includes AI targets that scored moderately in competence (M=59.33) while at the same time they exhibited high experience score (M= 30.74).

Finally, at the level of individual groups, e.g., within clusters, matched – pair t-tests compared competence and experience scores for each type of AI. Competence and experience scores differed significantly for all 23 types of AI. All these 23 types of AI were perceived to be significantly more competent than experienced.

Table 10

Mean Paired Differences (Agency - Experience)

	Competence	Experience	t	p
Robots used in manufacturing (e.g. digitally operated robotic arms)	70.55 (16.93)	17.85 (20.07)	20.782	p<.001
AI that acts as a typing assistant that reviews spelling, grammar, clarity, and corrects mistakes	70.83 (19.89)	18.42 (20.29)	17.141	p<.001
AI that acts as a personal assistant that takes voice commands	70.83 (19.89)	18.42 (20.29)	17.141	p<.001
AI that provides navigation services (e.g. how to go from point A to B, suggesting alternative itineraries etc.)	73.97 (14.96)	24.10 (21.13)	16.620	p<.001
AI that recommends music to listen to	62.50 (19.44)	23.03 (22.86)	14.611	p<.001
AI that matches people searching for a ride with potential drivers and also offers ridesharing services	61.93 (18.78)	20.09 (23.07)	14.584	p<.001
Facial recognition AI used to open digital devices	63.90 (19.65)	18.03 (16.20)	14.523	p<.001
Drones	56.19 (18.01)	18.52 (18.07)	14.219	p<.001
AI that filters and organises the content and the news feeds of social media sites	54.80 (22.45)	16.23 (17.10)	12.730	p<.001
AI that recommends movies, shows or series	56.79 (20.44)	21.72 (23.38)	12.573	p<.001
AI that categorises emails in your inbox and offers quick reply options	54.17 (20.77)	15.90 (17.82)	12.123	p<.001
Facial recognition AI used to identify potential suspects and conducts mass surveillance	63.87 (21.60)	21.61 (18.19)	11.595	p<.001
AI that connects you to potential friends/people you might know of	57.57 (20.80)	20.41 (20.71)	11.317	p<.001
Self-driving cars	63.21 (21.55)	27.44 (19.10)	11.181	p<.001
Autonomous robots that do vacuum cleaning in houses	58.89 (19.29)	26.10 (18.10)	11.123	p<.001
AI that calculates credit scores for granting credit cards, loans or mortgages	57.37 (23.82)	17.24 (20.65)	10.922	p<.001
Chatbots used in customer service to answer questions & provide information	47.70 (22.47)	21.84 (17.64)	10.893	p<.001
Domestic robots	62.55 (17.45)	31.82 (21.64)	10.535	p<.001
AI that recommends products or services to buy	54.19 (24.95)	21.02 (22.22)	9.778	p<.001
AI that recommends people to go on a date with	49.19 (22.70)	20.33 (16.56)	9.633	p<.001
Avatars used in video games to represent different players	51.82 (19.47)	32.41 (21.31)	7.555	p<.001
Non-player characters (NPC) in video games	50.01 (24.89)	33.11 (21.12)	7.382	p<.001
Avatars used in Internet forums, social media and other online communities	41.64 (23.86)	26.14 (21.04)	5.216	p<.001

In summary, the analysis conducted on clusters (using Analysis of Variance and matched-pair t-tests) and individual groups within clusters (employing matched-pair t-tests) indicated that a three-cluster solution was deemed the most appropriate for the competence x experience space. Among these identified clusters, two clusters (Cluster 1 and Cluster 3), which comprises the majority of AI targets (13 members in total), displayed characteristics of mixed clusters.

Discussion

AI's integration into daily life is steadily growing. Yet, how people differentiate between AI agents remains relatively limited. The current set of studies attempted to map human perception of AI across a wide range of AI agents, using well-validated person and mind-perception theoretical models from social psychology as well as a novel, data-driven model derived from the other two. The findings revealed that AI is not perceived homogeneously, with certain stereotypes emerging. Specifically, AI agents formed distinct clusters within each model suggesting the emergence of shared stereotype among those within the same cluster.

When mapping AI perception using each one of the three models examine, mixed (or ambivalent) clusters (e.g., groups of cognitive targets scoring high in one dimension while low in the other) formed from multivalent perceptions, did not emerge in the application of SCM. They did not emerge in the application of the MPD model either. Finding mixed clusters is essential because their existence challenges the idea of a univalent and unidirectional perception of AI which also appears to be highly unlikely based on existing empirical evidence. For example, based on its design features, a robot has been shown that can be simultaneously perceived as highly agentic but very low in experience (Gray & Wegner, 2012) and chatbots can be seen as highly competent but not so warm or friendly (Kim & Im, 2023).

In person perception, the discovery of mixed clusters showed that such ambivalence exists in prejudice towards different social groups (A. J. C. Cuddy et al., 2007) while the application of SCM in animal perception demonstrated that ambivalence also characterised the way people see animals e.g., separating them into low warmth/high competence (predators), high warmth/high competence (companions), high warmth/low competence (prey), low- warmth/low competence (pest). However, not finding mixed clusters in either the SCM or the MPD model for AI targets suggests neither model properly captured the ambivalence in AI perception at the time the mapping was performed. The novel AISM model introduced here, derived from the other two, identified some mixed clusters. This suggests that it may offer a more effective framework for mapping AI perception, at least at the time the mapping was conducted using the AI agents available then. Further replications of the current set of studies with diverse samples and over time could seek to further assess and validate the AISM model's applicability in mapping AI perception over time.

Also, the regression analysis of human likeness on the SCM and MPD models' dimensions (e.g., competence, warmth, agency, and experience) revealed noteworthy findings; perception of competence of an AI agent negatively predicted human likeness while perception of experience positively predicted human likeness. In other words, the results obtained in the current set of studies suggest that the less competent an AI is seen and the more it is perceived as capable of experiencing things, the more likely people are to view it as humanlike.

Implications and Directions for Future Research

The way various AI agents were positioned within each of the two-dimensional spaces of the SCM, MPD, and AISM reveals distinct patterns of AI perception across each model's dimensions. However, as AI agents become more integrated into daily life and people gain more experience with them, repeating the same study at different time points may show the same AI agents to occupy the two-dimensional spaces differently. Experience and familiarity with AI are likely to influence perceptions and conducting the study over multiple periods could help document how AI perception changes over time. Additionally, it would allow for updating insights on AI perception with the mapping of new AI agents that will be invented and reach the public.

Moreover, there is a methodological benefit from the mapping of AI across a range of different agents. It gives a 'forest' rather than a 'tree' view of AI perception, from which past findings can be framed and new research questions can emerge. For example, what factors drive AI agents that do not share morphological similarities (e.g., embodiment) or purposes of use to cluster together? Take for instance Cluster 3 in the AISM solution (Figure 4), which includes AIs that scored higher than those in the other two clusters in terms of experience. What might be driving high ratings of experience among these AIs? It might be that all of them are designed in a

way that resembles humans or simulate human-like behaviour, inevitably fostering the impression of experience, despite not actually having this capability. Future research should look at gaining a deeper understanding of the underlying factors that drive shared perceptions of competence or experience across AI agents sharing the same clusters.

Also, the aim of the current set of studies was to map AI perception and as such does not examine the factors driving those perceptions, nor does it distinguish between perceptions based on an accurate understanding of the AI's capabilities and those shaped by misperceptions. Particularly about misperceptions, however, it does bring to the surface the question of the extent to which an AI agent's perceived competence or experience are subject to misperception. For instance, AI in the form of Large Language Models (LLMs) has the potential to simulate human consciousness to the extent that people may struggle to distinguish them from humans (Bender et al., 2021) leading to potential misperception of its abilities, especially its ability to experience things. Conducting further research to understand and address public misperceptions around specific AI agents is important for fostering effective interaction, building public trust, and promoting acceptance of AI. It also plays a crucial role in shaping policy regulations and guiding moral considerations related to AI. For instance, public misperceptions can significantly influence policy decisions and regulatory frameworks. If people misunderstand an AI agents' capabilities and limitations, they may advocate for either overly restrictive regulations or insufficient safeguards, potentially hindering the responsible development and deployment of AI technologies. Similarly, in the case of self-driving cars, if people form misperceptions of competence or overestimate the agency of self-driving cars, they may inappropriately attribute moral responsibility to these systems, leading to ethical dilemmas. Identifying where an AI agent

is positioned in the two-dimensional space of competence x experience, will help assess any misperceptions that may arise.

Finally, the findings have implications for AI systems design, potentially confirming what AI engineers have already figured out by test and error. If the goal is to make AI be perceived more humanlike, introducing a degree of imperfection, or/and incorporating cues that suggest the AI's ability to simulate mental states seem to be keyways for achieving it.

Limitations

The current set of studies has limitations. First, every participant had at least one instance where they selected a 'Does not apply' answer for at least one of the adjectives used under each of dimensions (Table 1). Although no participant selected 'Does not apply' for all the adjectives under the same dimension, different adjectives were seen as not applicable for describing AI by different participants (e.g., not all participants considered the same adjectives as not applicable). This does not affect the validity of the findings as the 'Does not apply' ratings were not included in the data analysis; it nevertheless cautions about the appropriateness of some of the adjectives used. Future studies should seek to critically assess the subset of adjectives within each of the four dimensions to ensure their suitability for AI. Here we chose to use the exact adjectives used in the SCM and MPD studies. Future studies could seek to ask different questions to better operationalise each one of the four dimensions. For instance, the ability to learn is something that could be asked nowadays about AI as most people are likely to view Large Language Models (LLMs) as able to learn.

Secondly, the current set of studies deliberately presented different AI agents using high-level descriptions, avoiding the use of well-known commercial examples. For example, we aimed to evaluate perceptions of 'an AI that acts as a personal assistant taking voice commands'

rather than mentioning or explicitly asking people about e.g., ‘Alexa’ or ‘Siri’. This approach was chosen to minimise potential confounding factors arising from biased perception towards specific AI agents under the boarder category. An alternative approach could have involved presenting participants with multiple examples of AI agents falling into the same broader category to account for variations in perception within the range of a given category. In this approach for instance, in the case of AI that recommends movies, people would be asked about their perceptions of a variety of commercially available movie recommender systems (e.g., Netflix, Apple tv etc.) or in the case of social robots, they would be asked about a variety of social robots (e.g., Pepper, Nao, Paro, Aibo, ElliQ etc.) provided that the public knows about them. Future studies could attempt such variation in the experimental design. Additionally, alternative ways for presenting AI agents—such as with the use of images, videos, or direct interactions—should be explored.

Finally, it is advisable in self-report measures to ask multiple (usually two or three) questions to capture people's responses on the construct of interest, in order to account for potential errors arising from misunderstandings of the question. We used one question to assess human likeness: ‘*How much humanlike do you think most people find this AI*?’ (on a scale from 0 to 100). Ideally, we would have included at least one additional question using a synonymous or a word close in meaning to ‘humanlike’ to validate whether the responses to the two would be correlated (e.g., providing a Cronbach’s alpha of more than 0.6). A possible alternative question could have been, ‘*How much do you think most people would perceive this AI as resembling humans?*’. Future studies should include alternative questions to inquiry self-report perceptions of human likeness, as long as they can identify different ways to phrase the same question

without duplicating those already asked under the dimensions of experience, agency, warmth, and competence.

Conclusion

Three theoretical models were employed to map how humans perceive AI. The findings underscore that AI is not perceived homogeneously and reveal stereotype formation in AI perception. Future iterations of the studies in different points in time could reveal potential changes in the patterns observed as further integration of AI in human life and increased familiarity and experience with AI are likely to impact perceptions. Additionally, future iterations will enable the mapping of perceptions across new AI agents, as the technology continues to evolve and new, more sophisticated, and humanlike AI agents are expected to reach the public.

Chapter 3 The 3Ps of Trust in AI (Performance, Process and Purpose)

Introduction

Artificial Intelligence agents (AI agents) can be any system that perceives and acts on its environment to maximise its chance of achieving its goals (Poole et al., 1998). From Large Language Models (LLMs) like ChatGPT, the friendly voices of Alexa and Siri that guide us through our tasks, and efficient chatbots streamlining customer service to the lifelike avatars fostering connections in virtual communities, diverse AI agents are seamlessly integrating into our lives. We witness their proliferation and gradual integration into the social fabric where they evolve into novel interaction entities. At the same time, ongoing advancements in generative AI (Zhao et al., 2023), Human-Computer Interaction (HCI) (Helander, 2014) and Human-Robot Interaction (HRI) (Goodrich & Schultz, 2008) aim to enhance both the capabilities of these agents and the quality of the interactions people have with them, striving to make these interactions feel as intuitive and natural as human-to-human interactions do, if not more so.

Human-to-human interactions can be intricate, as they involve people making inferences about other peoples' thoughts, feelings, goals and intentions (Fiske, 1998; Frith & Frith, 2006). While many interactions with AI agents require inferring their mental states, like for instance when playing chess against a computer, AI agents do not possess mental states. Still, they are *agents* - entities endowed with the ability to gather, process information, make decisions, and dynamically engage with their environment, including humans and other AI agents, to accomplish shared objectives with varying degrees of autonomy (Russell & Norvig, 2016). As such, interactions with AI necessitate understanding features such as i.e., the goals instilled in the AI agent being interacted with.

How do people evaluate AI? We began by reviewing the literature on trust in automation to identify key features of automated systems that influence human evaluations, particularly evaluations of trustworthiness of the automation. We focused on the factors that influence trust, given its well-established role in decision-making involving human (Bonaccio & Dalal, 2006; Brynjolfsson et al., 2019; Haran & Shalvi, 2020a; Laban & Araujo, 2020; Rahwan, Cebrian, Obradovich, Bongard, Bonnefon, Breazeal, Crandall, Christakis, Couzin, & Jackson, 2019; Sniezek & Van Swol, 2001) and AI advisors (Brynjolfsson et al., 2019; Laban & Araujo, 2020; Rahwan, Cebrian, Obradovich, Bongard, Bonnefon, Breazeal, Crandall, Christakis, Couzin, & Jackson, 2019). Next, we turned to two more recent bodies of literature that focus on the factors influencing trust in AI and trust in HRIs. These literatures separately and collectively underscore three enduring features crucial in shaping evaluations of trustworthiness: performance, purpose, and process. In the rest of the chapter, we refer to these three features as key determinants of trust in AI or just '3Ps' for ease of reference.

Trust in Automation, AI, and Robots

In the context of trust in automation, literature examining the features of the automation that influence trust identifies *performance*, *purpose*, and *process* as key determinants (Chiou & Lee, 2023; Hoff & Bashir, 2015; Lee & Moray, 1994; Lee & See, 2004; Perkins et al., 2010; Schaefer et al., 2016). *Performance* refers to the current or historical operation of the automation Lee and See (2004). It includes traits such as reliability, predictability, and ability to achieve the specific goals the automation was designed to achieve. It can be considered as describing the '*what'* the automation does. *Process* refers to the automation's algorithms and the degree to which these algorithms are appropriate for the situation they are designed. It includes traits such as transparency, interpretability and explainability and can be considered as describing the '*how'*

the automation operates. *Purpose* refers to the degree to which the automation is being used within the realm of its designers' intentions. It can be considered as describing the '*why'* the automation was developed. Parallels can be drawn between trust in automation and trust in interpersonal relationships when considering the 3Ps. Just as trust between individuals is built upon factors like competence, integrity, or benevolence trust in automation hinges respectively on the performance, process, or purpose of the automation (Mayer et al., 1995).

A more recent meta-analysis on the antecedents of trust in AI also draw the spotlight on features of the AI agent, along with human-related and contextual antecedents of trust (Kaplan et al., 2023). Among the 67 studies examined, AI performance emerges as a noteworthy predictor of trust, showing a substantially large effect size (d=1.47). Also, an AI whose process is transparent was found to be more trustworthy compared to 'black-box' AI (d=0.24), and AI's behaviour was shown to significantly impacts trust (d=0.81) with, good-intentioned, honest, and rule-abiding AI perceived as more trustworthy than its deceptive, mal-intentioned counterpart.

In addition, two sequential and comparative meta-analyses (Hancock et al., 2011; Hancock et al., 2021) investigating the empirical evidence on the determinants of trust in HRI revealed that attributes associated with the robot itself exert a more substantial influence on trust compared to attributes related to the human or the contextual aspects of the interaction. Interestingly, within the realm of robot attributes shaping trust, two aspects emerge prominently: The *robot's personality* and its *communication style*. As outlined by the authors, the concept of *robot personality* emerged as particularly influential, demonstrating a significant correlation with trust. It encompasses attributes such as positive facial expressions, empathy, likability, and sociability. However, since robots embody AI, their facial expressions and empathetic traits can be viewed as reflections of their underlying purpose, indirectly highlighting again the importance

of cues that signal purpose, akin to the findings in both literatures concerning trust in automation and trust in AI. Similarly, while not explicitly detailed in the examined meta-analyses, one could interpret the term *'communication style'* as encompassing aspects of performance and process (the remainder two of the 3Ps). Notably, varying degrees of the robot's reliability and predictability of actions (traits related to performance), as well as the explainability or interpretability of its behaviour (traits related to process), were found to lead to different communication styles.

Overall, the above bodies of literature separately and collectively underscore the lasting importance of the 3Ps in shaping how people form evaluations, particularly those related to the trustworthiness of AI.

Research Question

Since the 3Ps (performance, process, and purpose) are fundamental in trust evaluations of automated systems, how do people evaluate AI models based on these trust determinants? This is the research question we sought to address here. This research question was further divided into the following three hypotheses (h1, h2, h3): reliance on an AI model's output will vary depending on evaluations of its performance, process, and purpose (h1), attitudes towards an AI model will vary depending on evaluations of its performance, process, and purpose (h2), trust in an AI model will vary depending on evaluations of its performance, process, and purpose (h3). The three hypotheses differentiated types of trust measurement. Reliance on AI advice (h1) was measured through a behavioural task, while attitudes towards AI (h2) and trust in AI (h3) were assessed using self-report measures. This distinction aimed to compare and complement findings from different measurement types. These hypotheses were addressed in two studies (Studies 1 and 2).

The Judge Advisor System (JAS) Paradigm

The Judge Advisor System (JAS) paradigm was used for measuring the extent to which participants followed AI advice in both Study 1 and Study 2. Based on the JAS paradigm participants make an initial decision under uncertainty and then they receive advice, by one or multiple advisors, before they are given the opportunity to make a final, possibly revised, decision (Sniezek & Buckley, 1995),. The JAS paradigm has been frequently used in advice uptake literature as a measure of the extent to which judges rely on advice from one or more human advisors (Haran & Shalvi, 2020b; Yaniv, 2004a, 2004b; Yaniv & Choshen-Hillel, 2012; Yaniv & Kleinberger, 2000). Unlike self-report measures, which typically involve asking participants about their actions in hypothetical scenarios, the JAS paradigm enables the direct observation of participants' behaviour.

Following the methodology of the JAS paradigm, reliance on advice from an AI was operationalised using the Weight of Advice (WoA). WoA is a ratio that compares the distance of the final estimate from the initial estimate to the distance of the advice from the initial estimate. WoA typically spans from 0 to 1 and has been previously used in several studies in the context of advice uptake (Harvey & Fischer, 1997; Yaniv, 2004b) as well as algorithmic advice uptake (Logg et al., 2019). A WoA of 0 occurs when participants ignore the advice and stick to their initial estimate. Conversely, a WoA of 1 occurs when participants update their final decision to match the advice given. Also, for numerical estimates, WoA of 0.50 occurs when participants average the advice given and their initial estimate. Indeed, if the goal is to maximise accuracy, previous research has shown that, for an individual seeking advice from a randomly selected person, it is recommended to average their own judgment with the advice received, resulting in a WoA of 0.50 (Dawes & Corrigan, 1974; Einhorn & Hogarth, 1975). However, people tend to

update their judgment only by 0.30-0.35 on average when considering advice, leading to decreased accuracy (Liberman et al., 2012; Soll & Larrick, 2009).

Study 1

Methods

Participants

UK participants were recruited through the Prolific experimental subject pool for compensation. Participants were paid based on an hourly rate (£9/hour) for the time spent in the study. The sample size was determined a priori with a goal of obtaining 0.95 power to detect a medium effect size of f^2= 0.17 taken from previous research that uses the JAS paradigm to measure algorithmic advice uptake at the standard 0.05 alpha error probability. Power analysis was conducted using statistical software program R, package '*pwd'*, function *'pwr. f2.test'* which is suitable for general linear models. It suggested a minimum sample size of 106 participants. In total, 134 participants were recruited to accommodate potential exclusions due to failures in attentional checks and completions in less than the minimum duration of 15 mins (accepted minimum duration was set to be the one-third of the estimated total duration of the study, which was found to be around 45 mins, based on the completion times gathered during a pilot run of the study). Two attentional checks were included in the study with the aim of removing participants who would fail both checks. Three out of the 134 participants failed both attentional checks. Additionally, one additional participant was excluded based on their response to the question regarding the overall study experience, suggesting they may have inferred the manipulation being employed. This resulted in a final sample of 130 participants, averaged 35-44 years of age, 82% female, 17% male, 1% non-binary and 1% preferred not to say of ethnicities: 86% White, 6% Black, 3% Asian, 2% Mixed and 2% Other Ethnicity. Finally, prior to the study, that, data

entries with WoA that is larger than 2 or less than -1 will be excluded. From our initial sample of 130 participants, there were 8 participants whose average WoA was above 2 and thus were excluded while there was no participant with an average WoA less than -1. Therefore, the final total sample size in Study 1was 122 participants. The exact materials and data for Study 1 are available in the Open Science Framework at https://osf.io/xtd4e

Materials and Procedure

The experiment and data collection were carried out using the Qualtrics survey platform. Upon accepting the task on Prolific, participants were directed to Qualtrics where they read the instructions and gave their consent to participate in the study. They were then presented with eight question blocks, each corresponding to one of the eight AI models (from now on referred to as AIs for simplicity) under evaluation, presented in a random order.

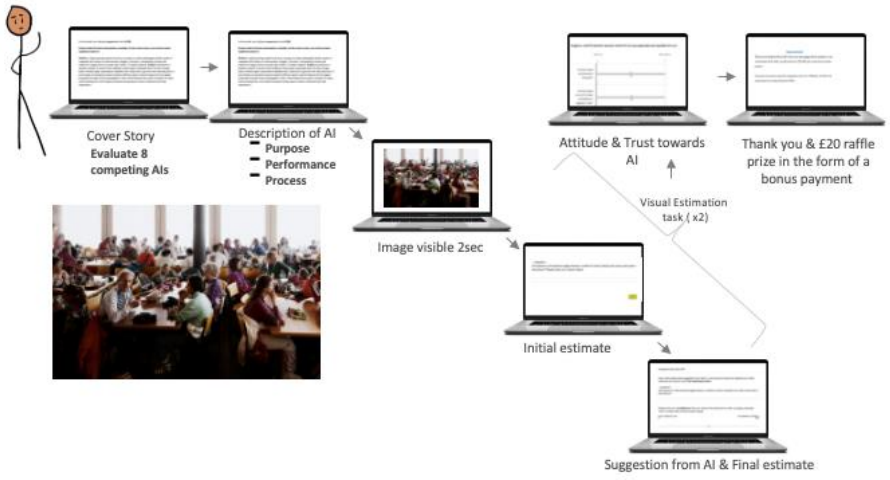
Participants were tasked with evaluating the eight competing AIs following the same structured process per AI. First, they read a paragraph describing the AI's performance (low vs high), purpose (good vs bad) and process (an explanation of its algorithm which was either complex/more technical or simple/less technical). Subsequently, they had to perform two visual recognition decision tasks. In each one, they viewed an image for a brief display time (2 sec) and were asked to estimate how many objects were depicted. After that, participants were presented with the AI's estimate. Finally, they were asked to provide their final, possibly revised estimate based on the AI's estimate. Essentially, participants needed to decide whether to revise their initial estimate to align with the AI's estimate based on how much they trusted it. Additionally, they indicated their confidence level in their final estimate on a continuous scale ranging from 0 to 100.

