ARTIFICIAL MINDS AND REAL BELIEFS: PERCEIVING MENTAL STATES IN AI

by

OLIVER JACOBS

B.Sc., Queen's University, 2018

M.A., The University of British Columbia, 2020

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Psychology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2025

© Oliver Jacobs, 2025

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Artificial minds and real beliefs: Perceiving mental states in AI

submitted by Oliver Jacobs in partial fulfilment of the requirements for the degree of Doctor of Philosophy
in Psychology

Examining Committee:

Alan Kingstone, Professor, Psychology UBC
Supervisor

Lawrence, Ward, Professor Emeritus, Psychology, UBC
Supervisory Committee Member

Julie Robillard, Associate Professor, Neurology, UBC
University Examiner

James Enns, Professor, Psychology, UBC
University Examiner

Benjamin Bergen, Professor, Cognitive Science, University of California, San Diego
External Examiner

Additional Supervisory Committee Members:

Friedrich Götz, Assistant Professor, Psychology, UBC
Supervisory Committee Member

Abstract

Mental attribution refers to ascribing or perceiving mental states in others—be it people, animals, or even non-sentient targets like artificial intelligence (AI), Large Language Models (LLMs), and social robots. These mental attributions can be categorized by distinguishing between agentic (the ability to do) or experiential (the ability to feel) attributions using the mind perception framework. Using this framework, over the course of eleven experiments (2020-2024), I systematically investigate mental attributions toward AI and LLMs, how such attributions affect how people perceive human minds, and the way these effects vary between different individuals. After an Introduction in Chapter 1, Chapter 2 starts by providing a taxonomic structure to categorize the ongoing psychological work with LLMs to situate the work reported in the subsequent chapters and to provide a roadmap for organizing future psychological research with LLMs. In Chapter 3, I discover that people ascribe agentic and experiential attributions toward a wide range of robotic and AI agents, including LLMs. In Chapter 4, I investigate how loneliness can influence mental attribution toward LLMs, finding that loneliness, moderated by prior exposure, predicts greater experiential attributions but not agentic attributions. In Chapter 5, I demonstrate that the mind perception framework can be used to investigate how one views their own mind. Then, I examine if exposure to LLMs can influence how people view their own mind and which attributes people consider uniquely human. I discover that, after being exposed to LLMs, people increase their self-evaluations of agency and experience, while reducing their belief that these features of mind are uniquely human. In Chapter 6, I find that a forced-choice design, in contrast to an absolute numerical scale, yields greater preferences for human-generated art compared to AI-generated art. Collectively, across the eleven experiments, I demonstrate that individuals frequently attribute both agency and

experience to AI and that these attributions, in turn, affect how people perceive human minds—in themselves and in others.

Lay Summary

People tend to attribute mental states to others, including animals, artificial intelligence (AI), and social robots—a form of anthropomorphism. These attributions can be broken down into components related to agency (the ability to do) and experience (the ability to feel). Across the eleven experiments in this thesis, I find that people to tend to attribute agency and experience toward a range of social robots and AI systems. However, I find that these attributions vary considerably between individuals and is related to factors like age, prior exposure to AI, and perceived social isolation. I also find that attributing agency and experience to AI can influence how people reflect on their own minds and the minds of other people.

Preface

All work presented in this thesis was conducted from the Brain, Attention, and Reality Lab at the University of British Columbia with Professor Alan Kingstone as my supervisor. I was the primary author for each of the studies and chapters. I was responsible for the writing, materials, data collection, and visualizations. My co-authors provided suggested edits and conceptual guidance. The eleven experiments were approved by the Behavioural Research Ethics Board of the University of British Columbia [Toward a More Natural Approach to Attention Research: H10-00527; H22-00572]. Generative AI and similar writing tools were used solely for providing spelling/grammar suggestions and debugging data analysis code but were not used for novel text or idea generation.

A version of Study 2 (Chapter 2) is published in The International Journal of Social Robotics.