Following the evaluation phase, participants were directed to a separate screen to express their attitude toward the AI they had evaluated. They indicated on a continuous scale from 0 to

100 how much they would like to see the AI utilised in real-life situations by responding to the question, '*Imagine a real-life scenario where this AI was applicable and available for use. How much would you like to see it being used?*'. Additionally, participants rated their level of trust in the AI by answering, '*Imagine a real-life scenario where this AI was applicable and available for use. How much would you trust it to make the decisions it's designed to make?'* also on a continuous scale from 0 to 100[3]. At the conclusion of the study, participants were requested to provide their demographic information and were presented with details about an online raffle. This raffle served as an incentive for participants to provide accurate estimate as each correct estimation they provided would earn them a virtual ticket for a £20 raffle. The more tickets they accumulated, the higher their chances of winning the £20 raffle. The above procedure is visually represented in Figure 5.

Figure 5

Experimental Procedure in Study 1



[3] In Study 1, we used a continuous scale but with a lower limit of 1 and an upper limit of 7 instead of 0 and 100. We adjusted this scale to range from 0 to 100 in Study 2, based on participants' feedback that a continuous scale from 0 to 100 scale made more intuitive sense than a continuous scale from 1 to 7. To ensure consistency across the studies, when analysing the data from Study 1 we rescaled the answers to reflect a 0 to 100 continuous scale for all reported results.

Study 1 employed a 2 (performance: high vs low) x 2 (purpose: good vs bad) x 2 (process: simple vs complex) within-subject variables design, resulting in 8 conditions (e.g., blocks of questions) which every participant went through in random order. Reliance on AI advice was measured by calculated the WoA in each one of the two visual recognition tasks per AI. Because in each condition participants did two visual recognition tasks, the overall WoA per condition and participant was determined as the mean of the two individual WoA values.

Also, to ensure that the estimate from each of the eight AIs was not a confounding factor, we employed a control mechanism. This involved programming the AI advice to be generated by adding a randomly generated number, denoted as N, to the participant's initial answer. This number N ranged from +5 to +9 or from -5 to -9. This approach aligns with strategies utilised by other researchers to control the AI output across conditions (Hou & Jung, 2021).

Data Analysis Strategy

Data were analysed using R statistical software (https://www.r-project.org/). Participants' WoA, Attitude towards the AI and Trust in the AI were analysed using three ANOVAs, each with three within-subjects factor variables: performance (high vs. low), purpose (good vs. bad), and process (complex vs. simple). Degrees of freedom for all within-subjects factor variables were corrected for sphericity violations using the Greenhouse-Geisser correction. To further investigate whether there is a different effect of process on WoA for each performance level, we calculated the simple main effect of process across the performance factor variable using 'emmeans'. We controlled for multiple testing in this analysis by adjusting the six p-values using the Bonferroni-Holm adjustment.

Results

Manipulation checks

To check whether the three manipulations introduced (e.g., high/low performance, good/bad purpose, and simple/complex process) were successful, relevant manipulation check questions were included. Specifically, to test the manipulation of performance (high vs low), we asked participants '*Based on the description you've just read, how would you rate the performance of this AI (on a scale of 0 –100)?'* To test the manipulation of purpose, we asked, '*Do you think this AI's purpose is morally good or bad (on a scale of 0 – 100)?'* with 0 being morally bad and 100 being morally good. Finally, to test the manipulation of process, we asked, *'How easy/difficult did you find it to understand how this AI works (on a scale of 0 – 100)?'* with 0 being 'Very difficult, I'm puzzled' and 100 being 'Very easy, it's clear to me how it works'. The relevant t-tests performed between the mean performance scores of AIs described as high vs low performing, t (519) = 39.88, p<.001, the mean scores of AIs described as having a good vs bad purpose, t (519) = 16.47, p <.001, and the mean level of difficulty in understanding the process of an AI described in a complex (more technical) vs a simple (less technical) way, t (519) = 16.96, p<.001, were all statistically significant (p<.001), suggesting that all three manipulations in place were successful.

Data Preparation

Typically, we anticipate a WoA value to fall within the range of 0 to 1, yet the WoA metric is subject to certain constraints (Gino & Moore, 2007). One such limitation is the potential for final estimates to fall outside the range between advice and initial estimate. While this occurrence has been rare in prior research (Gino, 2008; Harvey & Fischer, 1997), as a precautionary measure, we preregistered that we would adjust participants' WoA to 0 or 1 if their final estimate deviated from the range between advice and initial estimate. This adjustment adhered to the customary practice observed in the literature (Hou & Jung, 2021; Logg et al.,

2019), whereby a WoA below 0 (but above -1) is replaced with 0, and a WoA exceeding 1 (but less than 2) is replaced with 1. We made two replacements of WoA below 0 with but above -1 with 0. No WoA value exceeded 2.

WoA

The ANOVA revealed a significant main effect for *performance, F* (1, 121) = 99.37, η^2= 0.10, $p < .001$, indicating that on average participants' WoA increases with the AI's performance (Figure 6). There was also a significant main effect for *process*, *F* (1, 121) = 6.68, η^2= 0.003, $p <$ .05 suggesting that more complex (e.g., more technical) descriptions of the AI's algorithms led to higher WoA from participants (Figure 7), whereas the main effect of *purpose* was not found statistically significant, *F* (1, 121) = 2.10, η^2= .001, $p = .15$.

The main effects of *process* and *performance* were qualified by a significant *performance X* process interaction, *F* (1, 121) = 4.12, η^2 =.003, $p <$. 05. There was also a significant *process* X *purpose* interaction effect, *F* (1, 121) = 10.12, η^2= 0.01 $p < .01$. The remaining of the interactions *purpose* X *performance* interaction, *F* (1, 121) = 0.12, η^2 <.001, $p =$. 729, and *purpose* X *performance* X *process*, *F* (1, 104) = 1.47, η^2 <.001, $p =$. 023 were not statistically significant.

Figure 6

Reliance on Advice (WoA): Main Effect of Performance (Study 1)

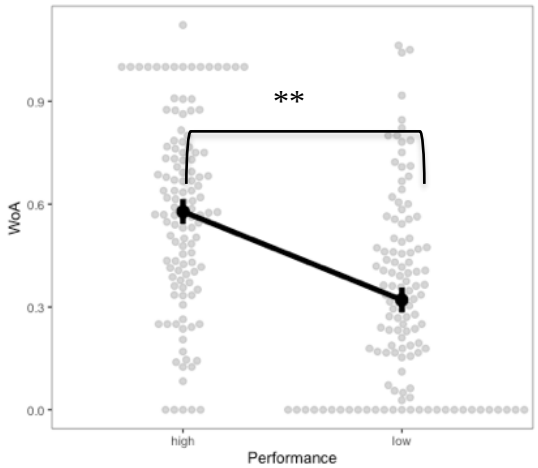
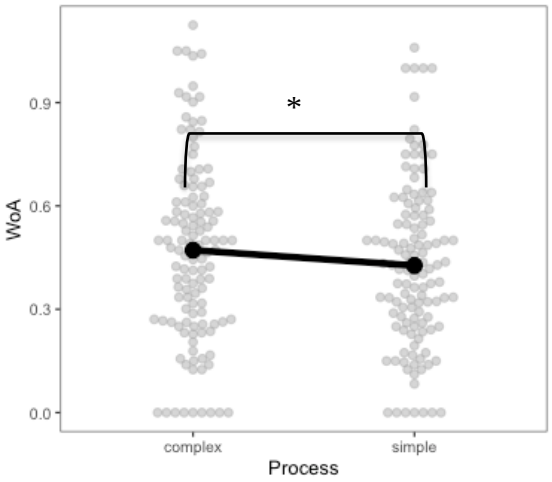


Figure 7

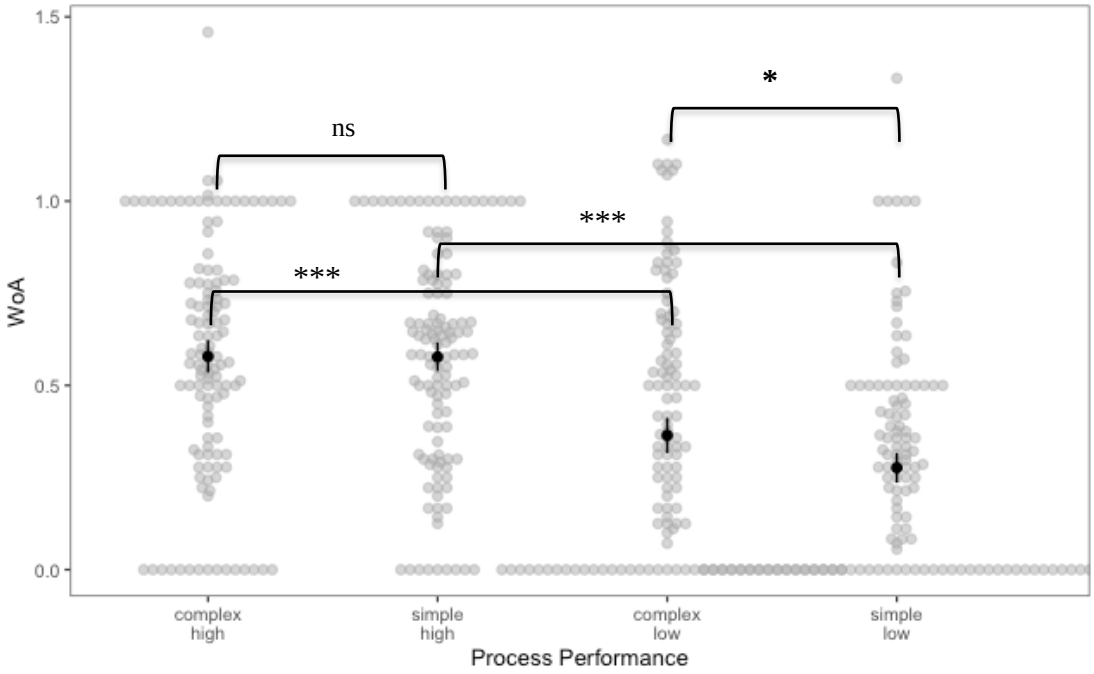
Reliance on Advice (WoA): Main Effect of Process (Study 1)



Note. Error bars represent 95% confidence intervals, * $p < 0.01$, ** $p < 0.001$

Figure 8

*Reliance on Advice (WoA): Process * Performance (Study 1)*



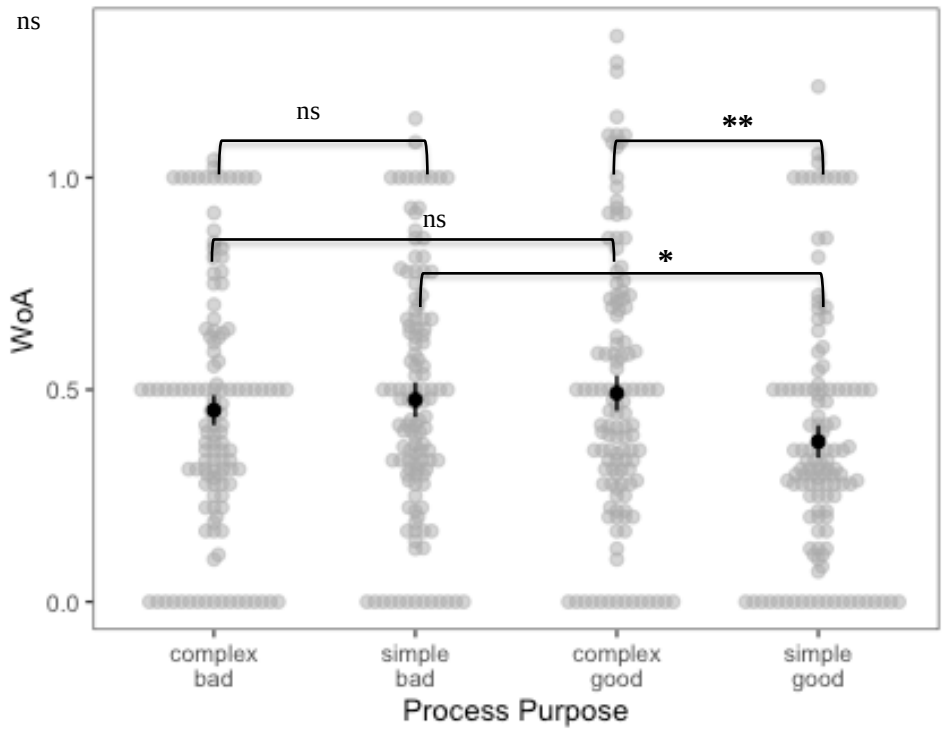
Note. Error bars represent 95% confidence intervals, * $p < 0.05$, ** $p < 0.001$, *** $p < 0.0001$, ns: non-significant

Looking closer at the *process* X *performance* interaction (Figure 8), a simple effect analysis showed that when the AI was described as low performing, there was a significant difference in WoA between AIs with complex vs simple explanations (t (121) = 3.05, $p < .01$) such that, for AIs described as having low performance, participants on average gave more weight to the AI's advice when it was accompanied with a complex (more technical) rather than a simpler (less technical) explanation of its process. This is an interesting finding since, perhaps contrary to what one might expect, it suggests that when the AI is low performing, reliance on AI increases with more complex (more technical) explanations. We further discuss this finding and provide plausible interpretations in the Discussion section.

When the AI was described as high performing, the complexity in the explanation of its algorithmic process (complex vs. simple) had no significant effect on WoA (t (121) = 0.04, p = 0.966). We also examined the simple effect of process on performance and found that, when the description of the process was complex, the average WoA was statistically significantly higher for AIs described as high vs low performing (t (121) =5.85, p<.0001). Likewise, when the description of the process was simple, again the average WoA was statistically significantly higher for AIs described as high vs low performing t (121) = 9.98, p <.0001. Both above results show that WoA increased with performance regardless of the level of complexity in the explanation of the process.

Figure 9

*Reliance on Advice (WoA): Process * Purpose (Study 1)*



Note. Error bars represent 95% confidence intervals, * p =.0, ** p< 0.001, ns: non-significant

Finally, looking closer at the *process* X *purpose* interaction (Figure 9), a simple effect analysis showed that when the AI was described as serving a good purpose, there was a significant difference in WoA between AIs with complex vs simple explanations (t (121) = 3.96 p < .001) such that, for AIs that described as having good purpose, participants on average gave more weight to the AI's advice when it was accompanied with a complex (more technical) rather than a simpler (less technical, high-level) explanation. This is again an interesting finding as it goes against the intuition that simple explanations are always preferred to complex and suggests that when AI is considered as serving a good purpose, reliance on AI increases with more complex (more technical) explanations. We further discuss this finding and provide plausible interpretations in the Discussion section.

When the AI was described as serving a bad purpose, the complexity in the explanation of its algorithmic process (complex vs. simple) had no significant effect on WoA (t (121) = -

0.92, p = 0.718). Finally, we also examined the simple effect of process on purpose and found that, when the description of the process was complex (more technical), the purpose of the AI (good vs bad) had no statistically significant effect (t (121) = -1.37, p = 0.518). However, when the description of the process was simple, purpose mattered such that the average WoA of AI with purpose: bad was found to be statistically significantly higher than the that of AI with purpose: good (t (121) = 3.284, p =.01), a puzzling finding as normally we would expect people to prefer AIs with good purpose.

Attitudes toward AI

When it comes to the self-report measure of attitudes towards AI, there was a significant main effect for *performance, F* (1, 129) = 143.19, η^2= 0.17, p < .001, showing that participants' attitudes on average became more positive with the AI's performance. Additionally, there was a significant main effect for *purpose, F* (1, 129) = 48.40, η^2= 0.7 p < .001, showing that, on average, more positive attitude was reported towards good-purposed than bad-purposed AIs. The main effects of *performance* and *purpose* were qualified by a significant *performance* X *purpose* interaction (F (1, 129) = 5.11, η^2= 0.02, p =0.03). A follow -up simple effects analysis performed for the *performance* X *purpose* interaction showed that when the performance was low, attitudes towards AI were more positive for AIs described as having a good rather than a bad purpose (t (129) = -5.54, p < .001) and when the performance was high, attitudes towards AI were more positive for AIs described as having a good rather than a bad purpose (t (129) = -6.69, p < .001).

Also, there was a significant *performance* X *process* interaction, F (1, 129) = 13.55, η^2= 0.04 p <.001. A follow -up simple effects analysis performed for the *performance* X *process* interaction showed that when the performance was low, attitude towards AI were more positive for AIs with complex rather than simple explanations (t (129) =2.79, p= 0.01) whereas when the

performance was high, attitude towards AI were more positive for AIs with simple rather than complex explanations (t (129) = -1.954, p= 0.05). The remaining interactions e.g., *purpose* X *process (F* (1, 129) = 0.69, η^2< 0.01, p =0.41) and *purpose* X *process X performance (*F (1, 129) = 2.49, η^2< 0.01, p =0.12*)* were not significant.

Trust in AI

When it comes to the self-report measure of Trust in AI, the only effect that was significant was the main effect for *performance*, F (1, 129) = 236.15, η^2= 0.40, p < .001, showing that participants' trust on average increased with the AI's performance, from M= 38.0, SD= 2.27 when the AI was described as low performing to M=75.2, SD= 1.19. The main effect of purpose (F (1, 129) = 0.76, η^2<.001, p = .384) and the main effect of process (F (1, 129) = 2.05, η^2< 0.001, p =.155) were not significant. Also, no interaction effect was found significant (e.g., *purpose* X *performance*, F (1, 129) = 2.75, η^2<.001, p = .100, *purpose* X *process*, F (1, 129) = 0.67, η^2<.001, p = .413, *performance* X *process*, F (1, 129) = 4.00, η^2=.001, p = .047 *purpose* X *performance* X *process*, F (1, 129) = 6.54, η^2 =003, p = .012).

Study 2

In Study 2, we replicated the experimental design of Study 1 with one key difference: the decision task. While Study 1 involved a visual recognition task without moral implications, Study 2 required participants to complete a resource allocation task, which by nature carries moral considerations.

Methods

Participants

UK participants were recruited through the Prolific experimental subject pool for compensation. Participants were paid based on an hourly rate (£9/hour) for the time spent in the

study. As in Study 1, the sample size was determined a priori with a goal of obtaining 0.95 power to detect a medium effect size of $f^2 = 0.17$ taken from previous research that uses the JAS paradigm to measure algorithmic advice uptake, at the standard 0.05 alpha error probability. Power analysis was conducted using statistical software program R, package '*pwd'*, function *'pwr. f2.test'* which is suitable for general linear models. It suggested a minimum sample size of 106 participants. We recruited in total, 120 participants, e.g., 14 more than the suggested minimum sample size, to accommodate potential exclusions due to failures in attentional checks and completions in less than the minimum duration. No participant had to be excluded due to failure of attentional checks or any other reason. This resulted in a final sample of 120 participants, averaged 25-34 years of age, 61% female, 38% male and 1% not to say. Study 2's sample comprised the following ethnicities: 84% White, 5% Black, 4% Asian, 3% Mixed, and 4% Other. Out of the final sample of 120 participants, 15 individuals were excluded because their average WoA exceeded 2, thus falling outside the predetermined range of -1 to 2. Therefore, the total sample size used for the WoA data analysis in Study 2 was 105 participants. The exact materials and data for Study 2 are available in the Open Science Framework at https://osf.io/quy2m.

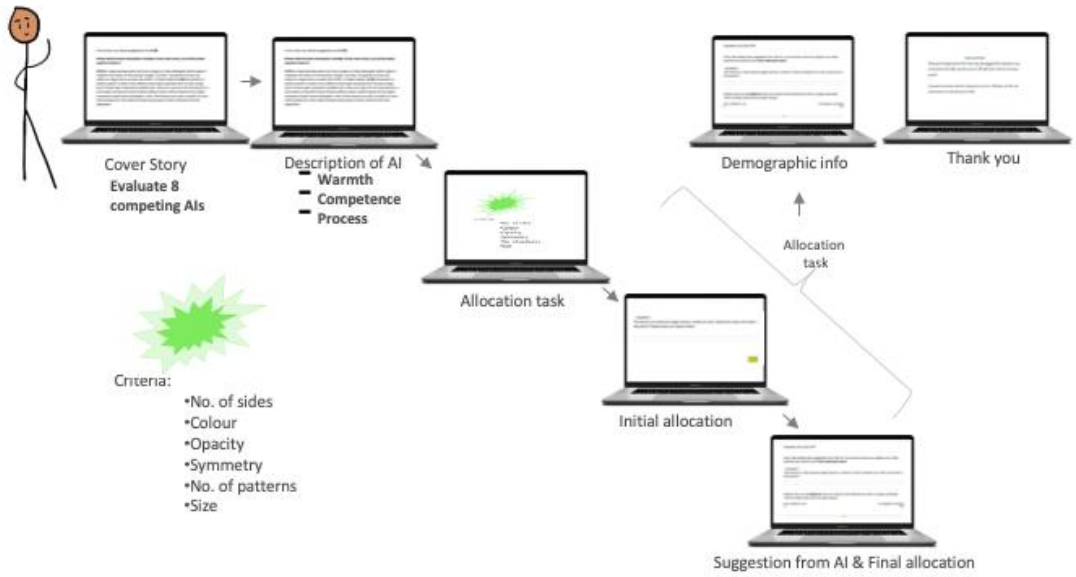
Materials and Procedure

Similar to Study 1, participants were given a cover story asking them to evaluate how well each of eight competing AIs performed. To assess each AI's performance, participants were presented with eight scenarios in which they acted as decision-makers with limited resources, tasked with finding the best way to allocate these resources to individuals in need. Rather than photographs of individuals, geometrical shapes were used to depict people in order to mitigate potential biases associated with visible characteristics such as gender, race, or age, which could

influence impressions and decisions (Gilovich et al., 2002; Martín & Valiña, 2023; Tversky et al., 1982). As such, six criteria pertaining to the shape—including the number of sides, colour, opacity, symmetry, patterns, and size—were used to reflect aspects of a person's circumstances affecting their financial needs. For example, shapes with more sides, warmer hues, greater symmetry, or more intricate patterns indicated higher levels of need. Prior to allocating points, participants were given an explanation of the six criteria and their meaning, and they were also informed that the shapes represented real individual in need of the resources and that the 50 points maximum that they would be allocating to the shapes represented monetary payments for these individuals. The procedure employed in Study 2 is visually represented in Figure 10.

Figure 10

Experimental Procedure in Study 2



For the eight competing AIs, participants were informed that the advice came from machine learning algorithms designed to make allocation decisions based on six criteria. Performance was explained as the number of criteria an AI used: more criteria meant higher

performance. A low-performing AI used one or two criteria, while a high-performing AI used five or six. Purpose was operationalised as generosity, with some AIs designed with the purpose of helping individuals in need of the resources by recommending favourable allocations while other AI were designed to be less helpful by consistently advising for less favourable allocations. For instance, a highly generous AI was consistently allocating the maximum points possible (e.g., from 47 to 50 out of 50 points). Conversely, a low-generosity AI was depicted as allocating fewer than 8 out of 50 points. Process was defined as explanation of the way the AI reached its advice in each scenario and it was either a high-level or a more technical explanation, with technical terminology used to reflect the two distinct levels (simple vs. complex). Study 2 did not include a raffle, as we wanted participants to be driven by moral considerations attached to the task at hand rather than being incentivised by winning a raffle.

Study 2 employed a 2 (performance: high vs low) x 2 (purpose: high vs low) x 2 (process: simple vs complex) within-subject variables design, resulting in 8 conditions which every participant went through in random order. Reliance on AI advice was measured by calculated the WoA. Finally, similar to Study 1, a control mechanism was employed to ensure the AI estimate did not act as a confound factor. As such, all low-generosity AI were programmed to suggest allocations that randomly ranged from 6 to 8 points, while the generous AI's allocations ranged from 47 to 50 points.