Jacobs, O. L., Gazzaz, K., & Kingstone, A. (2022). Mind the robot! Variation in attributions of mind to a wide set of real and fictional robots. *International Journal of Social Robotics*, *14*(2), 529-537.

A version of Study 5 (Chapter 5) is published in PLoS One.

Jacobs, O. L., Pazhoohi, F., & Kingstone, A. (2023). Self-discrepancies in mind perception for actual, ideal, and ought selves and partners. *Plos one*, *18*(12), e0295515.

A version of Study 6 (Chapter 5) is published in Consciousness and Cognition.

Jacobs, O. L., Pazhoohi, F., & Kingstone, A. (2024). Large language models have divergent effects on self-perceptions of mind and the attributes considered uniquely human. *Consciousness and Cognition*, *124*, 103733.

A version of Study 3 (Chapter 3) is also available as a preprint on PsyArXiv. Studies 1 (Chapter 2), 3 (Chapter 3), 4 (Chapter 4), and 7 (Chapter 6) are currently undergoing peer review.

Table of Contents

List of Tables

List of Figures

Acknowledgments

This dissertation would not have been possible without the immense support of those around me. First, I want to offer my deepest thanks to Alan Kingstone. From my very first moments in the Brain, Attention, and Reality lab to my very last, he has been an exceptional mentor. Whether it was his relentless passion for scientific research or his ability to see the bigger picture and promote well-being, Alan has been a role model in every sense. I could not have wished for a better supervisor.

I am also incredibly grateful for the unwavering support from my family, always cheering me on and having great faith in my ability to succeed throughout my graduate studies. I have been extraordinarily lucky to have my parents and siblings alongside me on this journey.

To my fellow lab members past and present—Kevin, Gray, Nicki, Farid, Leilani, Jake, Lucas, and Avi—thank you for being both collaborators and friends. I always enjoyed coming into the lab and seeing each and every one of you. Our conversations, whether quirky (like hours spent on the naturalistic fallacy), theoretical, or practical, made my graduate studies so much more enjoyable. My best friends outside the lab—Jessie, Anthony, Will, Chelsea—also deserve a heartfelt thanks for their support and aid in making my life outside academia richer.

Finally, I would like to thank my other co-authors on various projects not included in this dissertation as well as my wonderful committee members (Lawrence & Fritz) for their keenness, attention to detail, and insightful feedback. I was also fortunate that my research was graciously supported by grants from the Natural Sciences and Engineering Research Council of Canada, the Killam Foundation, Mitacs, and the University of British Columbia.

Chapter 1: Introduction

We live in a world full of emerging technologies that resemble realized science fiction from the past. One of these emerging technologies is generative artificial intelligence (AI). Large language models (LLMs) and their applications, such as ChatGPT, are at the forefront of generative AI. These systems are trained on billions of words to generate novel human-like text and images. LLMs have various use cases, including writing essays, solving math problems, providing feedback, and writing computer code. The widespread adoption of these models heralds a new age where millions of people interact and communicate with nonhumans in a way similar to how they might converse with a friend, therapist, coach, or life partner.

From a psychological perspective, the scientific study of generative AI can be insightful and consequential for several reasons. First, LLMs can be viewed as analogs or models of certain cognitive processes. This is valuable for probing the underlying mechanisms of topics such as the development and formulation of language, or decision-making under uncertainty. The ability of LLMs to simulate conversations also enables a lens to examine social interaction and questions of interest related to social cognition and group dynamics. Generative AI also raises questions related to ethics and human identity. For example, can LLMs and other generative AI technologies influence relationships or self-perception? Psychologists are also interested in generative AI for their practical applications such as their implications for mental health therapy or their potential efficacy in research pipelines.