Data Analysis Strategy

Data were analysed using R statistical software (https://www.r-project.org/). The data analysis strategy for analysing the three dependent variables: WoA, Attitudes towards AI and Trust in AI was the same as in Study 1 (refer to the Data Analysis Strategy section under Study 1).

Results

Manipulation checks

To check whether the three manipulations introduced (e.g., high/low performance, high/low purpose, and simple/complex process) were successful, relevant manipulation check questions were included. Specifically, to test the manipulation of performance (high vs low), we asked participants '*How competent do you find this AI to be? (on a scale of 0 –100)?'* with 0 being extremely competent and 100 being extremely competent. To test the manipulation of purpose (operationalised as generosity), we asked, '*How generous do you find this AI to be? (on a scale of 0 –100)?'* with 0 being extremely ungenerous and 100 being extremely generous, and to test the manipulation of process, we asked, '*How easy/difficult did you find it to understand how this AI works?*' with 0 being 'Very difficult, I'm puzzled' and 100 'Very easy, it's clear to me how it works'. The relevant t-tests performed between the mean competence scores of AIs described as high vs low performing, t (479) = 17.75, p<.001, the mean generosity scores of AIs described as high vs low in generosity, t (479) = 26.23, p <.001, and the mean level of difficulty in understanding the process of an AI described in a complex (more technical) vs a simple (less technical) way, t (479) = 4.87, p<.001, were all statistically significant (p<.001), suggesting that all three manipulations in place were successful.

WoA

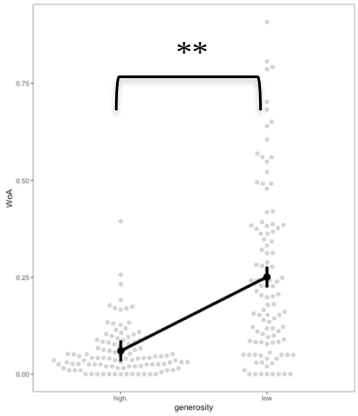
The ANOVA revealed a significant main effect for *purpose, F* (1, 104) = 96.92, η^2 = 0.12, p < .001 indicating that less generous AIs resulted in higher weight of advice (WoA) (Figure 11). This is an interesting finding as it goes against the intuition that AI suggesting higher allocations (e.g., high in generosity) would be preferred to AI suggesting lower allocations (e.g., low in generosity). We further discuss this finding and plausible interpretations

in the Discussion section. There was also a significant main effect for *performance, F* (1, 104) = 11.11, η^2 = 0.01, *p* = .001, showing that lower-performing AIs led to higher WoA (Figure 12). The main effect of *process* was not statistically significant, *F* (1, 104) = 0.01, η^2 < .001, *p* = .937.

The main effects of *purpose* and *performance* were qualified by a significant *purpose* X *performance* interaction, *F* (1, 104) = 11.44, η^2 = 0.01, *p* = .001. The remaining interactions were not found statistically significant: *purpose* X *process, F* (1, 104) = 0.02, η^2 <.001, *p* =. 884, *performance* X *process F* (1, 104) = 1.45, η^2 =.002, *p* =. 231, and *purpose* X *performance* X *process, F* (1, 104) = 1.02, η^2 = .001, *p* =. 315).

Figure 11

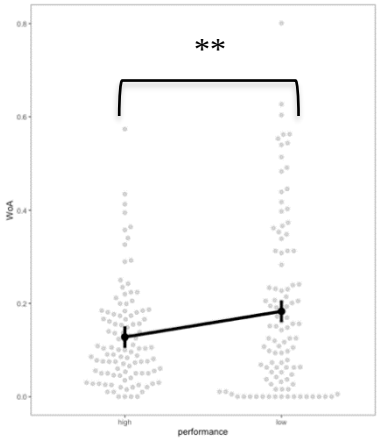
Reliance on Advice (WoA): Main Effect of Purpose (operationalised as generosity) (Study 2)



Note. Error bars represent 95% confidence intervals, ** *p*< 0.001

Figure 12

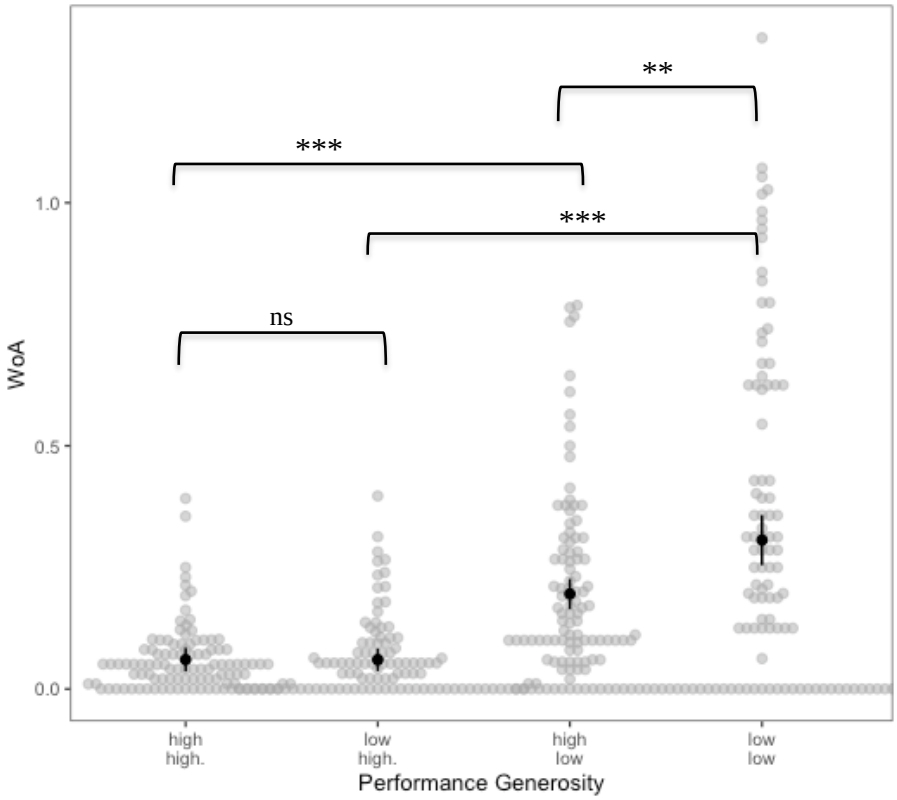
Reliance on Advice (WoA): Main Effect of Performance (Study 2)



Note. Error bars represent 95% confidence intervals, ** $p < 0.001$, *** $p < 0.0001$

Figure 13

*Reliance on Advice (WoA): Generosity * Performance (Study 2)*



Note. Error bars represent 95% confidence intervals, ** $p < 0.001$, *** $p < 0.0001$, ns: non-significant

Finally, looking closer at the *purpose* X *performance* interaction (Figure 13), a simple effects analysis showed that when the AI was described as low in generosity, there was a significant difference in WoA between high- vs low-performing AIs (t (104) = -3.46, $p < .001$) such that, participants placed more weight on its advice when the AI was described as low in performance rather than high. This was a puzzling finding as normally we expect people to prefer high to low performing AI. When the AI was described as high in generosity, its performance (high vs low) had no significant effect on WoA (t (104) = 0.03, $p = 0.977$). Finally, the simple effects analysis of performance on purpose showed that when the AI performance was high, the average WoA of AIs low in generosity was statistically higher than that those high in generosity, t (104) = - 7.51, $p <.0001$. Likewise, when the AI performance was low, again the average WoA of AIs low in generosity was statistically higher than that those high in generosity, t (104) = - 7.94, $p <.0001$. This suggested that participants trusted more the AI that was described as less generous, a finding that we further discuss in the Discussion section.

Attitudes towards AI

When it comes to the self-report measure of Attitudes towards AI, there was a significant main effect for *performance*, F (1, 119) = 176.82, η^2= 0.2, $p < .001$, showing that participants' reported attitudes were more positive for high performing (M= 58.6, SD= 1.70) compared to low performing AI (M= 33.6, SD= 1.94). Additionally, there was a significant main effect for *process*, F (1, 119) = 49.98, η^2= 0.03 $p < .001$, indicating that participants' attitudes on average were more positive when the AI's process was described in a simple (less technical) rather than a complex way (more technical). The *purpose* X *process* interaction was also significant, F (1, 119) = 49.28, η^2= 0.04, $p < .001$. A simple effects analysis showed that for AIs high in generosity, people on average reported more positive attitudes towards AI with simple (M=

55.90, *SD*= 1.94) rather than complex (*M*= 37.50, *SD*= 2.04) explanation (*t* (119) =-9.90, *p*< 0.0001) whereas for AIs low in generosity, the reported attitudes towards AIs with complex explanations (*M*= 46.30, *SD*= 2.03) did not differ significantly from those with simple explanations (*M*= 44.70, *SD*= 1.93), (*t* (119) =-9.90, *p*=0.371). The remaining interactions of *purpose* X *performance (F* (1, 119) = 0.30, η^2 <.001, *p* =. 584), *performance* X *process* (*F* (1, 119) = 0.001, η^2 <.001, *p* =. 955) and *purpose* X *performance* X *process* (*F* (1, 119) = 0.25, η^2 <.001 1, *p* =. 620) were not significant.

Trust in AI

The pattern of results for the self-report measure of Trust in AI were similar to the patterns of results for Attitudes towards AI. E.g., there was a significant main effect for *performance, F* (1, 119) = 196.94, η^2= 0.2, *p* < .001, showing that participants' trust on average increased with the AI's performance. Additionally, there was a significant main effect for *process, F* (1, 119) = 47.51, η^2= .03, *p* < .001, indicating that participants' trust on average was higher when the AI's process was described in a simple (less technical) rather than a complex way (more technical). The *purpose* X *process* interaction was also significant, *F* (1, 119) = 53.71, η^2= .05, *p* < .001. A simple effects analysis showed that for AIs high in generosity, people on average reported greater trust in AI with simple rather than complex explanation (*t* (119) 10.31, *p*< 0.0001) whereas for AIs low in generosity, the reported attitudes towards AIs with complex explanations did not differ significantly from those with simple explanations (*t* (119) = 1.41, *p*=0.160). The remaining interactions of *purpose* X *performance* (*F* (1, 119) = 0.13, η^2 <.001, *p* =. 714), *performance* X *process* (*F* (1, 119) = 0.07, η^2 <.001, *p* =. 785) and *purpose* X *performance* X *process* (*F* (1, 119) = 0.82, η^2 <.001 1, *p* =. 368) were not significant.

Discussion

Guided by the key determinants of trust in automation (Lee & See, 2004), we examined how information about an AI's performance, purpose, and process affects people's evaluations. This was done through two studies where we measured the weight participants placed on an AI model's output when making estimates in two tasks—a visual recognition task (Study 1) and a resource allocation task (Study 2). We conducted two studies (Study 1 and Study 2), varying only the nature of the decision tasks participants performed. E.g., in both studies, the decision tasks involved making numerical estimates; however, in Study 1, the estimates carried no moral implications for the participants (participants were asked to estimate the number of humans and objects in pictures), whereas in Study 2, the decision task involved numerical estimates carrying moral considerations, as participants were asked to estimate how to best allocate limited resources to people in need. Overall, the combined results of Study 1 and Study 2 suggested that people evaluate all three determinants of trust in automation (performance, process, and purpose) when evaluating AI models, however they weight each determinant differently depending on the moral nature of the decision.

Specifically, Study 1 (visual estimation task) revealed that participants' evaluation of the AI model's process explanation interacted with the evaluation of performance. E.g., when the AI was evaluated as low performing, participants gave higher WoA when the AI model's process explanation was detailed and technical rather than simple and high-level. However, when the AI was evaluated as high performing, the detail and technicality in the explanation no longer served as a heuristic (Figure 8). This finding holds a practical implication for explainable AI. It suggests that in the case of decisions that carry no moral considerations and when people evaluate an AI model as low performing (or still underperforming), they use the complexity of its explanation as

a heuristic, valuing more the advice that comes from an AI model with a more detailed and technical explanation rather a simple high-level one.

Moreover, in Study 1 (visual estimation task) the evaluation of the AI model's process explanation interacted with the evaluation of its purpose. When participants evaluated the AI model as overall serving a good purpose, they gave higher WoA when the explanation for its process was detailed and technical rather than simple and high-level. However, when participants evaluated the model as overall serving nefarious purposes, complexity no longer served as a heuristic (Figure 9). This finding holds another practical implication for explainable AI. It suggests that in the case of decisions without moral considerations and when people evaluate an AI model's purpose as good, they use the complexity of its explanation as a heuristic, valuing more the advice that comes from an AI model with a more detailed and technical explanation rather a simple high-level one.

Behavioural patterns, however, were different in Study 2, which was an exact replication of Study 1 with only one difference: the moral nature of the decision task at hand. This change of task (from one that did not carry moral considerations to one that did) resulted in a different weighting of the evaluations of an AI model's performance, process, and purpose. Complexity in the AI model's process explanation no longer served as a heuristic; complexity in the explanation of the AI model's process was not associated with greater reliance on its output as was the case with the task in Study 1. Instead, the AI model's purpose (operationalised as generosity) - became the new heuristic. Specifically, when the AI model was evaluated as low in generosity (e.g., when it was advising for lower resource allocations), its advice was followed more than when it was evaluated as high in generosity (e.g., when it was advising for higher resource allocations) regardless of whether its performance was rated as high or low (Figure 13). This

finding suggests that the importance placed on the explainability of an AI model's process depends on the moral nature of the decision. In decisions with more salient moral considerations, people prioritise their evaluations of the AI model's purpose over the explainability of its process. In such decisions, teleological explanations ('*Why did the AI make this decision?*') are more likely to increase trust in the AI model's output than mechanical explanations ('*How does the AI work?*'). We further discuss this in the 'Practical Implications and Directions for Future Research' section below.

What underlying mechanisms might explain the findings? First, the results from Study 1 seem to align with the psychological phenomenon of the *disfluency effect* (Alter et al., 2007). According to the disfluency effect, presenting information in a way that makes it more challenging to process can lead individuals to engage in deeper cognitive processing. This deeper processing has been shown to impact judgment and decision-making (Alter & Oppenheimer, 2009), enhance information retention (Diemand-Yauman et al., 2011), and improve learning (Bjork & Bjork, 2011). In the case of AI, by introducing a certain degree of cognitive difficulty or disfluency in the explanation of how the AI model works, through a more detailed and technical explanation, individuals may be prompted to engage more deeply with its explanation. As a result of this deeper processing, they may be willing to place more trust in the AI model's output, perceiving the time spent engaging with its explanation as a proxy for the quality of its algorithmic outcome.

Equally, the preference on more detailed and technical explanations may reflect a manifestation of the well documented phenomenon of the *illusion of competence* which is often created through the use of scientific terminology. Indeed, the phenomenon of the *illusion of competence* have been studied in psychology since the late seventies. Studies, such as those by

Naftulin et al. (1973), have shown that the use of complex, scientific jargon can create an illusion of competence, although a more recent study by Oppenheimer (2006) found that while scientific jargon can affect perceived intelligence, overly complex language can sometimes reduce this perception. In the case of AI, it's possible that the use of technical details in the explanation of the model's process is creating a similar illusion of competence, leading participants to view AI as more competent and as such more trustworthy.

Finally, an interesting finding from Study 2 was that when the AI model was evaluated as low in generosity (e.g., when it was advising for lower resource allocations), its advice was followed more than when it was evaluated as high in generosity (e.g., when it was advising for higher resource allocations). One potential explanation could be that participants may have approached the AI model exhibiting generosity with scepticism. For example, participants may have expected the AI model to be impartial and, therefore, more rather than less conservative in the resource allocation. Participants might have wondered, '*Why is an AI model being generous?*' and ultimately deemed it untrustworthy since its behaviour did not conform to their pre-existing beliefs. Future studies should seek to replicate and further explore the underlying psychological factors behind these observed behaviours. Pre-existing beliefs and expectations influencing the evaluation of AI would align with research with embodied AI such as robots, which has shown that when a robot appearance or behaviour resembles the human, but its capabilities fall behind it creates negative evaluations (De Graaf et al., 2017; Duffy & Joue, 2004; Pandey & Gelin, 2018) .

Limitations

Findings from Study 1 and Study 2 come with limitations. First of all, they come with small effect sizes - η^2 less or equal to 0.01, with an η^2 of 0.01 being considered as small effect

size (Cohen, 2009). While this is undoubtedly a limitation for deriving substantial practical implications, there is a discernible pattern in the experimental data of Studies 1 and 2 that should not be overlooked because of its small effect size. The data revealed an emerging pattern in how AI models are evaluated, suggesting a dynamic interplay between the 3Ps and the nature of the decision. This interplay warrants further exploration in future research, particularly research that seeks to put this interplay to test in real-life settings and across different decision environments, such as financial, medical, and legal decision-making contexts where the use of AI models has become prevalent. Secondly, another limitation of both studies is the fact that it remains open for investigation whether findings will resonate equally across different demographics. Characteristics such as digital literacy and familiarity with technical terminology, or other demographic differences such as i.e., generational differences could influence observed behaviours.

Thirdly, the current set of studies revealed behaviours rather than relying solely on self-report measures. In this regard, it contributes to a deeper understanding of how people trust AI based on their actions, rather than just what they say. As behavioural insights on trust in AI are currently underexplored compared to survey-based insights (Glikson & Woolley, 2020), this study contributes to a more nuanced understanding of trust in AI by uncovering behavioural patterns. However, further studies are needed to uncover the underlying mechanisms behind the observed behaviours. While some plausible mechanisms were discussed here, future research should aim to explore what drives these behaviours.

Finally, in both studies, cues about an AI's performance, process, and purpose were conveyed through written text. Future research should explore the impact of different mediums, such as images or sound, in communicating information about the 3Ps, and investigate how these

mediums affect evaluations of AI models. For instance, what would the pattern of results be if the information was presented with images rather than written descriptions? (i.e., through the use of graphs, especially for conveying information about the AI's performance or for explaining its process). Or what if an AI model could convey information about its performance, process, and purpose through its 'own' voice? Voice is a key design feature that has currently captured the attention of the AI research community, particularly in the development of Large Language Models equipped with voice (i.e., FunAudionLLM, mini-Omni), as it represents the next step toward enabling real-time conversational interactions with LLMs. Psychological research can provide insights into how auditory, human like cues influence users' perceptions and evaluations of AI models.

Practical Implications and Directions for Future Research

The current set of studies highlights the need for AI researchers, designers and deployers alike to adapt explanations of AI models to the decision context, particularly taking into consideration the moral considerations of the decisions at hand. In decision contexts where the moral stakes of decisions are prominent or high, people are likely to prioritise information more about the 'why' rather than the 'how' of an AI model. The focus of the '*why'* rather than the '*how'* is often referred to as teleological explanation and refers to explaining things in terms of their purpose or reason behind rather than the mechanism or internal processes by which they happen(Dennett, 1989). In the context of AI, this kind of explanation is often described as *goal-based* or *purpose-driven* (Miller, 2019). It answers the question: *'Why did the AI make this decision'* as opposed to '*How does the AI work?'*. For example, a teleological explanation for an AI model helping radiologists diagnose and treat a skin disease could be '*The AI recommends this treatment because it's associated with a higher survival rate for patients with similar*

medical histories' as opposed to a more mechanistic explanation of '*The AI uses a neural network trained on a dataset of 1 million patient records, analysing factors like age, medical history, and genetic data to predict which treatment is most likely to succeed'*. Future research should look at combining mechanistic and teleological explanations -the latter often appearing to be missing in today's AI models' explanations- to tailor transparency to a decision's moral dimension. This is particularly important in cases such as i.e., medical decision making or autonomous vehicles where the moral stakes of the decisions are high.

Moreover, it is noteworthy that when participants self-reported their attitudes and level of trust in the AI models, they expressed more positive attitudes and greater willingness to trust a high-performing over a low-performing model, a benevolent over a malevolent model, and an AI model with a simple rather than a complex explanation of its process. However, having a behavioural task in both studies allowed capturing people's behaviour alongside with these self-report measures. This proved to be insightful as it revealed that, in reality, and contrary to what people self- report when explicitly asked, it is not only high performance, or simple explanations or instilling a good moral character in an AI model that matters. The behavioural data painted a different picture; performance, process and purpose all play a role in how people evaluate AI models, with the moral nature of the decision at hand shaping which of these determinants takes precedence. This is a valuable finding from a methodological view, as it underlines the need for behavioural data in the study of decision making with AI.

Finally, the current set of studies revealed a positive reception of complexity in an AI model's explanation under specific conditions. E.g., when the decision task had no moral considerations, reliance on the AI model's output increased with the complexity in the explanation of its process when its performance was rated as low, or its purpose considered good

(Study 1). However, this finding needs to be further explored in future research that will seek to determine the optimal level of technical detail in an AI model explanation as moving away from '*black box'* AI is crucial for the AI community, however, too much complexity can equally overwhelm users, leading to negative experiences. This finding is also likely to be influenced by individual characteristics such as familiarity with AI models or level of experience with AI outputs in a specific decision context, which are worth exploring in future studies that will seek to contribute to a more nuanced understanding of how people evaluate AI models.

Conclusion

We explored how people form evaluations of AI models' outputs using the three determinants of trust in automation (performance, process, and purpose) and by using two different decision tasks that varied in terms of moral considerations (a visual estimation task and a resource allocation task). Results suggest that people factor in all three determinants (performance, process, and purpose) when evaluating an AI model's outputs with the nature of the decision task changing the importance put to each. The findings contribute to efforts towards a nuanced understanding of how people form evaluations of disembodied AI such as AI models.

Chapter 4 Decision Making with AI

Introduction

In this chapter, we explore how people interact with AI-generated outputs, particularly focusing on recommendations and advice. As AI increasingly influences daily life—offering guidance on entertainment options (Jesse & Jannach, 2021), informing health decisions (Obermeyer et al., 2019), shaping tastes and preferences (Yeomans et al., 2019), and even affecting romantic partner choices (Dellaert et al., 2020), understanding how people engage with AI outputs such as advice or recommendation is more relevant and intriguing than ever. If people do listen to them or at least listen to them more than they do to humans, then these outputs hold the potential to change people's lives. In addition, examining how people interact with AI outputs such as recommendations and advice not only enriches our understanding of the human perspective on AI, but also builds upon earlier insights into how people perceive (Chapter 2) and evaluate AI (Chapter 3) by shedding light on differences between self-report beliefs and actual behaviours toward AI outputs.

What We Know So Far About How People Respond to AI Advice?

Reactions to AI advice can vary widely, ranging from appreciation to aversion or indifference. Previous research has predominantly examined how people respond to AI-generated advice in comparison to advice from humans. This body of work suggests three possible outcomes: People may prefer AI advice *more* than human advice, a phenomenon termed as algorithmic appreciation (Logg et al., 2019). Equally, people may prefer AI advice *less* than human advice, something often termed in the relevant literature as algorithm aversion (Burton et al., 2020). Or, they may be indifferent to the source of advice, showing no preference for one over the other (Leib et al., 2024).

Two relatively recent comprehensive reviews on algorithmic aversion (Burton et al., 2020; Jussupow et al., 2020) highlight the plurality of factors at play as the main reason why people sometimes may be averse while other times appreciative towards AI advice. Specifically, they highlight factors having to do with the nature of the decision task (i.e., how subjective or objective people perceive the decision task at hand to be), the characteristics of the decision-makers (i.e., their level of familiarity or previous experiences with AI advisory systems), the unique aspects of each decision-making environment (i.e., the saliency of moral considerations in some decision-making environments more than others, such as is the case in medical decision contexts or legal decision contexts), as well as the complex interactions among the above factors.