The focus of the present thesis concerns one especially compelling direction: how humans perceive and attribute mental states to AI agents. Mental attribution is typically defined

as the capacity to reflect upon both one's own and others' mental states, including beliefs, desires, feelings, and intentions (Brüne et al., 2007). Despite the lack of consciousness or subjective experience, LLMs can invoke strong feelings of intentionality or intelligence, prompting us to question why we readily ascribe human-like qualities to non-sentient agents. This phenomenon touches upon long-standing questions related to mental attribution and how we infer cognition in others.

The Problem of Other Minds

One of the most quintessential problems in cognitive science has been the *problem of other minds* (Avramides, 2023). The problem starts with a simple premise: How does one come to believe that others have thoughts, feelings, or other mental activity? Without direct access, one cannot be certain that others have minds akin to their own mind. This problem is associated with solipsism or the deep skepticism regarding one's ability to know anything about the external world outside one's own mind. This form of radical doubt, or Cartesian doubt, is associated with Rene Descartes writings on being skeptical of all knowledge of the external world including any thoughts about other minds. Descartes' Second Meditation laid much of the groundwork for the ongoing debate around the problem of other minds alongside his famous conclusion 'Cogito, ergo sum' or 'I think therefore I am.' His establishment that his own existence is certain, unlike all else, highlighted the problem of other minds. His work continues to inspire work on the philosophical understanding of the other minds problem. However, modern philosophers like Thomas Nagel (1980) have turned away from the epistemological problem of other minds to focus on the conceptual problem: How can we understand the attribution of mental states to

others? In recent years, the problem has increasingly been understood by philosophers to be related to how knowledge arises through perception (Cassam, 2007; McNeill, 2012). This conceptual formulation of the problem becomes more addressable using empirical approaches to understand the psychological mechanisms at play.

Theory of Mind

In psychology, the tendency to bridge this problem and for people to attribute mental states to other people is referred to as theory of mind. As fundamentally social creatures, theory of mind lies at the heart of social functioning—permeating our daily lives in various interactions and activities. Theory of mind is evident in the language that we use regularly. For example, we will often use words when describing others like want, think, believe, feel, etc., that assume mental states without the direct observation of these states. In short, theory of mind is all around us and is integral to healthy social functioning. Conversely, deficits in theory of mind have long been associated with conditions related to social perception such as Autism Spectrum Disorder (ASD) and schizophrenia (Baron-Cohen et al., 1985; Corcoran et al., 1995; Frith & Corcoran, 1996; Garety & Freeman, 1999)

The origins of the term theory of mind are often associated with a 1978 paper by Premack and Woodruff called ‘Does a chimpanzee have a theory of mind?’ (Premack & Woodruff, 1978). Premack and Woodruff sought to investigate whether chimpanzees are capable of imputing mental states, regardless of the accuracy or completeness of the inference, in a way analogous to humans. Using a series of experiments based on the work of Köhler (1925), they investigated the problem-solving abilities of chimpanzees with the assumption that if a chimpanzee can make

sense of a human actor's intent or purpose and their knowledge or belief, this would provide evidence for a theory of mind over more associative explanations of the chimpanzee's behaviour. Premack and Woodruff stopped short of drawing strong conclusions based on the results of their experiment. Instead, they speculated, similarly to Darwin's views on the continuity between humans and animals (Darwin, 1853), that the physiological similarities between humans and animals are suggestive that theory of mind is not unique to the human species. Moreover, they suggested that most likely chimpanzees can attribute mental states related to wants, purposes, or affective attitudes but chimpanzees may be limited in their ability to impute states of knowledge.

Following the paper by Premack and Woodruff (1978), there was an increased interest in developing experimental tests to measure theory of mind, often by examining deception. Deceptive behaviours were considered an effective way of demonstrating theory of mind because deception requires a conceptualization of the deceived person's wrong beliefs (Wimmer, 1983). The most well-known method became the *False Belief* task which children typically pass beyond the age of four (Perner et al., 1987). In short, this test requires the test-taker to establish that another person's mental state, or knowledge, more specifically, is different from their own. A later version, called the Sally-Anne task, became the gold standard for evaluating theory of mind capabilities which involves testing whether a person can intuit whether another person will act on a mistaken belief about the location of an object.