Empirical evidence suggesting algorithmic appreciation comes from a series of experiments where, for numerical decision tasks (i.e., estimating a person's weight), participants relied more on advice labelled as coming from an algorithm than identical advice labelled as coming from humans (Logg et al., 2019). The same series of studies also showed that experts discounted algorithmic advice more than non-experts, although advice discounting was less when the advice was labelled as coming from an algorithm than when labelled as coming from a human. Discounting of advice by experts is consistent with existing literature on *egocentric advice discounting* (Yaniv & Kleinberger, 2000) as well as *overconfidence* (Johnson & Fowler, 2011; Moore & Healy, 2008; Moore et al., 2015; Russo & Schoemaker, 1992), which have both repeatedly shown that experts are more likely to discount advice compared to non-experts. The above studies showed that humans do the same with AI; when they consider themselves experts, they discount advice from AI as they do from humans, albeit to a lesser degree.

Furthermore, a series of studies by Castelo et al. (2019) highlights the role of the nature of the decision task, as participants reported greater willingness to rely on advice from an AI

advisor than a human advisor for decisions they perceived as more objective in nature (such as predicting a student's performance) than subjective (such as selecting a movie to watch). Also, for more subjective decisions such as decisions related to personal preferences, as is the case with books, movies, or jokes, participants were also more likely to seek advice from friends over recommender systems (Yeomans et al., 2019).

On the other hand, empirical evidence suggesting algorithmic aversion comes from a series of experiments by Dietvorst et al. (2015) where, after seeing an algorithm err, participant relied on humans for forecasting student performance, even when doing so resulted in suboptimal forecasts. Nevertheless, allowing participants to slightly modify the output of an algorithm made them more tolerant of errors and more likely to choose an algorithm for subsequent forecasts (Dietvorst et al., 2018). And demonstrating an AI advisor's ability to learn was shown to offset negative effects of familiarity and previous negative experience with its errors (Berger et al., 2020). Also, people report being averse to AI making decisions in moral domains where human lives are at stake, such as medical settings, parole sentences, military decision-making and self-driving cars (Bigman & Gray, 2018). And they can equally be averse to AI offering advice in these decision domains. For instance in medical decision-making, when recommendations came from an algorithm in a study by Longoni et al. (2019), people tended to trust it less than a human doctor's recommendations. And in the domain of employee selection and hiring decisions, where there also exist ethical considerations, Diab et al. (2011) found that participants thought of human interviewers' advice as being more useful, professional, fair, personal, flexible, and precise than AI advice.

Coupled with empirical evidence on both algorithmic appreciation and aversion, there is also evidence showing that people treat advice similarly, regardless of whether it comes from a

human or an AI agent. Empirical evidence suggesting that transparency about the AI source of advice does not influence people's subsequent uptake comes from a series of studies by Leib et al. (2024). In these studies, participants were exposed to advice generated by both a natural language processing algorithm (GPT-J algorithm) and a human equivalent. The findings showed that dishonesty-promoting advice increased dishonest behaviour, while honesty-promoting advice did not enhance honesty, and that pattern of results was the same regardless of whether the advice came from a human or an AI source.

Alongside the above empirical evidence which in their majority come from lab-based studies, we also observe, in our everyday lives, that people are increasingly turning to AI for guidance. This is also evidenced by the widespread popularity of AIs such as Alexa (Chalhoub & Flechais, 2020) or advanced Large Language Models (LLMs) (Radford et al., 2019) that can offer voiced or written advice. These AIs offer their expertise on a broad spectrum of topics, provided the user formulates the appropriate prompt. Moreover, seeing that AI advice comes with no or low cost, people may nowadays choose AI advice for its affordability and accessibility (an example of this is the AI-enabled financial app Wealthify.com which has a substantial customer base in the UK), even when expert human advice is available.

It is also well-supported that people follow social norms when making decisions (Bicchieri, 2016; Köbis et al., 2022; Schultz et al., 2018). Social norm theory categorises these norms into two types: proscriptive norms, which indicate what people believe they *ought* to be doing, and descriptive norms, which outline what people are *actually* doing. Empirical studies (Bobek et al., 2013; Cialdini et al., 1991) show that people are often more influenced by their perceptions of what others are doing or think they are doing – descriptive norms – rather than norms that refer to what they ought to be doing - proscriptive norms. When it comes to social

norms and AI advice, the hypothesis can go both ways; if people perceive AI advice as a better representation of collective beliefs and behaviours—since they are based on algorithms trained on vast amounts of human beliefs and behaviours—they may view such advice as a stronger indicator of what other people are doing compared to human advice coming from one or a few other people (i.e., human consensus advice). If this is the case, and based on social norms theory, we might expect people to be more likely to follow AI advice than human advice. However, the opposite might hold true too; people might see human advice as carrying a stronger cue for social norms, given that it originates from humans rather than algorithms. Consequently, they may be more likely to follow human advice than AI advice to align with social norms. As noted by Leib et al. (2024), the hypothesis could go either way, highlighting the need for further behavioural studies to clarify how individuals behave in response to AI advice.

Overall, the existing literature indicates that there is no consistent response to advice based solely on the identity of the advisor, whether it comes from an AI or a human. There is no clear trend of appreciation for AI-generated advice, aversion toward it, or significance placed on the transparency regarding the AI source. Instead, trust in advice, which includes both cognitive trust (based on a rational assessment of the AI, its advice and situational factors) and emotional trust (influenced by emotions or mood) appears to be heavily content-dependent (Glikson & Woolley, 2020).

Research Question and Overview of Studies

Given the context-dependent nature of AI advice uptake and, rather than taking a definitive stance (algorithmic appreciation, aversion, indifference), we decided to perform our own exploration of people's reaction to AI advice starting by focusing on one key factor identified in the existing literature: the perceived objectivity of the decision at hand (Studies 1

and 2). Previous studies suggest that the more objective a decision is perceived to be, the more likely people are to prefer AI over human advice(Castelo et al., 2019; Logg et al., 2019). We sought to examine this within our own experimental settings, adopting the same design as Castelo et al. (2019) in their Study 1 but incorporating an updated set of decision tasks. This approach allowed us to both replicate and extend the original study. Additionally, recognising the influence of social norms, we conducted two additional studies where we measured AI advice uptake in preference-based decisions where AI advice was compared to advice coming from other people (Studies 3 and 4). These studies examined advice uptake when validating advice was labelled as coming from an AI versus other people in preference-based decisions. We focused on advice that validates people's preferences, as this type of advice is common in AI as recommendations and advice are based on individuals' past behaviours, as these are reflected in their data. We therefore sought to address the following research question: How do people respond to AI advice that validates them? Are they more, less, or equally likely to listen to AI validating them as to other people?

Study 1

The aim of this study was to create a list of everyday decisions, categorised based on how subjective or objective people perceive them to be. This approach follows the example of previous research, which classified decisions according to their perceived level of subjectivity (Castelo et al., 2019). A total of 50 participants were asked to evaluate 49 decision tasks based on how subjective or objective they perceived them to be. The curated list of objective and subjective decisions of Study 1 was then used in Study 2.

Methods

Participants

UK participants, averaged 25-34 years of age, 64% female, 36% male, were recruited through the Prolific experimental subject pool for compensation. Since the purpose of the study was to generate information rather than test a hypothesis using inferential statistics, a rule of thumb (Baumol & Quandt, 1964) was used for deciding on the sample size of 50 participants. Participants were paid based on an hourly rate (£7.50/hour) for the time spent in the study. The exact materials and data for Study 1 are available in the Open Science Framework at https://osf.io/exbpt.

Materials and Procedure

At the start of the study, participants answered a set of demographic questions, and an attentional check question designed to remind them to read the questions carefully. If they answered the attentional check question incorrectly, though they were not disqualified from continuing, they were reminded to carefully read the questions before answering. Then, they rated each of 49 decision tasks on how subjective versus objective they found it to be, using a continuous scale from 0 ('Very Objective') to 100 ('Very Subjective'), and reported their level of confidence with their rating on a scale from 0 ('Not at all') to 100 ('Extremely'). The 49 decision tasks were randomly presented and before they started appearing, participants received the following brief explanation of the difference between *subjective* and *objective* decisions:

> 'In *broad terms, think of a SUBJECTIVE decision as one that is heavily influenced by one's personal feelings, perspectives, and interests and as such one to which, in principle, there is no right or wrong answer. On the contrary, an OBJECTIVE decision is usually viewed as relying solely on facts, data, and analysis and as such one to which, in principle, there is one verifiable and right answer*.'

Results and Discussion

The average rating for each decision task was determined using the following approach: If the mean rating was above 50 (the mid-point of the scale), a subjectivity score of x out of 50 was assigned. For instance, a task with a mean rating of 80 received a subjectivity score of 30 out of 50. Conversely, if the mean rating was below 50, an objectivity score of y out of 50 was assigned. For example, a task with a mean rating of 10 received an objectivity score of 40 out of 50. Additionally, the confidence ratings given by participants for each decision task, using a continuous scale from 0 ('Not at all confident') to 100 ('Extremely confident') were averaged. The group of subjective decision tasks exhibited a higher confidence score (M=82.06) compared to the confidence score (M=72.70) of the group of objective decision tasks, t (47) = -6.74, d = -0.194, p<0.001. This suggested that participants felt more confident when grouping a decision as subjective than objective.

The list of decision tasks with their corresponding subjectivity ratings are included in Table 11. Interestingly, two decision tasks—'*How much to save for retirement*' and '*Which retirement plan to invest in'*—had different subjectivity scores despite their similar meanings. The first got a subjective score of 53.96 (e.g., >50), while the second was rated as objective (score =44.02, e.g., <50). This discrepancy is likely to be a product of the framing effect (Tversky & Kahneman, 1981). E.g., '*choosing a retirement plan to invest in*' might be perceived as involving more careful benefits vs costs considerations, while '*deciding how much to save for retirement'* may be perceived as a decision made more on the basis of personal preferences. The framing effect is something that has also received attention in the realms of AI advice uptake. For instance, merely using the term 'expert system' instead of 'computer' to describe algorithmic advice was found to significantly increased trust in its recommendations compared to human advice (Hou & Jung, 2021).

Tables 11

List of Decision Tasks grouped by their Perceived Subjectivity (Study 1). Decision tasks are listed in decreasing order of perceived subjectivity

	Decision Task Subjectivity Rating	Decision Task Score	Average Confidence per Decision Task	Decision Task Category
Which song to listen to	90.80	40.80	90.60	Subjective
What movie to watch	88.24	38.24	85.38	Subjective
What to wear on a night out	87.04	37.04	87.96	Subjective
Who to go out on a date with	86.20	36.20	86.08	Subjective
Which book to buy	83.72	33.72	86.32	Subjective
What to cook for dinner	83.36	33.36	84.44	Subjective
Which restaurant to go to	83.30	33.30	84.66	Subjective
What to get a friend on their birthday	82.18	32.18	85.04	Subjective
Which joke to use in a zoom call at work	80.32	30.32	84.88	Subjective
Where to go on holiday	77.64	27.64	84.38	Subjective
Which career to follow	76.28	26.28	81.00	Subjective
Making decisions with regards to the planning of a birthday party	76.06	26.06	81.42	Subjective
Whether to end a relationship	75.12	25.12	85.42	Subjective
Predicting someone's personality	74.64	24.64	77.92	Subjective
Deciding who to vote for	68.74	18.74	76.38	Subjective
Which job to apply for	68.62	18.62	77.16	Subjective
Which neighbourhood to move to	66.18	16.18	84.20	Subjective
Which job offer to accept	64.04	14.04	76.94	Subjective
Solving ethical problem	61.84	11.84	72.20	Subjective
Making political decisions	58.76	8.76	76.38	Subjective
Scheduling one's week	57.44	7.44	79.42	Subjective
How much to save for retirement	53.96	3.96	77.12	Subjective
How to plan your monthly budget	49.80	0.20	75.48	Objective
Predicting parole violation	48.32	1.68	69.74	Objective
Predicting recidivism	48.22	1.78	71.90	Objective
Predicting a student's performance	47.32	2.68	74.02	Objective
Which credit card to apply for	46.20	3.80	72.08	Objective
Deciding who should be given a fellowship	45.62	4.38	62.10	Objective
Which retirement plan to invest in	44.02	5.98	75.36	Objective
Decide which stocks to buy	43.96	6.04	68.26	Objective
Predicting an employee's performance	43.34	6.66	70.88	Objective
Administering justice and rehabilitation	42.44	7.56	66.72	Objective
Predicting an election	41.78	8.22	64.14	Objective
Predicting a stock price	41.68	8.32	63.18	Objective
Hiring a new employee	41.30	8.70	77.16	Objective
Deciding on the performance of an employee	36.96	13.04	78.12	Objective
Firing an employee	35.76	14.24	72.50	Objective
Deciding on the strategic plan for a company	33.10	16.90	70.82	Objective
Performing fact checking on news feeds to decide what qualifies as fake news	32.54	17.46	75.00	Objective
Deciding the bonus payment of an employee	32.20	17.80	74.26	Objective
Which credit card to apply for	31.82	18.18	77.00	Objective
Which medical treatment to undergo	31.22	18.78	75.44	Objective
Deciding on a treatment of a disease	29.00	21.00	75.20	Objective
Deciding on the sentence in a legal case	28.72	21.28	71.58	Objective
Predicting the weather	27.10	22.90	75.92	Objective
Deciding whether a customer qualifies for a specific insurance plan	22.12	27.88	75.98	Objective
Piloting a plane safely to protect its passengers	21.28	28.72	75.30	Objective
Deciding on the shortest route between two points on a map	20.34	29.66	84.62	Objective
Diagnosing a disease	15.80	34.20	70.20	Objective

Study 2

In Study 2, a new group of participants was asked to indicate their willingness to rely on advice coming from an AI algorithm versus a human (or a well-qualified human) in 14 decision-making scenarios than varied in terms of their perceived subjectivity. The decision scenarios were taken from Study 1. The purpose of Study 2 was to investigate whether people are willing to rely on AI rather than humans (or well-qualified humans) more for objective than subjective decisions since prior research has indicated that trust in AI advice increase with the perceived objectivity of a decision (Castelo et al., 2019; Logg et al., 2019).

Methods

Participants

430 UK participants, averaged 25-34 years of age, 78% female, 20% male, and 2 % non-binary, were recruited through the Prolific experimental subject pool for compensation. Participants were paid based on an hourly rate (£7.50 /hour) for the time spent in the study. The sample size was determined a priori with a goal of obtaining 0.80 power to detect a medium effect size of d= 0.35 (e.g., f= 0.17) taken from previous research (see Logg et al. 2019, study 6) at the standard 0.05 alpha error probability. The software program G*Power was used to conduct an a-priori power analysis. The suggested minimum sample size was 337 participants and a total of 430 participants were recruited to account for potential exclusions due to failures in attentional checks and completions outside the window of accepted durations. The exact materials and data for Study 1 are available in the Open Science Framework at https://osf.io/tjgvd .

Materials and Procedure

Participants began the experiment by answering demographic questions, and an attentional check question designed to remind them to read the questions carefully. Before

proceeding with the remainder of the questions, they were provided with the following brief definition of an AI algorithm:

> *'A brief note before you start:*
>
> *In the rest of the survey, you will see Artificial Intelligence algorithm (AI algorithm) being mentioned quite often. In broad terms, you can think of an Artificial Intelligence algorithm (AI algorithm) as a set of rules given to a computer program to enable it to learn on its own by finding useful patterns in the data and coming up with a process to make a decision. These AI algorithms are nowadays applied in various fields such as finance, marketing, business analytics, agriculture, healthcare, etc. (source: EDUCAB, Artificial Intelligence tutorial)'.*

Participants were then asked to put themselves in the shoes of the decision maker in 14 hypothetical decision scenarios, seven subjective and seven objective decision scenarios, selected from Study 1 among the top 10 most highly subjective and the top 10 most highly objective decisions, respectively (see Table 12 for the list of the 14 decision scenarios used in Study 2). In each decision scenario, they were asked the following question: *'In the above decision scenario, when you make your decision, you can get advice from an AI algorithm or a human. Would you rely more on advice from an AI algorithm or a (well-qualified) human?'* using a continuous scale from 0 ('*Rely more heavily on a (well-qualified) human*') to 100 ('*Rely more heavily on an AI algorithm*.' The decisions scenarios were counterbalanced.

The study employed a 2x2 mixed design, examining the factors of subjectivity of the decision (subjective vs. objective) and advisor (human vs. well-qualified human). Subjectivity served as a within-subjects factor, while expertise of the human advisor was a between-subjects factor. The dependent variable was the participants' reported willingness to rely more on advice from an AI algorithm than a human, the latter being labelled for half of the participants as advice

coming from a ‘human’ and for the other half as advice coming from ‘a well-qualified human’ (expertise was a between-subjects variable).

Data Analysis Strategy

Data were analysed using R statistical software (https://www.r-project.org/). First, data were averaged across objective and subjective decisions per participant. This resulted in two values per participant: one representing the average proportion where the participant favoured one advisor over the other in the objective decisions, and another representing the average proportion where the participant favoured one advisor over the other in the subjective decisions. A mixed ANOVA was employed for the analysis, with the subjectivity of the decision (subjective vs. objective) as a within-subjects factor variable, and human expertise (human vs well-qualified human) as a between-subjects factor. We also calculated the simple main effect of perceived subjectivity across the human vs well -qualified human factor variable using ‘emmeans’. We controlled for multiple testing in this analysis by adjusting the six p-values using the Bonferroni-Holm adjustment. In all the statistical inference tests, a *p*-value of less than 0.01was used as the threshold for statistical significance in any reported results, following the precedent established in similar studies (Castelo et al., 2019).

Tables 12

List of the 14 decision scenarios used in Study 2

Decision Scenarios	Category
Imagine that you are thinking of a career change, and you have come up with three alternative career options. You need to decide which one to choose.	Subjective
Imagine you are a medical patient and you have received test results for treating tendonitis in your forearm. You need to decide between three alternative treatments.	Objective
Imagine you are a judge in a legal case. To decide on the sentence, you need to predict the offender's risk of reoffending and factor that in your final decision.	Objective
Imagine that you want to know what the weather will be like tomorrow so that you can decide what to wear.	Objective

Imagine that you are in a city you have never been before, and you want to get to the nearest train station.	Objective
Imagine that you have invited friends over to your place for dinner and you want to prepare a playlist of songs to listen to. You need to decide what's the best songs to include in your playlist.	Subjective
Imagine you find three books online that look interesting. You need to decide which book to buy.	Subjective
Imagine that you want to buy new clothes for a night out. You search online and you find three different outfits that you like and that cost approximately the same. You need to decide which one to buy.	Subjective
Imagine you are thinking of going out for dinner. You search online and you end up with a couple of restaurants that look appealing and need to decide which one to go to.	Subjective
Imagine you are looking for a private health insurance plan. You search online and you find three appealing options with similar premium cost. You need to decide which one to apply for.	Objective
Imagine that you are thinking of going away on holiday for a few days. After some initial research, you come up with a couple of places that are worth visiting and need to decide which one to choose as your next holiday destination.	Subjective
Imagine you are facilitating a zoom meeting at work, and you are thinking of possible jokes to use for breaking the ice at the start of the meeting. You need to decide which joke to include.	Subjective
Imagine you are thinking of getting a new credit card. You go on to your bank's website and you find three credit cards that look appealing. You need to decide which you would like to apply for.	Objective
Imagine you are the HR manager of a company, and you need to decide how to allocate the annual bonus payment to selective employees.	Objective

Results

Manipulation checks

Although the 14 decision scenarios in Study 2 were taken from Study 1, where they were rated on their perceived subjectivity by a different sample of participants, we also tested the subjectivity manipulation with Study 2's participants. At the end of the study, participants rated each of the 14 decisions on a scale from 0 (*'Very Objective'*) to 100 (*'Very Subjective'*) after receiving a brief description of subjective versus objective decisions, identical to the one used in

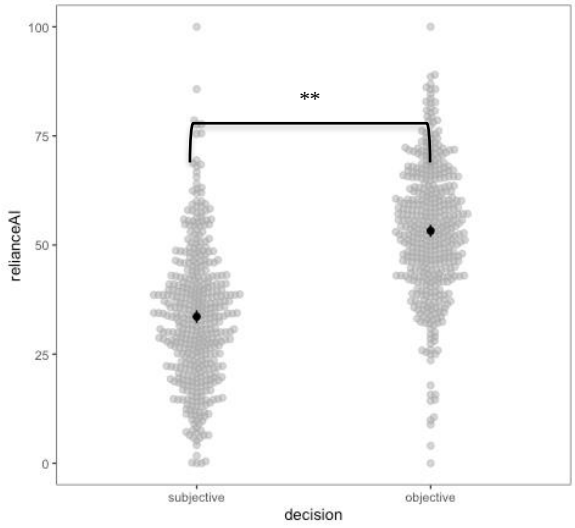
Study 1. We tested the manipulation by performing a *t*-test to determine if there was a significant difference between the means of subjective and objective decisions. The t-test revealed a significant difference, indicating the manipulation was successful (*t (*858) = -34.89, *d*= -2.38, *p*<.001).

Reliance on AI Advice

The ANOVA revealed a significant main effect of *perceived subjectivity*, *F* (1, 428) = 404.10, η^2= 0.30, *p* <.001 (Figure 14), suggesting that for decisions that are perceived as more objective than subjective, participants were more willing to rely on advice from an AI algorithm than humans (see Table 13 for means and standard deviations).

Figure 14

Main Effect of Perceived Subjectivity on Willingness to Rely on AI Advice (Study 2)



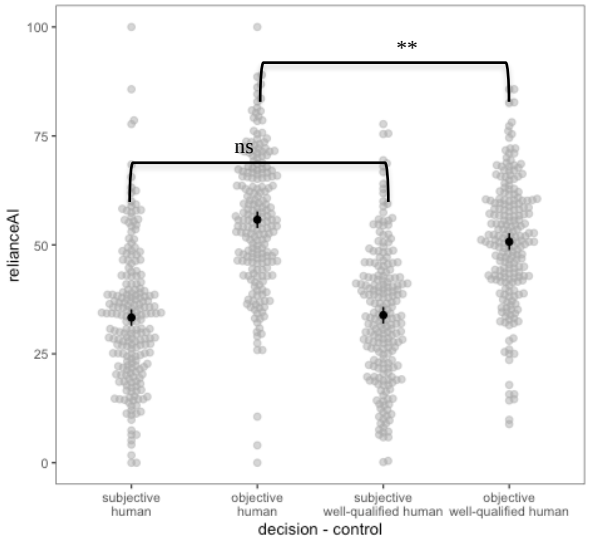
Error bars represent 95% confidence intervals, ** *p*< 0.001

There was also a significant *subjectivity* X *human expertise* interaction, *F* (1,428) = 8.17, η^2= 0.01 *p* < .01. A simple effect analysis showed that for objective decisions, the difference between reliance on a human (*M* = 55.8, *SD* =1.01) and a well-qualified human (*M* = 50.7, *SD* = 1.01) was significant (*t* (428) = - 3.54, *p*<.001). For subjective decisions, there was no significant

difference (t (428) = 0.37, p=0.7) When advice from an AI was pitched against advice from a human, the difference in reliance between objective decisions (M = 55.8, SD =1.01) and subjective decisions (M = 33.3, SD =1.05) was significant (t (428) = 16.24, p<.001) and, when advice from an AI was pitched against advice from a well-qualified human, the difference in reliance between objective decisions (M = 50.7, SD =1.01) and subjective decisions (M = 33.9, SD =1.05) was also significant (t (428) = 12.19, p<.001) (Figure 15).

Figure 15

*Willingness to rely more on an AI vs a human advisor: Perceived Subjectivity * Expertise of Human Advisor (Study 2)*



Error bars represent 95% confidence intervals, ** p< 0.001; ns: not statistically significant

Table 13

Willingness to rely more on an AI algorithm vs a human advisor: Means and Standard Deviations (Study 2)

	Willingness to Rely on Advice from an AI algorithm vs (well-qualified) human	
1	Type of Decision: Objective, advisor: human	55.8 (1.05)
2	Type of Decision: Objective, advisor: well-qualified human	50.7 (1.05)
3	Type of Decision: Subjective, advisor: human	33.3 (*1.05*)

4	Type of Decision: Subjective, advisor: well-qualified human	33.9 (1.05)

Discussion

Study 2 revealed that willingness to rely more on advice from an AI algorithm than a human increase with perceived objectivity of the decision at hand. The results also indicated that for objective decisions, the human expertise plays a significant role; people are less willing to follow AI advice when alternative human advice comes from someone they regard as an expert, compared to when it comes from a layperson. These findings align with previous research on the role of task subjectivity (Castelo et al., 2019) and the perceived human expertise (Logg et al., 2019) on AI advice uptake. In the subsequent two studies, Study 3, and Study 4, we concentrated solely on subjective, preference-related decisions.

Study 3

Study 3 sought to examine how people behave in interaction with AI output that comes in the form of recommendation on a preference-related topic, such as when choosing one's preferred coffee, and when presented alongside human recommendation. Both types of recommendations (from an AI algorithm and from other people with similar coffee preferences) were shown to participants simultaneously. Study 3 employed the judge–advisor system (JAS) paradigm (Sniezek & Buckley, 1995)**,** a paradigm commonly used to measure reliance on advice by computing how much participants revise their initial choice in response to external advice (Haran & Shalvi, 2020b; Yaniv, 2004a, 2004b; Yaniv & Choshen-Hillel, 2012; Yaniv & Kleinberger, 2000). This paradigm allows us to determine whether a final choice is influenced by advice, as it also captures the initial choice e.g., the choice before receiving advice.

Methods

Participants

UK participants averaged 18-24 years of age, 86% female, 14% male, were recruited through the UCL Psychology Subject Pool (SONA). Participants received 0.5 course credit and a coffee bag randomly selected from their actual choices to incentivise them throughout the study. The coffee blend was shipped to the UK address they provided. The sample size was determined a priori with a goal of obtaining 0.80 power to detect a medium effect size of $f^2 = 0.15$ (Cohen, 1988) at the standard 0.05 alpha error probability. The suggested minimum sample size was 98, and 110 participants were recruited to account for potential exclusions due to failures in attentional checks and completions outside the window of accepted durations. E.g., participants that took more than 1 hour to finish the study were automatically excluded (3 were excluded) leading to a sample size of 107 participants in total. The study was preregistered on AsPredicted.org (https://aspredicted.org/see_one.php).

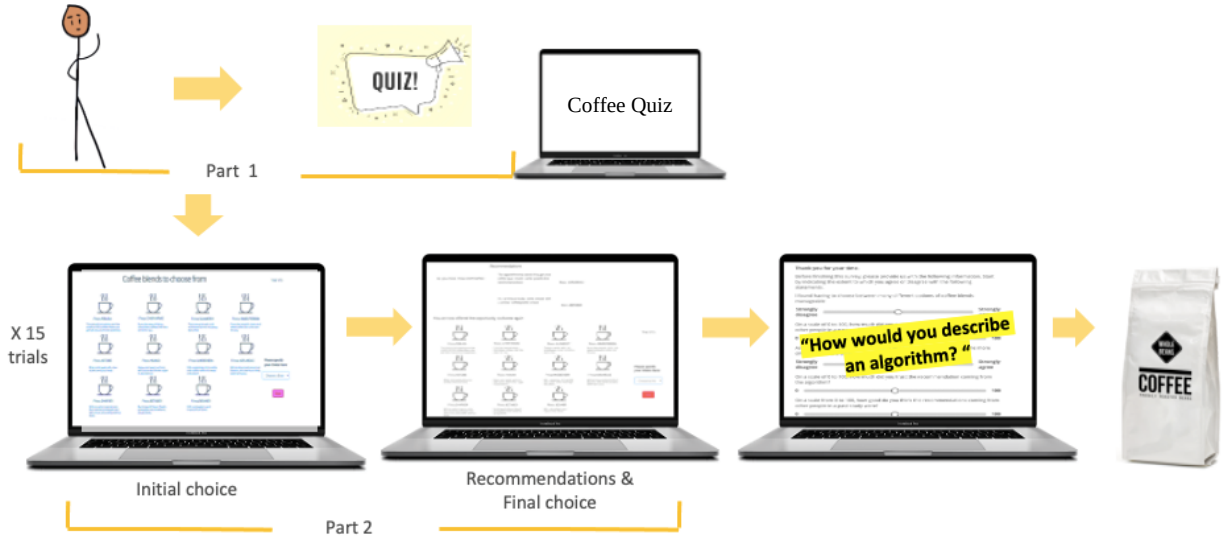
Materials and Procedure

The study was conducted online using the Gorilla survey platform. After consenting, participants completed two parts. In Part 1, they completed a coffee quiz to produce their coffee profile. Then, in Part 2, they chose their favourite coffee blends in 15 trials, with the number of available choices per trial varying randomly between 6 and 11. In each trial, they made an initial choice and then they were directed to a second screen where they were presented with two recommendations: one coming from an AI algorithm that used their coffee profile as an input to match them with a coffee recommendation, and another one coming from other similar people, e.g., people with similar coffee quiz results.

After seeing both recommendations, participants were asked to make a final choice. All participants completed both parts of the study. Finally, participants reported their age and gender

and were also asked to provide an UK address for the shipment of the bonus coffee bag. A graphical representation of the study's procedure is included in Figure 16.

Figure 16
Graphical Representation of Procedure followed in Study 3



Each participant completed 15 trials. In each trial participants were presented with different choice sets varying from 6 to 11 coffee varieties. While we aimed for more than 15 trials per participant to obtain participant's choice patterns, average duration from the pilot run of the study indicated that 15 trials were enough for keeping the study within a manageable duration (e.g., for keeping the study's duration within max 40 mins). Moreover, the advice labelled as coming from an 'AI algorithm' and 'other similar people' was programmed so that the likelihood of them differing from the participant's initial choice and the likelihood of them differing from each other followed predefined probabilities- e.g., Scenario 1: Participant choice = AI recommendation = Human recommendation (likelihood of appearing 5%), Scenario 2: Participant choice = AI recommendation (10%), Scenario 3: Participant choice = Human recommendation (likelihood of appearing 10%), Scenario 4: Participant choice ≠ AI recommendation ≠ Human recommendation (likelihood of appearing 75%), Scenario 5:

Participant choice ≠ (AI recommendation = Human recommendation) (likelihood of appearing 5%). Also, the names and descriptions of the coffee varieties used in the study were fabricated to prevent any familiarity with existing brands that could potentially introduced a confounding factor, and they were unique to each trial.

The design employed was a 2 (AI recommendation: Matches vs Does not Matches participant's initial choice) X 2 (human recommendation: Matches vs Does not Matches participant's initial choice) within-subject variables design. The dependent variable was the 'Choice Update' (Yes/No), a dichotomous categorical variable.

Data Analysis Strategy

Data were analysed using IBM SPSS Statistics (Version 29) predictive analytics software. To analyse repeated measurements with a categorical dependent variable a Generalized Estimating Equations (GEE) logistic regression model was used. Upon confirming that the assumptions for a GEE model were met, the GEE model was used to fit the data.

Results

AI advice had a statistically significant effect on final choice. The *human advice* and the AI *advice* X *human advice* interaction were not significant (see Table 15 for the Wald Chi Square tests of model effects). The post-hoc test comparing choice update between *AI advice - validates* and *AI advice - does not validate* revealed that the odds of updating one's initial choice dropped by 0.41points, when the recommendation from the AI algorithm validated participants (e.g., matched the participant's initial choice), e.g., a 41% decrease in the probability of choice update (X^2 (1,107) =28.717, p<.001). These results show that when AI advice was validating participants' initial decisions, the likelihood of changing those decisions was reduced and

suggest that the advice had a validation effect on participant's final choices when it was coming from AI but not from humans. Human advice did not have a validation effect (p=0.49).

Table 14

Test of Model Effects on Choice Update (Study 3)

	Wald Chi-Square	*df*	*p*
AI_validates	13.864	1	<.001
human_validates	3.884	1	.049
AI_validates participant * human_validates	0.644	1	.422
age	7.633	1	.006
gender	0.062	1	.803

To further test that there was indeed no validation effect when the human validated participants (p = 0.49) e.g., to test that there was indeed no statistically significant difference between *human recommendation -validates* and *human recommendation - does not validate*, a Bayesian inference test for binomial proportions was used. The test examined whether the data follows the null distribution model Beta (2,2), which assumes an equal probability of choice updating (50/50) under the 'human_validates: yes' and 'human_validates: no' conditions, or an alternative distribution model Beta (5, 2) whereby there is a higher probability of choice update in the 'human_validates: yes' condition. The estimated Bayes factor was 3.040, which exceeds 1. This provides additional evidence for the null hypothesis, as it suggests that the observed data is approximately 3 times more likely under the null model than under the alternative hypothesis.

Participant's Definition of the AI Algorithm. In this study, we chose not to provide participants with a description of AI. This approach was intended to capture participants' natural responses to the term '*AI algorithm*' and to conduct a thematic analysis on their interpretations.

As such, an open-ended question at the end of the study was included: '*How would you define the AI algorithm used in the study? Please feel free to give a definition in your own words*.' Participants' responses were coded using ATLAS.ti Mac (version 23.2.1) (https://atlasti.com) and categorised into four broad themes, as presented in Table 15 below.

Table 15
Participants' definitions of an AI Algorithm (Study 3)

		AI Algorithm defined as a ...	Example
Category 1	43 %	system/ computer/ tool/program/ procedure	*'It is a programme that outputs information according to the data that it is given.'*
Category 2	20 %	mathematical relationship/ formula/equation	*'A mathematical equation used to make predictions.'*
Category 3	18 %	calculations/ sequence of steps/set of instructions	*'A process where calculations are made through previous data and information to produce a rule or a set path.'*
Category 4	17%	other (specific to the experiment)	*'It used a blend of my previous choices on trials and my preferences which I had input at the very beginning'*

Discussion

Overall, in Study 3 we compared AI and human advice for a decision very much based on personal preferences, such as choosing one's preferred coffee. Our results showed that when it comes to receiving advice on coffee preferences, AI is more effective validating people's decisions that other people. In Study 4, we explored more this finding by broadening the range of preference-related decisions under evaluation from one to twelve covering a variety of preferences.

Study 4

Study 4 sought to further examine how people behave in interaction with AI advice compared to human advice on preference-related topics. Like Study 3, it employed the judge–advisor system (JAS) paradigm (Sniezek & Buckley, 1995) which measures whether participants' initial opinions were swayed or validated after advice (see Study 3 for more details on JAS).

To complement Study 3 and account for individual differences in participants' AI knowledge, an AI literacy index was also introduced in Study 4. This index comprised 12 multiple-choice questions, each with three options and one correct answer, designed to assess fundamental AI knowledge and its various applications. For example, questions included: '*What is an AI algorithm?*' with answer choices (a) '*a set of hardware components that enable computer programs to run*,' (b) *'a set of instructions that can be implemented to perform a specific task*,' and (c) *'a specific type of coding language.'* Other questions asked were of the following type: *'When someone selects a show recommended by Netflix, are they engaging with AI?'* with answer choices *'Yes*,' '*No*,' and '*I am not sure*.' (A complete list of the questions forming the AI literacy index is provided in the Appendix 2, Table 16). The AI literacy index scores ranged from 0 (no correct answers) to 1 (all 12 questions correct).

Methods

Participants

130 UK participants averaged 25-34 years of age, 68% female, 32% male were recruited through the Prolific experimental subject pool for compensation. Participants were paid based on an hourly rate (£9.00 /hour) for the time spent in the study. The sample size was determined a priori with a goal of obtaining 0.95 power to detect a medium effect size of d= 0.35 (e.g., f=

0.17) at the standard 0.05 alpha error probability. The software program G*Power was used to conduct an a-priori power analysis. The suggested minimum sample size was 116, and 130 participants were recruited to account for potential exclusions due to failures in attentional checks and completions outside the window of accepted durations. E.g., one participant who completed the study in less than 10 minutes was excluded leading to a sample size of 129 participants in total . The exact materials and data for Study 4 are available in the Open Science Framework at https://osf.io/hvy6g.

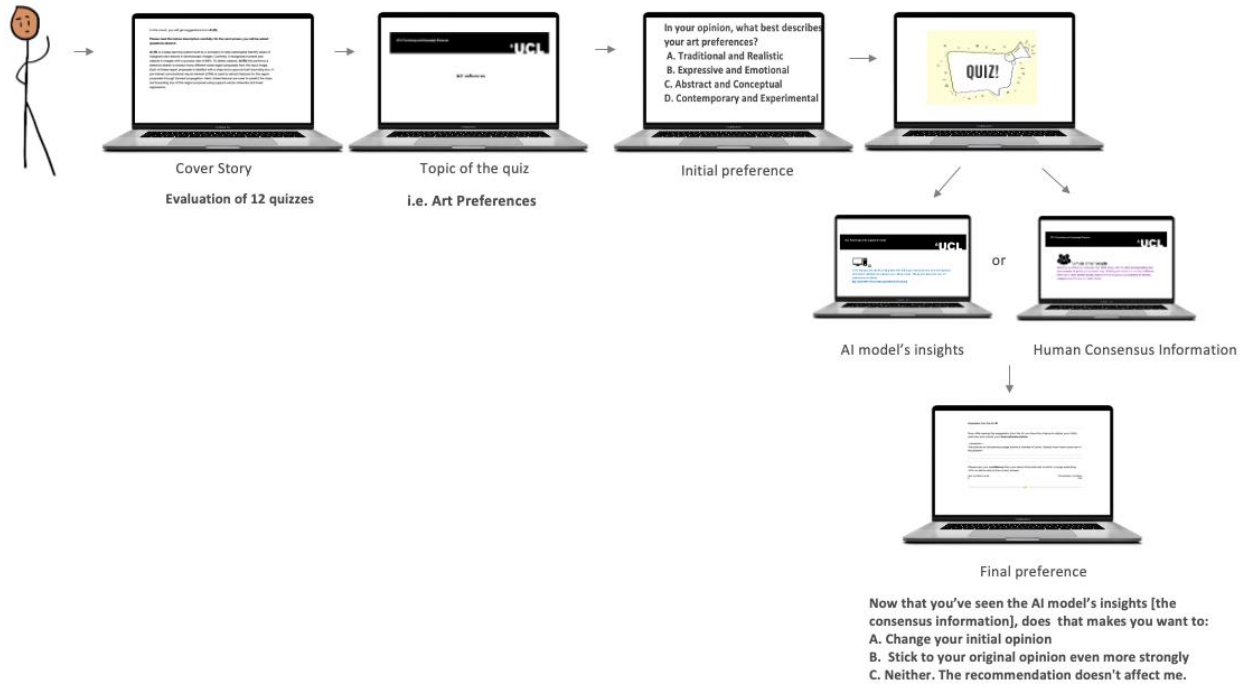
Materials And Procedure

The study was conducted online using the Qualtrics survey platform. Upon accepting the task on Prolific, participants were directed to Qualtrics where they read the instructions and gave their consent to participate in the study. Participants were tasked with evaluating 12 quizzes, each following a structured process. Initially, they encountered a title page that introduced the topic of the quiz (e.g., 'A Quiz to Help Uncover Your Art Preferences'). The quiz topics were subjective, focusing on personal preferences, and spanned various subjects such as art, sleep, meditation styles, stress and anxiety management, lifestyle choices, and communication styles.

Next, participants were required to record their perceived preferences related to the quiz topic before proceeding to fill out the quiz. After completing the quiz, they received insights labelled either as derived from an *AI model's analysis of 5,000 responses* from participants with similar demographics and quiz results, or as being *what most similar people prefer* based on the analysis of 5,000 responses from participants with similar demographics and quiz results. Although the insights were fundamentally the same (originating from the responses of 5,000 participants similar demographics and quiz results), the presentation differed, being either the output of an AI model or what most similar people prefer. Finally, participants were asked to

provide their final opinion regarding their preferences, which could have been revised based on the insights received. They also indicated their confidence level in their final opinion on a continuous scale ranging from 0 to 100. A graphical representation of the study's procedure is included in Figure 17.

Figure 17
Graphical Representation of Procedure followed in Study 4



The 12 quizzes used throughout the study were created using the '*ChatGPT3*' open-source NLP algorithm (OpenAI,2024). The prompt used to inquire ChatGPT3 per quiz was the following: '*Can you provide a quiz to discover one's [e.g., learning] preferences with eight questions where participants select one correct answer per question, and the results are based on the sum of scores?* The insights from the AI model validated participant's initial preferences in 25% of the cases and in 25% of the cases they didn't. Similarly, the human consensus information validated participant's initial preferences in 25% of the cases and in the rest 25% of the cases they

didn't. The above four scenarios (e.g., AI validates/AI invalidates/human validates/ human invalidates) were counterbalanced to account for order effect.

The design employed was a 2 (source of advice: AI vs human consensus) x 2 (validation: yes vs no) within-subject variables design, with 12 repeated measurements per participant. The dependent variable was the response to the question, '*After reviewing the advice, does that make you want to: a. Change your initial opinion, b. Stick to your choice even more strongly or c. Neither. the advice does not affect me.*' As such, the dependent variable was polytomous and categorical, with three possible responses: '*Change my initial choice'* (coded as 1), '*Stick to my choice even more strongly'* (coded as 2), or '*Neither. the advice does not affect me*' (coded as 3).

Data Analysis Strategy

Data were analysed using IBM SPSS Statistics (Version 29) predictive analytics software. To analyse repeated measurements with a categorical dependent variable a Generalized Estimating Equations (GEE) logistic regression model was used. Upon confirming that the assumptions for a GEE model were met, the GEE model was used to fit the data.

Results

Validation had a statistically significant effect on participant's final choice. The *source of advice* (e.g., 'who' validated them) had no statistically significant effect nor did the *source of advice X validation* interaction (see Table 16 for the Wald Chi Square tests of model effects). More specifically, the post-hoc test comparing choice update between *validation- yes* and *validation- no* revealed that the relative probability of changing one's opinion about their preferences versus not changing decreased by 0.32 points when the advice validated participants (e.g., the advice matched participant's initial opinion about their preferences), a 32% decrease in the probability of choice update ($X^2(1,129) = 119.819$, $p < .001$). These results show that when

advice was validating individuals' initial decisions, the likelihood of changing those decisions was reduced, regardless of whether that advice was labelled as coming from an AI or human advisor.

Table 16

Test of Model Effects on Choice Update (Study 4)

	Wald Chi-Square	*df*	*p*
source of advice	0.279	1	.598
validation	119.839	1	<.001
source of advice * validation	0.008	1	.928
AI literacy index	2.104	1	.147
age	0.057	1	.811
gender	10.318	1	.001

To further test that there was indeed no difference between the AI and the human condition, a Bayesian inference test for binomial proportions was used. The test examined whether the data follows the null distribution model Beta (2,2), which assumes an equal probability of choice updating (50/50) under the AI and human conditions, or an alternative distribution model Beta (5, 2) whereby there is a higher probability of choice update in the AI condition. The estimated Bayes factor was 1.597, which exceeds 1. This provides additional evidence for the null hypothesis, as it suggests that the observed data is approximately 1.6 times more likely under the null model than under the alternative hypothesis. Study 4 expanded on Study 3 by examining a broader range of preference-related decision tasks, encompassing preferences across various topics. The findings of Study 4 suggest that AI was just as effective validating people’s decisions as were other people.

Discussion

AI is increasingly taking on the role of an advisor (Rahwan, Cebrian, Obradovich, Bongard, Bonnefon, Breazeal, Crandall, Christakis, Couzin, Jackson, et al., 2019). However, its power to influence people's behaviour for good or bad rests on whether people follow its advice; at least as much, if not more, than they do with advice from other people. We started looking at what influences AI advice uptake by replicating and expanding on existing studies that have explored the role of the perceived objectivity of the decision at hand (Study 1& 2). Our results indicated that people's willingness to rely more on advice from AI compared to humans increases with the perceived objectivity of the decision at hand and that, for objective decisions, the expertise of the alternative human advisor plays a significant role. For objective decisions, people were less willing to follow AI advice when alternative human advice came from someone they regard as an expert, compared to when it comes from a layperson. These findings are consistent with previous research which also provides evidence that perceived objectivity and human expertise affect reliance on an AI advisor(Castelo et al., 2019; Logg et al., 2019).

In Study 3, we examined advice uptake in preference-based decisions, driven by the growing prevalence of AI offering abundant advice in areas such as food, travel, music, movies, career choices, and even romantic partnerships. We first examined people's behavioural responses to coffee recommendations generated by an AI algorithm and presented alongside human recommendations (Study 3). We found that people were less likely to change their coffee choices when these choices were validated by an AI algorithm (e.g., when the AI algorithm recommendations matched participant's initial choices). That was not the case when other people validated participant's initial choices. This finding suggested that AI was more effective validating people than other people. To further test this finding, in Study 4, we expanded the

range of preference-based decisions under examination from one (e.g., coffee preference) to twelve covering a large array of preferences (from sleep preferences to art preferences and preferred ways to deal with stress etc.). We found AI to be just as effective validating people's preference-based decisions as were other people. We further discuss these findings below.

First, an interesting observation emerges from the critical examination of the combined findings from Studies 2 to 4. Study 2 relied on participants' self-reported intentions rather than actual behaviours, whereas the latter two studies (Studies 3 and 4) measured behaviour in relevant decision tasks by using the JAS paradigm. By juxtaposing the findings, we see that while self-reports suggested people would be less likely to rely on AI than humans for advice in subjective decisions (Study 2), behavioural data painted a different picture. People listened to AI that validated them but not to humans that validated them when selecting their favourite coffee (Study 3), and AI was as effective validated people's preference- based decisions as were other people (Study 4). This highlights that self-reported intentions do not always translate into actual behaviour and underscores the need for further research that captures behavioural data to better understand how people trust AI. The behavioural data collected will complement—and potentially challenge—data on people's reported attitudes toward AI by revealing potential discrepancies between stated attitudes and actual behaviours. Furthermore, behaviour data, gathered either in lab or real-life settings, could then be used to inform interventions designed to encourage or discourage AI advice uptake in specific contexts such as e.g., financial, medical, or educational contexts and among different cohorts (e.g., young children and adolescents).

Findings that vary depending on the type of measure used (self-report, behavioural or neural), is a known challenge in research looking at how people perceive AI where complimenting and juxtaposing measurements of the same construct (i.e., behavioural and neural

data) is often used (Blut et al., 2021; Thellman et al., 2022). For instance, in a study where the effect of anthropomorphism on trust in a computer (2D) vs a virtual (3D) agent was studied with both subjective and behavioural measures, anthropomorphism did not affect people's behavioural trust, however, anthropomorphism increased self-reported trust in the each AI agent (Kulms & Kopp, 2019).

Secondly, Study 2 highlighted the contextuality of AI advice uptake, indicating that factors such as the type of decision (objective vs. subjective) and the expertise of the alternative human advisor significantly influence the willingness to trust AI. This suggests that findings may vary considerably depending on decision-specific, contextual factors and warns against generalising results across decision contexts. Especially in cases where contextual factors have been shown to be easily malleable as is the case with perceived subjectivity of a decision. Indeed, perceived task objectivity was shown to be malleable. For example, framing a task as benefiting from quantitative analysis rather than intuition increases its perceived objectivity and that led to subsequent increase in AI advice uptake (Castelo et al., 2019).