At the same time as these developments, there was an increased interest in developing a better understanding of why people possess theory of mind and how it may arise in development. One influential framework addressing the *why* question of theory of mind, was the intentional stance framework presented by Daniel Dennett (1988). Following his earlier work related to intentionality inspired by the work of Franz Bertano, Dennett proposed that theory of mind can

develop out of the principles of information maximizing with minimal cognitive effort. Dennett proposed that understanding other people's behaviours is often optimized by concerning the intentions of other agents, unlike understanding the behaviour of natural or designed objects. These differences are the three stances put forth: the intentional, the design, and the physical stances. In essence, to understand how inanimate objects like rocks or rainbows behave, it is best to refer to a naïve or folk theory of physics. For things such as dishwashers or printers, it is best to predict how they behave by using information about how we think they are designed. And for people, the best approach is to understand how their behaviour is related to their thoughts, feelings, and desires. In other words, ascribing mental states renders a person's behaviour more understandable, and critically, more predictable. As Adam Waytz (2007) put it: "Mental states turn others' actions into something meaningful, orderly, and seeming comprehensible, an outcome that is deeply satisfying to perceivers so deeply motivated to understand."

Mind Perception

Following the renewed interest in mental state attribution, the 21st century began with a reevaluation of theory of mind. Several major problems were identified with the false belief task paradigm. First, the false belief task led to an all-or-nothing verdict. There are several reasons to be skeptical of conceptualizing theory of mind in this way. Consistent with an evolutionary perspective, there is likely some degree of theory of mind that animals are capable of following the continuity of humans and other animals. Furthermore, the influential findings by Baron-Cohen et al.(1985) suggesting that individuals with ASD displayed impaired false belief task performance failed to be replicated and, at best, was associated with arrested development rather

than an absence of passing the test (e.g., Gernsbacher & Yergeau, 2019). These findings highlight that theory of mind appears to be more of a spectrum than a binary trait.

Another significant issue is that mental attribution can appear in various forms, perhaps suggesting that mental attribution can occur along multiple dimensions. For example, Premack and Woodruff (1978) from the beginning believed that theory of mind could involve separate components for attributing desires or wants as opposed to attributing knowledge. Robbins and Jack (2006) argued for the existence of the phenomenological stance in addition to the three stances put forth by Dennett (1988). This phenomenological stance differed from the intentional stance by centering on the attribution of consciousness and qualia—features related to the explanatory gap of Chalmer's hard problem of consciousness (Chalmers, 2017). Later empirical research supported this dissociation between phenomenal and non-phenomenal mental attribution (Huebner, 2010).

The landmark study in partitioning types of mental attribution was a paper by Gray, Gray, and Wegner (2007). Gray et al. use the term mind perception, stemming from its use in Wegner's book *The Illusion of Conscious Will* (2003), to refer to the attributing, or perceiving, of mental states in others (and to a limited extent in oneself). The novelty of their approach was the use of factor analysis to partition explained variance among attributions of 78 pairwise comparisons on a five-point scale for 13 characters and 18 mental capacities. Gray et al. found that the rotated solution could explain 97% of the observed variance with just two factors. The first factor, coined Experience, was loaded with items related to capacities such as hunger, fear, thirst, pleasure, and pain. The second factor, Agency, was loaded with items related to self-control, morality, memory, and planning. These factors were associated with moral judgements and

appeared related to Aristotle's classical distinction between moral agents (Agency) and moral patients (Experience).