Thirdly, in Study 3 participants listened to AI that validated them but not to humans that validated them whereas Study 4 indicated that AI is as effective validating people's decisions as are other people in a wider array of preference-based decisions. These findings suggests that AI can provide similar social functions, such as validation, during decision making as other people, which in turn, carries ethical implications for the design of advisory AI and recommender systems. Recommender systems in particular, are typically designed with a specific objective function, such as e.g., maximising profit, and are trained on large amounts of users' data (Alm & Sheffrin, 2017). Ethical implications arise from that, however, as recommender systems come with biases inherent in the data they are trained on and objective functions that can (perhaps)

unintentionally alter people's preferences in an effort to maximise their objective function (Abdollahpouri et al., 2020; Adomavicius et al., 2013; Agan et al., 2023; Chaney et al., 2018; Jesse & Jannach, 2021; Kramer et al., 2014; Merrill & Oremus, 2021). Indeed, there have been cases where recommender systems, especially those trained using reinforcement learning, engage in what researchers call *'user tampering*,' where they polarise users to increase success with subsequent recommendations that align with this induced polarisation (Evans & Kasirzadeh, 2021). And natural language processing (NLP) models have been shown to have the ability to detect and strategically use deception as a beneficial tactic in negotiation tasks (Lewis et al., 2017). The above, when considered alongside the finding that AI can effectively provide social validation during decision-making, at least as effective as other people, underscores the need for responsible designing of AI advisory and recommender systems.

Furthermore, AI being as effective in validating people's preference -related decisions as other people (Study 4) appears to contradict previous research indicating that people generally treat AI and human advice similarly, although they tend to discount AI advice to a lesser extent than human advice– a phenomenon that the researchers termed as algorithmic appreciation (Logg et al., 2019). However, a closer examination of the differences between the studies reveals that there is no true contradiction. In the aforementioned studies, the identical advice—labelled as coming either from an AI algorithm or a human—was always an average of estimates from past participants, which did not necessarily match the participant's initial estimate. E.g., the impact of advice that matches the participant's initial estimate (e.g., validation effect) was not something that researchers chose to investigate and as such manipulate. This is however different to what was done here. This highlights how sensitive the findings on AI advice uptake are to contextual differences and serves as a caution to carefully consider variations in experimental designs when

comparing findings across studies or equally when attempting to generalise lab-based results to real-world settings.

Finally, the finding that AI being as effective in validating people as other people (Study 4) invites careful testing of future behavioural interventions that may seek to increase advice uptake on the basis of the identity of the advisor. For example, based on our results, it seems unlikely that health-related interventions offering advice on issues such as sleep preferences, eating habits, or exercise routines would benefit from either concealing the fact that the advice comes from an AI model or emphasising the presence of a human behind it.

Limitations and Directions for Future Research

The studies under this chapter have several limitations. Firstly, using a coffee preference task in Study 3 may not have been ideal, as people often hold strong views about their coffee preferences. Perceiving oneself expert has consistently been shown to lead to greater discounting of advice (Johnson & Fowler, 2011; Moore & Healy, 2008; Moore et al., 2015; Russo & Schoemaker, 1992). Study 4 accounted for this by including a broader range of subjective decisions across different preference domains. However, future studies could examine advice coming from AI vs human advisor, while also varying the level of uncertainty associated with the decision at hand. In situations where people feel less confident about their decisions, they might be more likely to seek advice that presents alternatives rather than validates them. In these cases, it will be interesting to explore whether advice in the form of alternatives options coming from an AI model are valued more, less, or equally as those offered by humans.

Moreover, exploring preference- based decisions that involve trade-offs—such as balancing food preferences with health considerations or travel choices with climate change concerns—could influence how people view the source of validating advice. It's possible that, in

trade-off scenarios, people may place more value on social advice rather than from an AI model, even if the AI is trained on human data. This could be because the label 'human' may carry a stronger association with social norms than the label 'AI'. Further research is needed to clarify whether AI can fulfil the same social functions as humans in preference- related decisions involving trade-offs.

Furthermore, in the current set of studies the spotlight was drawn on preference-related decisions. Future research should explore other decision-making contexts where i.e., the decisions are more objective and consequential in nature such as medical, legal, or financial decisions. Previous studies have noted a stated aversion to AI advice in moral contexts (Bigman & Gray, 2018). It will be interesting to explore whether AI fulfils the social function of validation in the same way humans do when making decisions involving moral considerations. For these decisions, people may feel more reassurance from human validation confirming that their decisions are reasonable or acceptable. Equally, human validation might be more highly valued in these contexts, as it allows sharing of responsibility with other people.

Another limitation is that in Studies 3 and 4, which employed behavioural measures, both AI and human advice were based on answers from relevant quizzes. For instance, in Study 4 the AI advice was framed as coming from *'an AI trained on 5,000 quiz responses* 'and the human advice was framed as coming from *'5,000 people who completed the same quiz'* Thus, the quiz was integral to both sources of advice, serving as input data for each. However, the effectiveness of a quiz as a foundation for advice may be questionable. If participants viewed the quiz as an inadequate tool for revealing their preferences, this perception could have influenced their responses in the study, although it would have similarly affected the comparison of interest (AI vs human validation). Future studies aiming to replicate the current set of studies should do so in

more ecologically valid settings, where AI advice either comes from an actual AI advisory model or the AI advisor in these studies is presented as having the capabilities of most of today's multimodal AI models that are trained on far more, both in quantity and variety, data than *5,000 quiz responses*. In this case, we might expect people to perceive AI as a more capable advisor than a human (even an expert) or a group of people.

Furthermore, participants were informed about the high-level process by which AI advice was generated in both Study 3 and 4. However, they were not made aware of the specific incentive structure of the advisors. Individuals may behave differently when they know that a human or AI advisor stands to benefit from their actions. This scenario is common in many recommender systems that operate with an objective function, such as maximising profits. Recent research has indicated that although individuals take into account the payoffs for machines, they place greater importance on the payoffs for humans (von Schenk et al., 2023). Future studies could therefore expand the current findings by also looking at the effect of communicating information about the incentive function of the AI advisor when contrasting it with human advice (that may or may not stand to benefit).

Finally, the responses of younger cohorts (e.g., young children and adolescents) to AI advice should be examined as these cohorts are expected to be more familiar with AI tools while they are highly susceptible to social influence by their peers (Knoll et al., 2015). Also, we do not know whether people will continue to perceive advice from AI in the same way over time, as AI becomes more sophisticated and integrated into human decision-making, and as people gain more experience interacting with AI outputs. Investigating AI advice uptake over time will require longitudinal studies, which, although more challenging to conduct, will provide valuable insights into how perception and use of AI advice evolves.

Conclusion

AI has the potential to transform people's lives. As people continue to interact with AI outputs that have the form of advice, gaining a better understanding of how people respond to AI advice is necessary for designing and implementing more effective and responsible AI advisory and recommender systems. Ultimately, this will help ensure that AI fulfils its potential in transforming people's lives by helping them make better decisions. In four studies we investigated how humans respond to AI advice capturing attitudes and behaviour and while comparing AI to alternative human advice. The studies reveal that findings on AI advice uptake are highly susceptible to contextual factors (e.g., nature of the decision), and the type of measurement (e.g., self-report, behavioural). When it comes to AI that offers validation, our findings suggest that AI can fulfil this social function as effectively as other people, if not more. People are just as likely to listen to AI that validates their decisions as to other people who provide validation. This, however, can be problematic especially in online spaces, where AI is often designed as a mirror which, trained on people’s individual and collective data, can only reflect those data back to them and as such is more likely to validate them than not. The findings contribute to efforts towards managing AI advice and recommender systems responsibly.

Chapter 5: Discussion of Empirical Evidence

This thesis set out to explore how people perceive and make decisions with AI, using methods from psychology to map the perceptions and investigate the behaviours. Through a series of eight studies, it investigated the human perspective on AI, offering empirical insights.

Studies 1 and 2 in Chapter 2 investigated AI perception using two conceptual frameworks from social psychology, the Stereotype Content Model (SCM) (Fiske et al., 2002) and the Mind Perception Dimensions (MPD) model (Gray et al., 2007), and a novel, data-driven model, derived from the other two, the AI Stereotype Model (AISM). These models were assessed in terms of how well they could map perception across a diverse range of AI agents, considering variations in design features (such as form, movement, and social interaction), embodiment (both physical and virtual), and intended purpose of use. The findings indicated that AI perception is not homogeneous; instead, distinct stereotypes emerge based on perceptions of competence and experience—the two key dimensions of AISM. Also, this model proved to be more effective than the other two models (MDP and SCM) in mapping AI perception.

Studies 3 and 4 in Chapter 3 examined how people form evaluations about the trustworthiness of AI models based on the three determinants of trust in automation - performance, process, and purpose (Lee & See, 2004). The findings suggested that trust in AI models is influenced by evaluations of its performance, process and purpose with different weight placed on each determinant of trust depending on the moral nature of the decision. In decisions with moral considerations, the AI's moral stance (the 'why' behind its decisions) was found to drive trust. In non-moral decisions, trust was driven by the AI's process (the 'how' it decides), with detailed explanations being trusted more.

Studies 5–8 in Chapter 4 explored how people respond to AI-generated advice compared to human advice. Consistent with previous research, the findings suggested that willingness to trust AI more than a human advisor increased with the perceived objectivity of the decision (Studies 4 and 5). Studies 6 and 7 expanded existed research by taking a closer look at advice that validates people in subjective decisions, such as personal preferences. The findings indicated that people treat advice that validates them the same, regardless of whether it is AI or other people that validates them.

In the rest of this chapter, we discuss the emergence of AI stereotypes, building on the initial evidence for their existence presented in Chapter 1. Key theoretical insights and practical implications drawn from the collective work presented in this thesis are discussed next. Finally, we conclude with future directions for research.

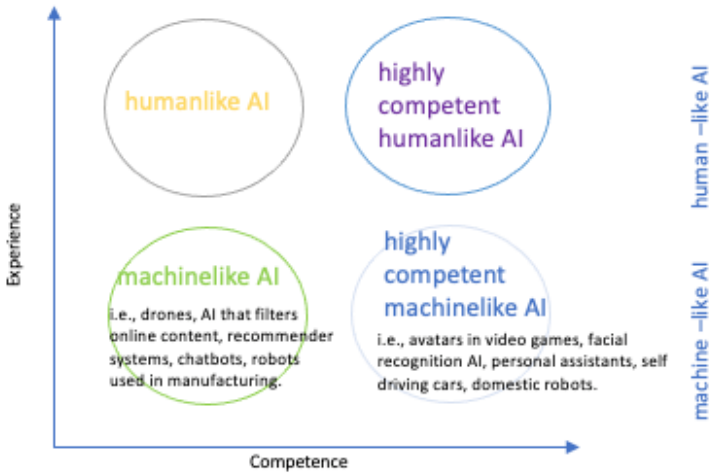
The Emergence of AI Stereotypes

Based on the AISM (Chapter 2), the following four stereotypes are starting to emerge in the two-dimensional space of competence x experience (Figure 18). The low competence / high experience cluster which will (currently this cluster is empty, as seen in Chapter 2) potentially include AI agents that will be perceived as lacking high competence but be rated high in experiential traits relative to other AI agents. We coin this category as the ‘humanlike AI’ cluster because, as seen in Chapter 2, perceived human likeness of AI decreased with perceived competence and increased with perceived experience. Next is the high competence / high experience AI cluster which will potentially (currently this cluster is empty, as seen in Chapter 2) include AI that will be perceived as high both in competence and experience relative to other AIs. We coin this category as the ‘high competent humanlike AI’ as it will comprise AI that will be seen as surpassing the ‘humanlike AI’ cluster in competence while still being rated high in

experience relative to other AI agents. Artificial General Intelligence (AGI), defined as AI capable of surpassing human in practically every task (Pennachin & Goertzel, 2007) and currently a topic of interest within the AI research community, is likely to fall into this cluster, when and if realised and introduced to the public.

Figure 18

The AI Stereotypes Model (AISM)



In the lower part of the two-dimensional space, AI perceived as low in experience is positioned. This is where all of today's AI was found to cluster, as seen in Chapter 2. We term the AI as 'machine-like AI', since the perception of experience, which as seen in Chapter 2 plays a significant role in perceived human likeness, is either very low or non-existent. Within the 'machine-like AI' clusters, the level of competence further differentiates AI into 'machine like AI' and 'high competence machinelike AI'. We refer to the 'lowest of the low' AI as mere 'machinelike AI' as it is AI that is perceived as low in both competence and experience relative to the AI in the other clusters, and as such regarded as mere machinery.

The work under this thesis provides first evidence for the emergence of the above AI stereotypes. Interesting paths to take this first evidence further is to replicate the mapping of AI perception in different points in time, with different samples and different AI agents. Given the current pace of technological advancements, new AI agents are expected to reach the public. And perhaps in 20 years from now more sophisticated algorithms or more human-like robots and avatars will have already become part of the human experience. Repeating the mapping of AI perception will give us insights into how humans perceive AI and how this perception evolves over time. Perhaps, it will also show us member in the stereotype clusters that are currently empty; something that will certainly carry many other interesting questions with it.

Methodologically, the emergence of AI stereotypes underscores that some AI (and certainly most of the AI available back in July 2022 when we performed the first mapping) might be more likely to share perceptual similarities with machines. As such, they may be more effectively studied through that lens. Or equally, it points to the fact that the current focus of research in comparing responses to human targets - particularly brain responses; for a detailed overview of studies comparing brain processing of responses to human and to AI targets see Harris (2024) - might be obscuring other research questions that could be asked, when not comparing with the human perceptual target but instead making comparisons across different AI perceptual targets.

Key Theoretical Insights

The first key theoretical insight is the need to map AI perception across its diversity. As seen in Chapter 2, AI is not a homogenous perceived target. Mapping AI perception across the diverse range of AI agents can however help reveal differences and identify potential similarities in the way people perceive them that go beyond design or contextual aspects. This approach can

give rise to new research questions, such as i.e., why is certain AI perceived similarly in terms of traits of experience and competence? what causes AIs with different embodiment and/or purpose of use to be grouped within the same perceptual category? Adopting a vertical approach in the mapping of AI perception—spanning across different AI agents rather than focusing on just one— can give rise to such questions. In addition, it helps avoid the tendency to focus exclusively on a single type of AI agent when studying AI perception. It also enables the mapping of new AI agents as it requires updating to reflect the perception of new AI agents that are being invented and integrated into people's lives.

The second key theoretical insight is the need to update the mapping in order to capture shifts in perception. Research findings need to be updated to keep pace with the technological advancements as new and more sophisticated AI agents are nowadays reaching the public and as people's increased familiarity with AI is likely to affect their perceptions and behaviours. For instance, the studies under Chapter 2 which explored stereotype formation in AI were conducted in July 2022, prior to the surge in public interest in large language models like ChatGPT-3 and ChatGPT- 4 in late 2022 and early 2023. As a result, ChatGPT was not even ranked among the top 23 AIs that people were familiar with in July 2022. If these studies were repeated today, ChatGPT would have likely been ranked as one of the AIs that people are very familiar with. This underscores the shifting landscape of AI perception. If anything, these shifts should encourage researchers to exercise caution when sharing their findings. Including a timestamp with research results could be particularly valuable for future comparisons. It also points to the need for longitudinal studies. While more difficult to conduct, longitudinal studies will track changes in perception as new AI agents reach the public and as the level of integration of the existing ones and people's familiarity with them increases.

How people perceive and make decisions with AI is expected to change over time, not only due to the invention and integration of new AI agents into human society and the growing familiarity with AI, which will likely influence these perceptions and behaviours, but also because of the potential developmental impact of interacting with AI from an early age. For instance, today's schoolkids or adolescents have a completely different experience with AI having started interacting with AI agents much earlier in life than previous generations, and next generations are expected to be even more apt with AI. These generational differences can however shape stereotypes around AI (Chapter 2) and can have an impact on how people trust and decide with AI (Chapters 3 and 4). This is the third key theoretical insight. Perception and behaviours towards AI need to start being studied across different cohorts, including more studies looking at younger generations. These studies could reveal how early interaction with AI influences stereotype formation and may uncover distinct behaviours toward AI outputs such as recommendations and advice.

Practical Recommendations

When it comes to practical recommendations, recognising the diversity in AI perception (Chapter 2) can inform more nuanced AI design. For example, perception of human likeness was found to be shaped by perception of imperfection (perceived human likeness decreased with perceived competence) and perception of experience (perceived human likeness increased with perceived experience). These insights can guide AI development to better align with human expectations—if making AI more human-like is contextually appropriate, ethical, and beneficial. And recognising that trustworthiness evaluations are influenced by the combined assessments of an AI's purpose, performance, and process, depending on the moral nature of the decision (Chapter 3), provides valuable insights into the types of explanations—such as teleological

(purpose-driven) and mechanical (process-driven)—required to make AI explainable, with a balance tailored to the moral context of the decision.

Recognising the diversity in AI perception (Chapter 2) can also guide the development of targeted legal frameworks for AI. Governing AI is not an easy task and the best approach to doing so remains a topic of ongoing debate (Clarke, 2019; Zaidan & Ibrahim, 2024). A primary challenge involves the fact that there is no such thing as 'one AI', given the vast diversity of AI agents and the different contexts they can be applied (Smuha, 2021). The mapping of AI perception provides evidence that further reinforces the above observation about the need for targeted regulation of AI. What is more, it invites for the exploration of the behavioural and emotional reactions towards the four AI stereotypes (Figure 18) which are likely to be different for each group, leading to different considerations about moral responsibilities and protections towards AI. For instance, it is unlikely all AI agents to engage with moral, societal, and philosophical question based on how people see them. Perhaps, the ones that are seen as more human like than others (upper part of Figure 18) will be the ones that will invite considerations for expanding the scope of moral responsibilities and protections typically reserved for humans to include AI.

Moreover, any attempt to regulate AI should ideally be informed by work similar to the work undertaken in this thesis e.g., work aiming at understanding how humans perceive and make decisions with AI. This approach will ensure that AI systems are aligned with societal values and needs in ways that are trustworthy, legitimate, and beneficial for humans. Nevertheless, Chapter 4 highlights a key challenge in this type of work: people's stated opinions may not always align with their behaviours in specific contexts (e.g., discrepancies observed between self-report and behavioural data). To address this, a practical recommendation from

Chapter 4 is to complement survey-based research with behavioural data. Taking this recommendation a step further, and by drawing from psychology research on social robots where neuroscientific measures have been extensively applied (Henschel et al., 2020), a further recommendation would be to combine different types of measurements, including more neural data . Combining self-report and behavioural data with neural data, although the latter are more difficult and costly, will provide insights into the nuances of perception and behaviour, nuances that behavioural and self-report data alone cannot capture. Also, neural data will shed light to the underlying brain mechanism supporting perception and behaviours. In addition, given the strong influence of contextual factors on trust in AI (as seen in Chapters 3 and 4), a further recommendation is to investigate phenomena like trust in AI and AI advice uptake in the specific real life decision domains (e.g., financial, medical, legal) by complementing lab studies with field studies conducted in real-world decision-making contexts.

Finally, Chapter 4 highlighted that AI is as effective validating people's decisions as other people. However, this poses a risk, as AI provides validation with a scale, ease, and consistency that humans cannot match. Indeed, AI's validation effect has already raised concerns about keeping people in their own 'opinion bubbles', polarising beliefs and voting behaviours (O'neil, 2017). Concerns have also been raised in terms of AI's validation effect posing a threat to the freedom of thought, especially when the way AI 'filters' reality obscure alternative perspectives and choices (Vallor, 2024). While people may generally understand that AI relies on vast datasets of past behaviours, they may not realise that its 'reflection' lacks the diverse opinions and choices it fails to present. This highlights the need to educate people about AI's validation effect. Public awareness campaigns, such as documentaries, public debates, and updates to school curricula, could help address this issue. These initiatives should emphasise that

while AI is a powerful tool, it should not dictate what individuals think, prefer, or do, but rather be used critically. AI primarily reflects existing knowledge, much like a parrot echoing familiar ideas (Bender et al., 2021) and while this is not inherently negative, it is crucial for people to recognise its limitations and actively seek diverse perspectives to challenge their ideas and foster critical thinking.

Future Directions

First, this thesis can be expanded in two ways: content and temporal scope. A content-scope expansion involves replicating studies with methodological improvements to address limitations and testing the hypotheses posed in this thesis through new experimental approaches. For example, future research that seeks to explore the effect of the 3Ps (Chapter 3) could examine interactions with real-world AI, such as large language models (LLMs), manipulating factors like error rates (e.g., performance), moral character instilled in the AI (e.g., purpose), and explanation provided (e.g., process). This would enhance ecological validity by integrating real-world AI into experimental settings. A temporal-scope expansion would involve mapping AI perception along the identified dimensions of competence x experience (Chapter 2) over time and with diverse samples, and monitoring changes in perception of AI, as people become more experienced with AI and the technology evolves.

Secondly, coupled with how people perceive and make decisions with AI, future research should also aim to investigate how interactions with AI reshape human cognition. Increasing reliance on AI tools like Google, Wikipedia, and LLMs such as ChatGPT may be altering information encoding, retrieval, and processing. For example, studies have shown that access to Google changes memory patterns, with people recalling how to find information rather than the information itself (Sparrow et al., 2011). Similarly, reliance on AI for cognitive tasks has been

shown to affect perceived knowledge (Fisher & Oppenheimer, 2021) while the danger of 'illusions of understanding 'in scientific knowledge generation has also been highlighted (Messeri & Crockett, 2024). Memory, metacognition, and knowledge formation are just parts of the broader cognitive picture. AI's impact on learning, thinking, and problem-solving remains underexplored. Systematic behavioural and neurological research is needed to assess the impact of repeated interactions with AI on cognition. Starting now will also allow us to track changes in cognitive functioning due to repeated interactions with AI and the human brain's adaptation to AI over time.

Finally, future research, coupled with validating the emergence of AI stereotypes using diverse samples, should also seek to understand how the brain differentiates between different AI agents based on the dimensions of competence and experience and over time, treating the findings under Chapter 2 as behavioural data to further investigate using neuroimaging techniques. Neuroimaging techniques, like fMRI or fNIRS may seek to explore the neural regions that correlate with the dimensions of competence and experience of the AISM (Figure 18), giving us insights on what allows the brain to differentiate AI agents. Brain imaging technique fNIRS in particular, which similarly to fMRI tracks the blood oxygen level dependent response, could be a more suitable technique for studying human perception of artificial entities in dynamic and interactive settings due to its portability (Henschel et al., 2020).

Future research may also seek to expand existing literature on how the brain differentiates between AI and humans, seeking to understand the neural correlates of AI presence. An interesting approach would be to see what the brain imaging studies will reveal when the AI becomes the main topic of interest treating the human target as the baseline in the comparisons. So far, what is known about how the brain processes AI agents is heavily based on

studies that compare AI to human target (Harris, 2024; Vaitonytė et al., 2023) And in these studies, AI has served as the non-human baseline. Perhaps new existing research questions will emerge when the AI is not compared to human targets or if compared it serves as the main topic of interest and human becomes the control condition (Harris, 2024) . This line of research will ultimately shed light to how the human brain differentiates between humans and AI, enabling us to determine whether AI falls within the scope of moral responsibilities and protections typically reserved for humans.

Conclusion

AI is becoming an increasingly integral part of human life, with more sophisticated algorithms and more humanlike avatars and robots potentially becoming a common reality within the next 20 years—or even sooner. However, our understanding of how people perceive AI and how the brain processes it and adapts to it remains relatively limited. It is the author's genuine aspiration this thesis to contribute to a series of ongoing research, including longitudinal studies and comprehensive programs, which will diligently chronicle the evolving human perspective on AI, capturing its changes over time. Ultimately, with endeavours as such in place, we will be able to, one day, narrate the psychological story of humanity's relationship with its most remarkable creation. Not through anecdotes, sensational media, or pop culture and fictional portrayals—though the author of this thesis holds great affection for some of the latter—but by capturing the human experience with AI as it truly is; as it unfolds in the perceptions, behaviours, and is supported by the underlying brain mechanisms of humans across different points in time and generations.

References

Abdollahpouri, H., Mansoury, M., Burke, R., & Mobasher, B. (2020). Addressing the multistakeholder impact of popularity bias in recommendation through calibration. *arXiv preprint arXiv:2007.12230*.

Abubshait, A., & Wiese, E. (2017). You look human, but act like a machine: agent appearance and behavior modulate different aspects of human–robot interaction. *Frontiers in psychology*, *8*, 1393.

Abubshait, A., & Wykowska, A. (2020). Repetitive robot behavior impacts perception of intentionality and gaze-related attentional orienting. Front Robot AI 7: 565825. In.

Abubshait, A., Perez-Osorio, J., De Tommaso, D., & Wykowska, A. (2021). Collaboratively framed interactions increase the adoption of intentional stance towards robots. 2021 30th IEEE international conference on robot & human interactive communication (RO-MAN),

Adomavicius, G., Bockstedt, J. C., Curley, S. P., & Zhang, J. (2013). Do recommender systems manipulate consumer preferences? A study of anchoring effects. *Information Systems Research*, *24*(4), 956-975.

Adomavicius, G., Bockstedt, J. C., Curley, S. P., & Zhang, J. (2013). Do recommender systems manipulate consumer preferences? A study of anchoring effects. *Information Systems Research*, *24*(4), 956-975.

Agan, A. Y., Davenport, D., Ludwig, J., & Mullainathan, S. (2023). *Automating automaticity: How the context of human choice affects the extent of algorithmic bias*.

Alm, J., & Sheffrin, S. M. (2017). Using behavioral economics in public economics. In (Vol. 45, pp. 4-9): SAGE Publications Sage CA: Los Angeles, CA.

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General*, *136*(4), 569.

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(5), e1424.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2022). Machine bias. In *Ethics of data and analytics* (pp. 254-264). Auerbach Publications.

Asch, S. E. (1946). Forming impressions of personality. *The journal of abnormal and social psychology*, *41*(3), 258.

Ayeni, O. O., Al Hamad, N. M., Chisom, O. N., Osawaru, B., & Adewusi, O. E. (2024). AI in education: A review of personalized learning and educational technology. *GSC Advanced Research and Reviews*, *18*(2), 261-271.

Babel, F., Hock, P., Kraus, J., & Baumann, M. (2022). Human-robot conflict resolution at an elevator-the effect of robot type, request politeness and modality. 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI),

Banks, J. (2020). Theory of mind in social robots: replication of five established human tests. *International Journal of Social Robotics*, *12*(2), 403-414.

Baraka, K., Alves-Oliveira, P., & Ribeiro, T. (2020). An extended framework for characterizing social robots. *Human-Robot Interaction: Evaluation Methods and Their Standardization*, 21-64.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest*, *20*(1), 1-68.

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, *1*, 71-81.

Bartneck, C., Suzuki, T., Kanda, T., & Nomura, T. (2007). The influence of people's culture and prior experiences with Aibo on their attitude towards robots. *Ai & Society*, *21*, 217-230.

Baumol, W. J., & Quandt, R. E. (1964). Rules of thumb and optimally imperfect decisions. *The American economic review*, *54*(2), 23-46.

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., & Tanaka, F. (2018). Social robots for education: A review. *Science Robotics*, *3*(21), eaat5954.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). Proceedings of the 2021 ACM conference on fairness, accountability, and transparency,

Berger, B., Adam, M., Rühr, A., & Benlian, A. (2020). Watch Me Improve—Algorithm Aversion and Demonstrating the Ability to Learn. *Business & Information Systems Engineering*, 1-14.

Bianco, F., & Ognibene, D. (2019). Transferring adaptive theory of mind to social robots: Insights from developmental psychology to robotics. Social Robotics: 11th International Conference, ICSR 2019, Madrid, Spain, November 26–29, 2019, Proceedings 11,

Bicchieri, C. (2016). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21-34. https://doi.org/10.1016/j.cognition.2018.08.003

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society*, *2*(59-68).

Blashfield, R. K., & Aldenderfer, M. S. (1988). The methods and problems of cluster analysis. In *Handbook of multivariate experimental psychology* (pp. 447-473). Springer.

Blut, M., Wang, C., Wünderlich, N. V., & Brock, C. (2021). Understanding anthropomorphism in service provision: a meta-analysis of physical robots, chatbots, and other AI. *Journal of the Academy of Marketing Science*, *49*, 632-658.

Bobek, D. D., Hageman, A. M., & Kelliher, C. F. (2013). Analyzing the Role of Social Norms in Tax Compliance Behavior. *Journal of Business Ethics*, *115*(3), 451-468. https://doi.org/10.1007/s10551-012-1390-7

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

Bossi, F., Willemse, C., Cavazza, J., Marchesi, S., Murino, V., & Wykowska, A. (2020). The human brain reveals resting state activity patterns that are predictive of biases in attitudes toward robots. *Science Robotics*, *5*(46), eabb6652.

Broadbent, E. (2017). Interactions with robots: The truths we reveal about ourselves. *Annual Review of Psychology*, *68*(1), 627-652.

Broadbent, E., Kumar, V., Li, X., Sollers 3rd, J., Stafford, R. Q., MacDonald, B. A., & Wegner, D. M. (2013). Robots with display screens: a robot with a more humanlike face display is perceived to have more mind and a better personality. *PloS one*, *8*(8), e72589.

Bruce, A., Knight, J., Listopad, S., Magerko, B., & Nourbakhsh, I. R. (2000). Robot improv: Using drama to create believable agents. Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065),

Brynjolfsson, E., Hui, X., & Liu, M. (2019). Does machine translation affect international trade? Evidence from a large digital platform. *Management science*, *65*(12), 5449-5460.

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.

Cao, W., Song, W., Li, X., Zheng, S., Zhang, G., Wu, Y., He, S., Zhu, H., & Chen, J. (2019). Interaction with social robots: Improving gaze toward face but not necessarily joint attention in children with autism spectrum disorder. *Frontiers in psychology*, *10*, 1503.

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

Chalhoub, G., & Flechais, I. (2020). "Alexa, are you spying on me?": Exploring the Effect of User Experience on the Security and Privacy of Smart Speaker Users. HCI for Cybersecurity, Privacy and Trust: Second International Conference, HCI-CPT 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22,

Chaminade, T., Rosset, D., Da Fonseca, D., Nazarian, B., Lutcher, E., Cheng, G., & Deruelle, C. (2012). How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. *Frontiers in human neuroscience*, *6*, 103.

Chaney, A. J., Stewart, B. M., & Engelhardt, B. E. (2018). How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. Proceedings of the 12th ACM conference on recommender systems,

Chang, S. E., Liu, A. Y., & Shen, W. C. (2017). User trust in social networking services: A comparison of Facebook and LinkedIn. *Computers in Human Behavior*, *69*, 207-217.

Chen, G.-D., Nurkhamid, & Wang, C.-Y. (2011). A survey on storytelling with robots. Edutainment Technologies. Educational Games and Virtual Reality/Augmented Reality Applications: 6th International Conference on E-learning and Games, Edutainment 2011, Taipei, Taiwan, September 2011. Proceedings 6,

Chevalier, P., Kompatsiari, K., Ciardo, F., & Wykowska, A. (2020). Examining joint attention with the use of humanoid robots-A new approach to study fundamental mechanisms of social cognition. *Psychonomic Bulletin & Review*, *27*, 217-236.

Chiou, E. K., & Lee, J. D. (2023). Trusting automation: Designing for responsivity and resilience. *Human factors*, *65*(1), 137-165.

Choi, A., Melo, C. D., Woo, W., & Gratch, J. (2012). Affective engagement to emotional facial expressions of embodied social agents in a decision-making game. *Computer Animation and Virtual Worlds*, *23*(3-4), 331-342.

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In *Advances in experimental social psychology* (Vol. 24, pp. 201-234). Elsevier.

Ciardo, F., De Tommaso, D., & Wykowska, A. (2022). Joint action with artificial agents: Human-likeness in behaviour and morphology affects sensorimotor signaling and social inclusion. *Computers in Human Behavior*, *132*, 107237.

Cifuentes, C. A., Pinto, M. J., Céspedes, N., & Múnera, M. (2020). Social robots in therapy and care. *Current Robotics Reports*, *1*, 59-74.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (Rev. ed.). Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203771587
Cohen, J. (2009). *Statistical power analysis for the behavioral sciences / Jacob Cohen* (Second edition. ed.). Psychology Press.
Collins, E. C., Prescott, T. J., Mitchinson, B., & Conran, S. (2015). MIRO: a versatile biomimetic edutainment robot. Proceedings of the 12th international conference on advances in computer entertainment technology,
Cross, E. S., & Ramsey, R. (2021). Mind meets machine: Towards a cognitive science of human–machine interactions. *Trends in cognitive sciences*, *25*(3), 200-212.
Cross, E. S., & Ramsey, R. (2021). Mind meets machine: Towards a cognitive science of human–machine interactions. *Trends in cognitive sciences*, *25*(3), 200-212.
Cross, E. S., Liepelt, R., de C. Hamilton, A. F., Parkinson, J., Ramsey, R., Stadler, W., & Prinz, W. (2012). Robotic movement preferentially engages the action observation network. *Human brain mapping*, *33*(9), 2238-2254.
Cross, E. S., Ramsey, R., Liepelt, R., Prinz, W., & Hamilton, A. F. d. C. (2016). The shaping of social perception by stimulus and knowledge cues to human animacy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1686), 20150075.
Cross, E. S., Riddoch, K. A., Pratts, J., Titone, S., Chaudhury, B., & Hortensius, R. (2019). A neurocognitive investigation of the impact of socializing with a robot on empathy for pain. *Philosophical Transactions of the Royal Society B*, *374*(1771), 20180034.
Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2007). The BIAS Map: Behaviors From Intergroup Affect and Stereotypes. *Journal of personality and social psychology*, *92*(4), 631-648. https://doi.org/10.1037/0022-3514.92.4.631
Cuddy, A. J., Fiske, S. T., & Glick, P. (2007). The BIAS map: behaviors from intergroup affect and stereotypes. *Journal of personality and social psychology*, *92*(4), 631.
Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in experimental social psychology*, *40*, 61-149.
Cuddy, A. J., Fiske, S. T., Kwan, V. S., Glick, P., Demoulin, S., Leyens, J. P., Bond, M. H., Croizet, J. C., Ellemers, N., & Sleebos, E. (2009). Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology*, *48*(1), 1-33.
Davis, F. (1986). A technology acceptance model for empirically testing new end-user information systems. *Theory and Results/Massachusetts Institute of Technology*.
Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *Mis Quarterly*, 319-340.
Dawe, J., Sutherland, C., Barco, A., & Broadbent, E. (2019). Can social robots help children in healthcare contexts? A scoping review. *BMJ paediatrics open*, *3*(1).
Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological bulletin*, *81*(2), 95-106. https://doi.org/10.1037/h0037613
de Graaf, M. M., Ben Allouch, S., & Van Dijk, J. A. (2015). What makes robots social?: A user's perspective on characteristics for social human-robot interaction. Social Robotics: 7th International Conference, ICSR 2015, Paris, France, October 26-30, 2015, Proceedings 7,
De Graaf, M., Ben Allouch, S., & Van Dijk, J. (2017). Why do they refuse to use my robot? Reasons for non-use derived from a long-term home study. Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction,

De Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*(3), 331.

Dellaert, B. G., Shu, S. B., Arentze, T. A., Baker, T., Diehl, K., Donkers, B., Fast, N. J., Häubl, G., Johnson, H., & Karmarkar, U. R. (2020). Consumer decisions with artificially intelligent voice assistants. *Marketing Letters*, *31*, 335-347.

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Dereshev, D., Kirk, D., Matsumura, K., & Maeda, T. (2019). Long-term value of social robots through the eyes of expert users. Proceedings of the 2019 CHI conference on human factors in computing systems,

Di Dio, C., Ardizzi, M., Schieppati, S. V., Massaro, D., Gilli, G., Gallese, V., & Marchetti, A. (2023). Art made by artificial intelligence: The effect of authorship on aesthetic judgments. *Psychology of Aesthetics, Creativity, and the Arts*.

Diab, D. L., Pui, S. Y., Yankelevich, M., & Highhouse, S. (2011). Lay perceptions of selection decision aids in US and non-US samples. *International Journal of Selection and Assessment*, *19*(2), 209-216.

Diemand-Yauman, C., Oppenheimer, D. M., & Vaughan, E. B. (2011). Fortune favors the (): Effects of disfluency on educational outcomes. *Cognition*, *118*(1), 111-115.

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.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management science*, *64*(3), 1155-1170.

Drexler, N., & Lapré, V. B. (2019). For better or for worse: Shaping the hospitality industry through robotics and artificial intelligence. *Research in Hospitality Management*, *9*(2), 117–120-117–120.

Duchowski, A. T., Krejtz, K., Krejtz, I., Biele, C., Niedzielska, A., Kiefer, P., Raubal, M., & Giannopoulos, I. (2018). The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation. Proceedings of the 2018 CHI conference on human factors in computing systems,

Duffy, B. R., & Joue, G. (2004). I, robot being. Intelligent Autonomous Systems Conference (IAS8), Amsterdam, The Netherlands,

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International journal of human-computer studies*, *58*(6), 697-718.

Einhorn, H. J., & Hogarth, R. M. (1975). Unit weighting schemes for decision making. *Organizational behavior and human performance*, *13*(2), 171-192. https://doi.org/10.1016/0030-5073(75)90044-6

El Naqa, I., Haider, M. A., Giger, M. L., & Ten Haken, R. K. (2020). Artificial intelligence: reshaping the practice of radiological sciences in the 21st century. *The British journal of radiology*, *93*(1106), 20190855.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, *114*(4), 864.

Evans, C., & Kasirzadeh, A. (2021). User tampering in reinforcement learning recommender systems. *arXiv preprint arXiv:2109.04083*.

Eyssel, F., & Kuchenbrandt, D. (2012). Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology*, *51*(4), 724-731.
Eyssel, F., & Reich, N. (2013). Loneliness makes the heart grow fonder (of robots)—On the effects of loneliness on psychological anthropomorphism. 2013 8th acm/ieee international conference on human-robot interaction (hri),
Eyssel, F., Kuchenbrandt, D., & Bobinger, S. (2011). Effects of anticipated human-robot interaction and predictability of robot behavior on perceptions of anthropomorphism. Proceedings of the 6th international conference on Human-robot interaction,
Eyssel, F., Kuchenbrandt, D., & Bobinger, S. (2011). Effects of anticipated human-robot interaction and predictability of robot behavior on perceptions of anthropomorphism. Proceedings of the 6th international conference on Human-robot interaction,
Eyssel, F., Kuchenbrandt, D., Bobinger, S., de Ruiter, L., & Hegel, F. (2012). If you sound like me, you must be more human: on the interplay of robot and user features on human-robot acceptance and anthropomorphism. New York, NY, USA.
Ferrari, F., Paladino, M. P., & Jetten, J. (2016). Blurring human–machine distinctions: Anthropomorphic appearance in social robots as a threat to human distinctiveness. *International Journal of Social Robotics*, *8*, 287-302.
Festinger, L. (1957). A Theory of Coijnitive Dissonance. In: Stanford, Cal.: Stanford niversity Press.
Fiske, S. T. (1991). Social cognition. In: McGraw-Hill.
Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination.
Fiske, S. T. T., & Taylor, S. E. (2020). Social cognition: From brains to culture.
Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A Model of (Often Mixed) Stereotype Content: Competence and Warmth Respectively Follow From Perceived Status and Competition. *Journal of personality and social psychology*, *82*(6), 878-902. https://doi.org/10.1037/0022-3514.82.6.878
Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, *11*(2), 77-83.
Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2018). 7 A model of (often mixed) stereotype content. *Social Cognition: Selected Works of Susan Fiske*, 163.
Fitrianie, S., Bruijnes, M., Li, F., Abdulrahman, A., & Brinkman, W.-P. (2022). The artificial-social-agent questionnaire: establishing the long and short questionnaire versions. Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents,
Fleisher, W. (2022). Understanding, idealization, and explainable AI. *Episteme*, *19*(4), 534-560.
Fong, K., Quinlan, J. A., & Mar, R. A. (2023). Select your character: Individual needs and avatar choice. *Psychology of Popular Media*, *12*(1), 30.
Franklin, S., & Graesser, A. (1996). Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents. International workshop on agent theories, architectures, and languages,
Fraune, M. R. (2020). Our robots, our team: Robot anthropomorphism moderates group effects in human–robot teams. *Frontiers in psychology*, *11*, 1275.
Fraune, M. R., Šabanović, S., & Smith, E. R. (2020). Some are more equal than others: Ingroup robots gain some but not all benefits of team membership. *Interaction studies*, *21*(3), 303-328.

Frith, C. D., & Frith, U. (2006). How we predict what other people are going to do. *Brain research*, *1079*(1), 36-46.
Frith, U., & Frith, C. (2001). The biological basis of social interaction. *Current Directions in Psychological Science*, *10*(5), 151-155.
Fussell, S. R., Kiesler, S., Setlock, L. D., & Yew, V. (2008). How people anthropomorphize robots. Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction,
Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press.
Gino, F. (2008). Do we listen to advice just because we paid for it? The impact of advice cost on its use. *Organizational behavior and human decision processes*, *107*(2), 234-245. https://doi.org/10.1016/j.obhdp.2008.03.001
Gino, F., & Moore, D. A. (2007). Effects of task difficulty on use of advice. *Journal of behavioral decision making*, *20*(1), 21-35. https://doi.org/10.1002/bdm.539
Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627-660.
Go, E., & Sundar, S. S. (2019). Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior*, *97*, 304-316. https://doi.org/10.1016/j.chb.2019.01.020
Goetz, J., Kiesler, S., & Powers, A. (2003). Matching robot appearance and behavior to tasks to improve human-robot cooperation. The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003.,
Gollwitzer, P. M., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in experimental social psychology*, *38*, 69-119.
Gollwitzer, P. M., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in experimental social psychology*, *38*, 69-119.
Goodrich, M. A., & Schultz, A. C. (2008). Human–robot interaction: a survey. *Foundations and Trends® in Human–Computer Interaction*, *1*(3), 203-275.
Goudey, A., & Bonnin, G. (2016). Must smart objects look human? Study of the impact of anthropomorphism on the acceptance of companion robots. *Recherche et applications en marketing (English edition)*, *31*(2), 2-20. https://doi.org/10.1177/2051570716643961
Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of Mind Perception. *Science (American Association for the Advancement of Science)*, *315*(5812), 619-619. https://doi.org/10.1126/science.1134475
Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, *125*(1), 125-130.
Greulich, R. S., & Brendel, A. B. (2022). " Feel, Don't Think" Review of the Application of Neuroscience Methods for Conversational Agent Research. ECIS,
Gursoy, D., Chi, O. H., Lu, L., & Nunkoo, R. (2019). Consumers acceptance of artificially intelligent (AI) device use in service delivery. *International Journal of Information Management*, *49*, 157-169.
Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California management review*, *61*(4), 5-14.
Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (1998). Multivariate data analysis. Uppersaddle River. *Multivariate Data Analysis (5th ed) Upper Saddle River*, *5*(3), 207-219.

Hancock, P. A., Billings, D. R., & Schaefer, K. E. (2011). Can You Trust Your Robot? *Ergonomics in design*, *19*(3), 24-29. https://doi.org/10.1177/1064804611415045
Hancock, P. A., Kessler, T. T., Kaplan, A. D., Brill, J. C., & Szalma, J. L. (2021). Evolving Trust in Robots: Specification Through Sequential and Comparative Meta-Analyses. *Human factors*, *63*(7), 1196-1229. https://doi.org/10.1177/0018720820922080
Haran, U., & Shalvi, S. (2020). The Implicit Honesty Premium: Why Honest Advice Is More Persuasive Than Highly Informed Advice. *Journal of experimental psychology. General*, *149*(4), 757-773. https://doi.org/10.1037/xge0000677
Haran, U., & Shalvi, S. (2020a). The implicit honesty premium: Why honest advice is more persuasive than highly informed advice. *Journal of Experimental Psychology: General*, *149*(4), 757.
Haran, U., & Shalvi, S. (2020b). The Implicit Honesty Premium: Why Honest Advice Is More Persuasive Than Highly Informed Advice. *Journal of experimental psychology. General*, *149*(4), 757-773. https://doi.org/10.1037/xge0000677
Haring, K. S., Silvera-Tawil, D., Watanabe, K., & Velonaki, M. (2016). The influence of robot appearance and interactive ability in HRI: a cross-cultural study. Social Robotics: 8th International Conference, ICSR 2016, Kansas City, MO, USA, November 1-3, 2016 Proceedings 8,
Harris, L. T. (2024). The neuroscience of human and artificial intelligence presence. *Annual Review of Psychology*, *75*(1), 433-466.
Harvey, N., & Fischer, I. (1997). Taking Advice: Accepting Help, Improving Judgment, and Sharing Responsibility. *Organizational behavior and human decision processes*, *70*(2), 117-133. https://doi.org/10.1006/obhd.1997.2697
Haslam, N., & Loughnan, S. (2014). Dehumanization and infrahumanization. *Annual Review of Psychology*, *65*(1), 399-423.
Helander, M. G. (2014). *Handbook of human-computer interaction*. Elsevier.
Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, *57*(3), 407-434.
Hou, Y. T.-Y., & Jung, M. F. (2021). Who is the Expert? Reconciling Algorithm Aversion and Algorithm Appreciation in AI-Supported Decision Making. *Proceedings of the ACM on human-computer interaction*, *5*(CSCW2), 1-25. https://doi.org/10.1145/3479864
Jesse, M., & Jannach, D. (2021). Digital nudging with recommender systems: Survey and future directions. *Computers in Human Behavior Reports*, *3*, 100052.
Johnson, D. D., & Fowler, J. H. (2011). The evolution of overconfidence. *Nature*, *477*(7364), 317-320.
Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion.
Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2023). Trust in artificial intelligence: Meta-analytic findings. *Human factors*, *65*(2), 337-359.
Kiesler, S., Powers, A., Fussell, S. R., & Torrey, C. (2008). Anthropomorphic interactions with a robot and robot–like agent. *Social cognition*, *26*(2), 169-181.
Kim, J., & Im, I. (2023). Anthropomorphic response: Understanding interactions between humans and artificial intelligence agents. *Computers in Human Behavior*, *139*, 107512.
Knoll, L. J., Magis-Weinberg, L., Speekenbrink, M., & Blakemore, S.-J. (2015). Social influence on risk perception during adolescence. *Psychological science*, *26*(5), 583-592.

Köbis, N. C., Troost, M., Brandt, C. O., & Soraperra, I. (2022). Social norms of corruption in the field: social nudges on posters can help to reduce bribery. *Behavioural Public Policy*, *6*(4), 597-624.

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.

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788-8790.

Kulms, P., & Kopp, S. (2018). A social cognition perspective on human–computer trust: The effect of perceived warmth and competence on trust in decision-making with computers. *Frontiers in Digital Humanities*, 14.

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. In *Proceedings of mensch und computer 2019* (pp. 31-42).

Laban, G., & Araujo, T. (2020). The effect of personalization techniques in users' perceptions of conversational recommender systems. Proceedings of the 20th ACM international conference on intelligent virtual agents,

Laban, G., George, J., Morrison, V., & Cross, E. (2021). Tell me more! Assessing interactions with social robots from speech. Paladyn J Behav Robot 12 (1): 136–159. In.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International journal of human-computer studies*, *40*(1), 153-184. https://doi.org/10.1006/ijhc.1994.1007

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50-80.

Lee, V. K., & Harris, L. T. (2014). Sticking with the nice guy: Trait warmth information impairs learning and modulates person perception brain network activity. *Cognitive, Affective, & Behavioral Neuroscience*, *14*, 1420-1437.

Leib, M., Köbis, N., Rilke, R. M., Hagens, M., & Irlenbusch, B. (2024). Corrupted by algorithms? how ai-generated and human-written advice shape (dis) honesty. *The Economic Journal*, *134*(658), 766-784.

Lewis, M., Yarats, D., Dauphin, Y. N., Parikh, D., & Batra, D. (2017). Deal or no deal? end-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125*.