Another major observation in Gray et al.'s (2007) work highlighted a third limitation with theory of mind—at least in the way it was being measured. Humans are not the only target of mind perception. A beneficial aspect of their approach was that it demonstrated that the mind perception framework can be easily applied to probe anthropomorphism. Anthropomorphism, or 'seeing human in nonhumans', can encompass a range of perceptions and behaviours relating to both mental and non-mental features, with only the former being measured with mind perception scales or questions (Epley et al., 2007; Waytz, Cacioppo, et al., 2010). This mental component of anthropomorphism is also frequently referred to as psychological anthropomorphism (Thellman et al., 2022). Similar to theory of mind more broadly, the history of anthropomorphism is heavily related to philosophical treatises. Xenophanes, the pre-Socratic thinker, has been identified as amongst the first to note the dubious tendency for gods to bear a striking resemblance to their believers (Lesher, 2023; Waytz et al., 2010). While later scholars, such as Baruch Spinoza, focused on the irreducible problem of the Bible's anthropomorphism (Koch, 2021; Preus, 1995).

Modern empirical research into anthropomorphism is oft-associated with the early work by Heider and Simmel (1944) on apparent behaviour from moving shapes. Their study showed subjects a motion picture depicting the movements of three geometrical figures with the task that subjects ought to describe it. Regardless of whether specific instructions were given to interpret the movements in terms of human actions, almost all the subjects defaulted to this strategy to communicate the movements of the shapes. Mental attribution to the shapes was frequently reported such as with descriptions of the shapes being 'scared or frightened' or 'aggressive or

bullying'. In effect, Heider and Simmel's work highlighted previous ideas about the relative pervasiveness of anthropomorphism.

Much of the later empirical developments during the 20th century in the study of anthropomorphism coincided with technological developments following milestones like the Turing test, the creation of the ELIZA program, and more recently the proliferation of smartphones and computers in the 21st century. It was amidst these latter technological advancements that the mind perception framework emerged and began to spark general interest. While many computer scientists sought to develop more empathetic (or affective) computing systems (Picard, 2000), psychologists began to search for the psychological mechanisms underpinning anthropomorphism. For example, Waytz et al. (2007) provided a 3-factor theory citing elicited agent knowledge, effectance, and sociality as the principle reasons for anthropomorphizing. These reasons respectively are prior knowledge about a nonhuman agent, the need for predicting nonhuman agent behaviour, and the need for social connection. This latter rationale presupposes that the same mechanism by which people infer mental states in other people to foster social connection is likely also applied to nonhumans.

The popularization of Gray et al.'s 2-factor model of mind perception thereby emerged at a time of renewed interest in the study of anthropomorphism. The mind perception framework (or Mind survey) became a popular tool for investigating human-computer interaction and affective computing systems (e.g., Eyssel et al., 2012; K. Gray & Wegner, 2012; Wang & Krumhuber, 2018). For example, the framework was used to examine how a range of behaviours, human features, or manipulations to descriptions of robots can impact subsequent mind perception (e.g., Malle, 2019; Thellman, 2017). Critically, much of the early work using the mind perception framework to probe anthropomorphism toward machines concerned either relatively

primitive social robots (NAO; Eyssel & Kuchenbrandt, 2011; Kismet; Xu & Sar, 2018) or fictional robots (e.g., a future conscious robot; Gray et al., 2007). Even today, there are few social robots and most people have not had any exposure interacting with social robots. This absence of exposure is part of why the LLM era of the last few years is remarkable. Probing mental attribution to highly sophisticated non-sentient social agents is no longer purely hypothetical, but relatively few researchers have begun to focus their work on this new opportunity.

Terminology

Throughout this thesis, I occasionally use different terminologies to refer to mental state attribution toward AI and robots. A recent meta-analysis found that mental state attribution, anthropomorphism, mind perception, mentalizing, intentional stance, and folk psychology reflect subtle semantic differences but are often used synonymously, with different terms often being used in the same paper (Thellman et al., 2022). Typically, I refer to the broader component of mental attribution when referring to the phenomenon of ascribing or perceiving mental states in others while using the more specific term mind perception when discussing the Gray et al. (2007) 2-factor framework. Similarly, throughout this thesis I occasionally use AI and LLMs interchangeably despite their differences. I define AI as digital systems that can imitate intelligent-like qualities or features. LLMs are a form of AI but are much more specific. These differences are described in more detail in Chapter 2.