Liberman, V., Minson, J. A., Bryan, C. J., & Ross, L. (2012). Naïve realism and capturing the "wisdom of dyads". *Journal of experimental social psychology*, *48*(2), 507-512. https://doi.org/10.1016/j.jesp.2011.10.016

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

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, *46*(4), 629-650.

Manzi, F., Peretti, G., Di Dio, C., Cangelosi, A., Itakura, S., Kanda, T., Ishiguro, H., Massaro, D., & Marchetti, A. (2020). A robot is not worth another: Exploring children's mental state attribution to different humanoid robots. *Frontiers in psychology*, *11*, 2011.

Martín, M., & Valiña, M. D. (2023). Heuristics, biases and the psychology of reasoning: state of the art. *Psychology*, *14*(2), 264-294.

Martini, M. C., Gonzalez, C. A., & Wiese, E. (2016). Seeing minds in others–can agents with robotic appearance have human-like preferences? *PloS one*, *11*(1), e0146310.
Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, *20*(3), 709-734.
Merrill, J. B., & Oremus, W. (2021). Five points for anger, one for a 'like': How Facebook's formula fostered rage and misinformation. *The Washington Post*, *26*.
Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1-38.
Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological review*, *115*(2), 502.
Moore, D. A., Tenney, E. R., & Haran, U. (2015). Overprecision in judgment. *The Wiley Blackwell handbook of judgment and decision making*, *2*, 182-209.
Naftulin, D. H., Ware, J. E., & Donnelly, F. A. (1973). The doctor fox lecture: A paradigm of educational seduction. *Journal of medical education*, *48*(7), 630-635.
Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447-453.
Okanda, M., Taniguchi, K., Wang, Y., & Itakura, S. (2021). Preschoolers' and adults' animism tendencies toward a humanoid robot. *Computers in Human Behavior*, *118*, 106688.
Oppenheimer, D. M. (2006). Consequences of erudite vernacular utilized irrespective of necessity: problems with using long words needlessly. *Applied cognitive psychology*, *20*(2), 139-156. https://doi.org/10.1002/acp.1178
Pak, R., McLaughlin, A. C., & Bass, B. (2014). A multi-level analysis of the effects of age and gender stereotypes on trust in anthropomorphic technology by younger and older adults. *Ergonomics*, *57*(9), 1277-1289.
Pandey, A. K., & Gelin, R. (2018). A mass-produced sociable humanoid robot: Pepper: The first machine of its kind. *IEEE Robotics & automation magazine*, *25*(3), 40-48.
Perkins, L., Miller, J. E., Hashemi, A., & Burns, G. (2010). Designing for human-centered systems: Situational risk as a factor of trust in automation. Proceedings of the human factors and ergonomics society annual meeting,
Poole, D. I., Goebel, R. G., & Mackworth, A. K. (1998). *Computational intelligence* (Vol. 1). Oxford University Press Oxford.
Pozharliev, R., De Angelis, M., Donato, C., & Rossi, D. (2023). Do not put the blame on me: Asymmetric responses to service outcome with autonomous vehicles versus human agents. *Journal of Consumer Behaviour*, *22*(2), 455-467.
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.
Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. S., . . . Wellman, M. (2019). Machine behaviour. *Nature (London)*, *568*(7753), 477-486. https://doi.org/10.1038/s41586-019-1138-y
Rosenberg, S., Nelson, C., & Vivekananthan, P. S. (1968). A multidimensional approach to the structure of personality impressions. *Journal of personality and social psychology*, *9*(4), 283.
Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Pearson.

Russo, J. E., & Schoemaker, P. J. (1992). Managing overconfidence. *Sloan management review*, *33*(2), 7-17.
Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors*, *58*(3), 377-400.
Scheunemann, M. M., Cuijpers, R. H., & Salge, C. (2020). Warmth and competence to predict human preference of robot behavior in physical human-robot interaction. 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN),
Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2018). The constructive, destructive, and reconstructive power of social norms: Reprise. *Perspectives on psychological science*, *13*(2), 249-254.
Sevillano, V., & Fiske, S. T. (2016). Warmth and competence in animals. *Journal of Applied Social Psychology*, *46*(5), 276-293.
Sniezek, J. A., & Buckley, T. (1995). Cueing and Cognitive Conflict in Judge-Advisor Decision Making. *Organizational behavior and human decision processes*, *62*(2), 159-174. https://doi.org/10.1006/obhd.1995.1040
Sniezek, J. A., & Van Swol, L. M. (2001). Trust, confidence, and expertise in a judge-advisor system. *Organizational behavior and human decision processes*, *84*(2), 288-307.
Soll, J. B., & Larrick, R. P. (2009). Strategies for Revising Judgment: How (and How Well) People Use Others' Opinions. *Journal of experimental psychology. Learning, memory, and cognition*, *35*(3), 780-805. https://doi.org/10.1037/a0015145
Spatola, N., Marchesi, S., & Wykowska, A. (2021). Intentional and Phenomenal attributions in the light of the influence of personality traits, and Attitudes towards robots on pro-social behaviour in human-robot interaction.
Straub, I. (2016). 'It looks like a human!'The interrelation of social presence, interaction and agency ascription: a case study about the effects of an android robot on social agency ascription. *Ai & Society*, *31*, 553-571.
Sturgeon, S., Palmer, A., Blankenburg, J., & Feil-Seifer, D. (2019). Perception of social intelligence in robots performing false-belief tasks. 2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN),
Takahashi, H., Ban, M., & Asada, M. (2016). Semantic differential scale method can reveal multi-dimensional aspects of mind perception. *Frontiers in psychology*, *7*, 1717.
Tan, H., Wang, D., & Sabanovic, S. (2018). Projecting life onto robots: The effects of cultural factors and design type on multi-level evaluations of robot anthropomorphism. 2018 27th IEEE international symposium on robot and human interactive communication (RO-MAN),
Tanibe, T., Hashimoto, T., & Karasawa, K. (2017). We perceive a mind in a robot when we help it. *PloS one*, *12*(7), e0180952.
Thellman, S., De Graaf, M., & Ziemke, T. (2022). Mental state attribution to robots: A systematic review of conceptions, methods, and findings. *ACM Transactions on Human-Robot Interaction (THRI)*, *11*(4), 1-51.
Tversky, A., & Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science (American Association for the Advancement of Science)*, *211*(4481), 453-458. https://doi.org/10.1126/science.7455683
Tversky, A., Kahneman, D., & Slovic, P. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge.

Van Der Woerdt, S., & Haselager, P. (2019). When robots appear to have a mind: The human perception of machine agency and responsibility. *New Ideas in Psychology*, *54*, 93-100.

von Schenk, A., Klockmann, V., & Köbis, N. (2023). Social Preferences Toward Humans and Machines: A Systematic Experiment on the Role of Machine Payoffs. *Perspectives on psychological science*, 17456916231194949-17456916231194949. https://doi.org/10.1177/17456916231194949

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, *58*(301), 236-244.

Ward, A. F., Olsen, A. S., & Wegner, D. M. (2013). The harm-made mind: Observing victimization augments attribution of minds to vegetative patients, robots, and the dead. *Psychological science*, *24*(8), 1437-1445.

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in cognitive sciences*, *14*(8), 383-388.

Xu, X., & Sar, S. (2018). Do we see machines the same way as we see humans? a survey on mind perception of machines and human beings. 2018 27th IEEE international symposium on robot and human interactive communication (RO-MAN)

Xue, J., Niu, Y., Liang, X., & Yin, S. (2023). Unraveling the effects of voice assistant interactions on digital engagement: The moderating role of adult playfulness. *International Journal of Human–Computer Interaction*, 1-22.

Yaniv, I. (2004a). The Benefit of Additional Opinions. *Current directions in psychological science : a journal of the American Psychological Society*, *13*(2), 75-78. https://doi.org/10.1111/j.0963-7214.2004.00278.x

Yaniv, I. (2004b). Receiving other people's advice: Influence and benefit. *Organizational behavior and human decision processes*, *93*(1), 1-13. https://doi.org/10.1016/j.obhdp.2003.08.002

Yaniv, I., & Choshen-Hillel, S. (2012). When guessing what another person would say is better than giving your own opinion: Using perspective-taking to improve advice-taking. *Journal of experimental social psychology*, *48*(5), 1022-1028. https://doi.org/10.1016/j.jesp.2012.03.016

Yaniv, I., & Kleinberger, E. (2000). Advice Taking in Decision Making: Egocentric Discounting and Reputation Formation. *Organizational behavior and human decision processes*, *83*(2), 260-281. https://doi.org/10.1006/obhd.2000.2909 (Organizational Behavior and Human Decision Processes)

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of behavioral decision making*, *32*(4), 403-414

Zhang, Z., Chen, Z., & Xu, L. (2022). Artificial intelligence and moral dilemmas: Perception of ethical decision-making in AI. *Journal of Experimental Social Psychology*, *101*, 104327.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., & Dong, Z. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Appendices

Appendix 1
Table 1

The 67 Different AI targets Included In Study 1 (Chapter 2). Familiarity ratings of the 23 Most Popular AIs are indicated in Bold font.

no.	AI target	How familiar are you with this type of AI? (On a scale from 1 to 7)	Are you aware of AI like the above? (% of 'yes' answers)	Have you ever encountered or used an AI like the above? (% of 'yes' answers)
1	AI that provides navigation services	**5.54**	98%	100%
2	Chatbots used in customer service to answer noncomplex questions and provide information	**5.42**	100%	100%
3	Facial recognition AI used to open digital devices	**5.28**	98%	88%
4	AI that recommends movies, shows or series	**5.04**	94%	92%
5	AI that acts as a typing assistant that reviews spelling, grammar and corrects mistakes	**4.88**	94%	88%
6	AI that acts as a personal assistant taking voice commands (e.g., searches for Web information, orders products online, triggers events or plays movies and music when you ask it to)	**4.62**	90%	82%
7	AI that recommends music and artists to listen to	**4.52**	92%	90%
8	AI that connects you to potential friends/people you might know of	**4.38**	84%	82%
9	Avatars used in video games to represent different players	**4.18**	92%	72%
10	AI that recommends products or services to buy	**3.98**	80%	78%
11	AI that categorises emails in your inbox and offers quick - reply to options	**3.94**	78%	76%
12	Non-player characters (NPC) in video games. NPCs act as if they were controlled by human players but in reality, their behaviour is determined by artificial intelligence algorithms	**3.8**	76%	66%
13	AI that filter and organises the content on the news feed of social media sites	**3.72**	78%	78%

14	Drones	**3.52**	98%	50%
15	Autonomous robots that do vacuum cleaning in houses	**3.38**	92%	38%
16	AI that calculates credit scores for granting credit cards, loans, or mortgages	**3.3**	76%	64%
17	AI that matches people searching for a ride with potential drivers and also offers ridesharing services	**3.04**	78%	46%
18	Self-driving cars	**2.84**	96%	12%
19	AI that recommends people to go on a date with	**2.76**	72%	32%
20	Facial recognition AI used to identify potential suspects and conduct mass surveillance, which includes monitoring and tracking people	**2.74**	72%	26%
21	Avatars used in Internet forums, social media, and other online communities	**2.66**	64%	44%
22	Robots used in manufacturing (e.g., digitally operated robotic arms)	**2.54**	82%	14%
23	Domestic robots	**5.54**	98%	100%
24	AI that generates substantial passages of text in many different styles when prompted with a few initial words or lines	2.38	38%	24%
25	AI that creates paintings by being trained on thousands of paintings of different styles and aesthetics	2.38	40%	30%
26	AI used to predict someone's personality based on their data	2.38	48%	30%
27	Non-humanoid robots (= Robots that are not designed after the human body)	2.32	64%	14%
28	Pet-like robots developed for play	2.3	46%	26%
29	Robots developed for military purposes	2.2	58%	6%
30	AI used in medical diagnosis and treatment of diseases (i.e., in dermatology where AI algorithms are used to identify and propose treatments for skin lesions)	2.16	52%	18%
31	AI used in recruitment to screen candidates and identify the most qualified applicants	2.1	46%	18%
32	Teleoperated robots (=Robots that are operated remotely by a person, over the internet) used in harsh environments such as space or the sea	2.04	52%	6%
33	AI that provides financial advice (e.g., investment advice) or services (such as e.g., pension	2.04	34%	18%

	management or cash management services etc.)			
34	AI that generates photorealistic faces of people who never actually existed. These images are created by cleverly mixing features from large databases of actual faces	1.96	50%	20%
35	Healthcare robots made to assist with health management (e.g., monitoring blood pressure, detecting falls, and providing wellbeing advice)	1.94	42%	14%
36	AI that provides travel agent services (e.g., suggests holiday offers, hotels, books ticket and stays, organises car hires)	1.88	18%	14%
37	AI that automates low-value, repetitive back-office processes (such as i.e., the administration of benefits or the scheduling of interviews in an HR department)	1.88	20%	16%
38	AI used in video games to get feedback from a player's moves and techniques, and creates the landscape according to that	1.84	22%	16%
39	Robots used in robot-assisted surgery	1.74	56%	6%
40	AI used in education to produce personalised training based on students' learning pace and needs	1.72	22%	6%
41	Robots used as guides in public places such as shopping malls and museums	1.72	26%	6%
42	AI that composes music in a variety of music genres	1.7	22%	14%
43	Humanoid robots (= Robots that are designed after the full-human body). They are often referred to as anthropomorphic robots	1.7	64%	0%
44	Teleoperated robots (=Robots that are operated remotely by a person, over the internet) used for distant communication between people in business teleconferencing	1.7	18%	16%
45	AI used to predict performance (e.g., college students' performance, employees' performance, baseball players' performance)	1.66	18%	6%
46	Educational robots specifically designed to interact with children during their educational activities (i.e., robots that provide health education to children)	1.66	16%	12%

47	AI used in predictive policing to calculate where crimes are more likely to occur based on historical crime data	1.62	34%	10%
48	AI that provides personal styling services (e.g., discovers clothing for you and offers styling advice)	1.56	28%	6%
49	AI used in healthcare systems to predict health risk levels in the population and allocate resources accordingly	1.54	16%	6%
50	Androids (= Robots that strongly resemble the human outer appearance and are covered with flesh-or skin - like materials). When they possess male physical features, they are called androids whereas when they possess female physical features they are referred as gynoids	1.5	38%	2%
51	Healthcare robots made to assist with physical tasks (e.g., walking, fetching and carrying, and bathing)	1.48	26%	2%
52	Sex robots	1.46	46%	0%
53	Geminoids (= Robots built to look exactly like an existing person, but their behaviour is controlled by a human who teleoperates them)	1.4	32%	2%
54	Social robots (= Robots that are designed explicitly to interact with humans socially)	1.36	10%	4%
55	Facial recognition AI used in schools or university campuses to take attendance, permit access to facilities and monitor student behaviour, attention, and other emotional characteristics	1.36	18%	2%
56	Pet-like robots developed for therapy	1.36	16%	4%
57	Chatbots that deliver the onboarding process for new hires (e.g., walk new hires through the company's processes, answer questions)	1.32	6%	4%
58	Chatbots that provide legal services (e.g., drafting and submitting parking ticket claims, providing notary services, making tax appeals)	1.28	12%	2%
59	Healthcare robots for older people (= Robots used to help meet the healthcare needs of older people)	1.28	16%	2%
60	Companion robots in retirement homes	1.28	16%	2%

61	Healthcare robots made to assist with psychological issues (e.g., used in mental help therapy)	1.28	10%	4%
62	Telenoids (= Teleoperated robots with minimal human characteristics, usually ageless and genderless, but designed to have a head, torso, and short limbs)	1.26	16%	4%
63	AI that determines which job candidates will be successful based on analysis of video data (e.g., speech patterns, tone of voice, facial movements, and other indicators)	1.24	10%	4%
64	AI that helps you go through mental therapy	1.2	6%	2%
65	AI used in criminal justice to inform criminal sentencing decisions (i.e., parole decisions by predicting the risk of a defendant reoffending)	1.16	4%	0%
66	Teleoperated robots (=Robots that are operated remotely by a person, over the internet) used in telemedicine for specialist doctors to visit patients remotely in a hospital or at home	1.14	8%	2%
67	Robots for children with autism	1.08	4%	2%

Table 2

The 23 Most Familiar AI Targets Arranged In Descending Order of Familiarity (Study 1, Chapter 2)

no.	AI target	How familiar are you with this type of AI? (On a scale from 1 to 7)	Are you aware of AI like the above? (% of 'yes' answers)	Have you ever encountered or used an AI like the above? (% of 'yes' answers)	Examples given by participants (N=50)
1	AI that provides navigation services	5.54	98%	100%	*'Google maps', 'Apple maps', 'Sat Navs in cars', 'Strava', 'Map my run', 'TomTom', 'Waze- petrol stations and café nearby', 'Komoot'.*
2	Chatbots used in customer service to answer noncomplex questions and provide information	5.42	100%	100%	*'All customer service online chats on retail stores, utility companies, banks etc.', 'Amazon help chat, 'Most marketing websites, Virgin, Amazon, Lego, Curry's', 'Facebook chat bots', 'Twitter chat bots', 'Pretty much every single retailer has something like this: Amazon, Microsoft, Apple, Adobe'.*
3	Facial recognition AI used to open digital devices	5.28	98%	88%	'*Apple face ID', 'Microsoft Face recognition', 'Samsung Galaxy face recognition', 'banking app unlock'.*
4	AI that recommends movies, shows or series	5.04	94%	92%	*'Netflix', 'Amazon Prime', 'Apple TV', 'Disney+', 'BBC iPlayer', 'pretty much any recommendation on any streaming service'.*
5	AI that acts as a typing assistant that reviews spelling, grammar and corrects mistakes	4.88	94%	88%	*'Grammarly', 'Microsoft Word',' Apple Autocorrect', 'Microsoft office spell checker, sand pell checkers in other software',' Android keyboards e.g., SwiftKey, Gboard'.*
6	AI that acts as a personal assistant taking voice commands (e.g., searches for Web information, orders products online, triggers events or plays movies and music when you ask it to)	4.62	90%	82%	'*Alexa', 'Google Assistant', 'Siri', 'Google Home', 'Cortana', 'Amazon Echo', 'I use this feature on my Samsung phone a lot. Talk to text'.*
7	AI that recommends music and artists to listen to	4.52	92%	90%	*'Amazon music', 'Spotify', 'YouTube Music', 'Deezer', 'Tidal', 'Apple Music'.' Shazam'*
8	AI that connects you to potential friends/people you might know of	4.38	84%	82%	*'Facebook', 'Instagram', 'Snapchat', 'Twitter'.*

9	Avatars used in video games to represent different players	4.18	92%	72%	'Xbox games', 'Bots in shooter games like Pubg and CAll of duty', 'I would have to list every online video game I've ever played. For some examples that are at the top of my head: Diablo, Phantasy Star Online, Elden Ring, Monster Hunter Rise, Dark Souls.'
10	AI that recommends products or services to buy	3.98	80%	78%	*'Amazon, eBay and other online shopping sites', 'on supermarket websites, people who bought this also bought this',' suggested products on eBay, amazon prime, ads on social media', 'Things such as the Tesco Clubcard, M&S Sparks and Boots Advantage Card, they often tailor rewards and offers to you spending habit'.*
11	AI that categorises emails in your inbox and offers quick - reply to options	3.94	78%	*76%*	*'Gmail algorithm', 'My work inbox',' Outlook', 'Android mail', 'Out of office automatic response', 'yahoo mail'.*
12	Non-player characters (NPC) in video games. NPCs act as if they were controlled by human players but in reality, their behaviour is determined by artificial intelligence algorithms	3.8	76%	66%	*'NPcs in various games, the sims, runescape',' I've experienced them in games such as FFXIV', 'In games such as Grand Theft Auto, The Sims, Animal Crossing etc.', 'Dragon's Dogma, Diablo, Vampire The Masquerade Bloodlines, basically I would have to name every video game.'*
13	AI that filter and organises the content on the news feed of social media sites	3.72	78%	78%	'*Twitter feed', 'Reddit "popular" feed', 'Facebook', 'Instagram', 'YouTube', 'Algorithms on social media (Instagram and TikTok) that determine what comes up in your suggested posts/videos, Facebook newsfeed and 'people you may know' section', 'suggested posts on Instagram, news feed providing subjects you* have shown interest in or have talked about.'
14	Drones	3.52	98%	50%	*'Amazon delivery drones', 'Drones flown by individuals usually for photography/videography reasons', 'Delivery drones', 'surveillance drones', 'military drones', 'bombing drones',' Used by hobbyist but also for commercials uses such as photography', security, police investigations.*
15	Autonomous robots that do vacuum cleaning in houses	3.38	92%	38%	*'A vacuum cleaner that travels around the room in its own', 'Roomba', 'Vroom bot', 'eufy robot vacuum'.*
16	AI that calculates credit scores for granting credit cards, loans, or mortgages	3.3	76%	64%	*'Banking, credit card, mortgage, hire purchase applications', 'Experian', 'Clearscore', 'Money supermarket', 'credit karma', 'Money Supermarket Credit Club'.*
17	AI that matches people searching for a ride with potential drivers and also offers ridesharing services	3.04	78%	46%	*'Uber',' bolt', 'unicab',' Lyft', 'Swift',' bla bla car', 'rideshare'.*

18	Self-driving cars	2.84	96%	12%	*'Tesla', 'Google car', 'Heathrow Pods',' Audi TT assist'.*
19	AI that recommends people to go on a date with	2.76	72%	32%	*'Bumble', 'Hinge', 'Tinder', 'Plenty of Fish',' eHarmony',' Facebook dating', 'Grindr', 'Match', 'OKcupid'.*
20	Facial recognition AI used to identify potential suspects and conduct mass surveillance, which includes monitoring and tracking people	2.74	72%	26%	*'CCTV', 'Passport control', 'Airport security',' Government street and public transport cameras, passport control at airports, facial recognition on mobile phones.'*
21	Avatars used in Internet forums, social media, and other online communities	2.66	64%	44%	*'Facebook avatar', 'Snapchat bitmoji', 'twitter bots', 'Xbox 360','I've seen and used avatars on Facebook and on iPhone', 'chatbot for customer services', 'iPhone avatars', 'online games avatars'*
22	Robots used in manufacturing (e.g., digitally operated robotic arms)	2.54	82%	14%	*'Car manufacturing companies, Jaguar Land Rover', 'Used in vehicle and train manufacture, seen them on tv within programmes such as 'inside the factory', 'robots used in manufacturing such as on a car production line', 'I've seen videos from inside car factories'.*
23	Domestic robots	5.54	98%	100%	*''Carpet sweeper', 'Cleaning robots', 'Grass cutters hoovers', 'Hoover and lawnmowers', 'Vacuuming robots', 'iRobot Roomba', 'I have visited a restaurant where the meals are served by robots.*

Appendix 2
Table 1
List of the 15 questions used to assess AI literacy (Study 4, Chapter 4)

Question	Available multiple-choice options
What is an AI algorithm?	- a set of hardware components that enable computer programs to run - a set of instructions that can be implemented to perform a specific task - a specific type of coding language
Two important data sets for an AI algorithm include:	- Prototype and Launch - Training and Test - Weighted and Unweighted
What is the difference between machine learning (ML) and artificial intelligence (AI)?	- ML is a subset of AI - AI is a subset of ML - There is no difference
Smart speakers, like Alexa, use AI technology	- True - False - Not sure
The way AI learns is by consuming large amounts of data	- True - False - Not sure
When someone selects a show recommended by Netflix, they are engaging with AI	- True - False - Not sure
When someone unlocks their smartphone using face recognition, they are using AI.	- True - False - Not sure

AI technology is versatile, and a single AI program can perform many tasks, such as autocompleting sentences, controlling robots, etc.	- True - False - Not sure
AI can accurately identify common objects in an image at the level of an adult human	- True - False - Not sure
When someone uses an app like Facebook to tag a photo, they are engaging with AI technology	- True - False - Not sure