Despite efforts to avoid anthropomorphic language when discussing AI or LLMs, there is some anthropomorphic language used throughout (e.g., describing LLMs as having *human-like*

features). I also refer to LLMs as social agents—a contentious use of language that can be heavily criticized alongside any pseudo-agentic statements like what they can *do* or cannot do (Goddu et al., 2024). Nonetheless, I apply this for the sake of facilitating communication rather than to make claims about underlying sentience or mental states in any AI being discussed.

Thesis Overview

The basis for this thesis lies in the acknowledgment that recent developments in the field of AI and HCI have reshaped the study of mind perception by prompting a critical re-evaluation of the tools, methodologies, and, more broadly, the field's subsequent goals. This re-evaluation is being realized in light that the distinction between human and machine intelligence is becoming increasingly blurred with groundbreaking developments such as Large Language Models (LLMs). This backdrop of rapidly emerging technology offers a promising opportunity for re-examining quintessential problems in cognitive science related to mind perception. Seizing this opportunity, my thesis seeks to address the following guiding questions at the heart of these ongoing transformations:

1. To what extent do people perceive AI and robots as cognitive and phenomenal agents?
2. How does attribution of mind to nonhumans affect inter-and intrapersonal perceptions of mind?
3. What individual traits are most predictive of mind perception toward AI?

In Chapter 2, I present a novel taxonomy for understanding psychological research with LLMs to situate the following research concerning mind perception and AI. In Chapter 3, I

examine mind perception with regard to a diverse group of robots and computing systems in order to establish the range of mind perception targets (Study 1). In Chapter 3, I turn toward a specific focus on mind perception toward LLMs and the role of individual differences such as personality and social psychopathology (Study 2). I then investigate loneliness in the context of LLMs (Chapter 4). Chapter 5 focuses on the relationship between mind perception and inter-and intra-person perception. First, I examine whether or not the mind perception framework can be used to elucidate self- and other-perceptions of mind (Study 4). Next, I reexamine self-perceptions of mind in light of LLMs as a comparative other (Study 5). I then present a study highlighting the importance of methodological decisions involving comparative designs when it comes to perceiving and comparing AI-generated and human-generated creations (Study 6). Collectively, these chapters comprise 7 manuscripts (6 empirical studies) and 11 quantitative experiments.

Finally, I conclude with a discussion bringing together the insights garnered over each of the studies and what it means for the field of mind perception and HCI in regards to the questions proposed above. Implications, limitations, and future directions of the present research are also presented and discussed.

Historical Context for Data Collection

Data collection for each of the eleven experiments in this thesis took part from 2020 to 2024. The experiments are presented in a thematic order rather than a chronological order. The chronological order is: Experiment 1 (April 2020), Experiment 10 (April, 2020), Experiment 11 (November 2020), Experiment 7 (April 2022), Experiment 9 (March 2023), Experiment 2

(March 2023), Experiment 3 (April 2023), Experiment 6 (May 2023), Experiment 8 (October 2023), Experiment 4 (August 2024), and Experiment 5 (September 2024). Data was primarily collected using online recruitment platforms (cf. Experiments 2, 4) via MTurk, CloudResearch, and Prolific. The shift in which recruitment platform was used reflected growing evidence for the strength of CloudResearch and Profilic (Douglas et al., 2023). Differences between power analyses in experiments are in part explained by the specific analysis and the corresponding ease of use with either G*Power or WebPower (Zhang & Mai, 2018).

Chapter 2:

Categorizing psychological research with Large Language Models: AI-Centered, Human-Centered, and Tool-Centered approaches

The following chapter is a version of a manuscript submitted for publication. The aim of this review is to provide a taxonomic framework for organizing the different types of ongoing psychological research with LLMs.

Introduction

The introduction of ChatGPT and other Large Language Models (LLMs) has marked a significant turning point in the public's understanding and everyday access to highly sophisticated artificial intelligence (AI) with a wide range of applications. Amongst many other use cases, these models are highly capable of generating prose on virtually any subject; can analyze and run code; and can respond to images or generate their own (OpenAI, 2022.; Van Dis et al., 2023). Concurrent with the public's widespread adoption, has been an increased number of calls to urgently improve our understanding of how these models work and to investigate the potential consequences of these technologies in what has been oft-described as a 'paradigm shift' in AI and language processing (Dhar, 2024; Orrù et al., 2023). With these calls, there has been an abrupt rise in the scientific study of LLMs in human-computer interaction (HCI) and psychology in particular (Van Noorden & Perkel, 2023).

Situated among the calls for increased research efforts involving LLMs and AI more broadly, has been the recognition that the study of AI requires a collaborative interdisciplinary approach (Rahwan et al., 2019). The scientists that study AI behaviour have been the same scientists that have created these programs. However, the emerging reality is that the social, cultural, economic, and political consequences of this technology have become so large that it requires a broader, more intensive, effort from a wider range of scientists (Rahwan et al., 2019). Among these required actions from different disciplines is the need for psychologists to apply their domain of experience in probing human behaviour. The tools, methods, and theoretical knowledge developed over more than a century in psychology are essential for applying to LLMs to better understand their behaviour, and even to potentially better understand our own. These calls have been echoed in a variety of recent papers (Binz & Schulz, 2022; Kosinski, 2024; Rahwan et al., 2019).

The current published psychology papers involving LLMs have a wide range of goals and subtopics. Many of them aim to examine specific properties of LLMs or people's perceptions of them. These papers use a plethora of different psychological tools and methods, such as verbal tests or behavioural tasks with human participants (e.g., Dwivedi et al., 2023; Van Dis et al., 2023). There have also been broader review papers with various aims, such as seeking to provide future directions for research (Van Dis et al., 2023), providing specific tools or methods for conducting efficacious studies (Demszky et al., 2023), or raising awareness about important ethical considerations (Salah et al., 2023). There is, however, a lack of effort aimed at organizing the psychological research being conducted with LLMs. The present paper seeks to address this gap by presenting an organizational framework for categorizing the ongoing psychological research with LLMs before diving into comparisons between the approaches. The aim is to

provide a useful guide for understanding the landscape of psychology research involving LLMs, and by doing so, facilitate future research endeavours.

Classifying the Different Psychological Approaches

The basic proposal of this review is that the vast diversity in psychology research involving LLMs can be broken down into three distinct approaches or categories. These categories will be referred to as the AI-Centered, Human-Centered, and Tool-Centered approaches, which are distinguished by their principal questions, their research targets, and their measurements.

The AI-Centered approach can be thought of as research that involves treating LLMs as if they were a human participant in a psychological task and seeks primarily to examine the extent to which LLMs appear to manifest human cognitive abilities. The AI-Centered approach is most closely associated with cognitive psychology. The Human-Centered approach can be understood as research that explicitly aims to improve our understanding of humans' interactions with LLMs. This approach focuses on the perceptions and attitudes that arise from interacting with LLMs and is most closely aligned with social and personality psychology. Its key question concerns the degree to which people treat LLMs as cognitive agents and the repercussions thereof. The final approach, the Tool-Centered approach, contains studies that examine the practical applications of LLMs for subtopics related to psychology and human behaviour in certain environments, such as education and healthcare settings. The key question associated with the Tool-Centered approach is: To what extent can LLMs help us in psychologically meaningful ways? The Tool-Centered approach is not closely aligned with any subfield in psychology but rather draws from a mix of many subfields. Before diving into more detailed

explanations, examples, and comparisons between approaches, below are brief summaries describing key terms and the history behind the rise in LLMs.

Background

What are Large Language Models?

LLMs are a modern development in the field of natural language processing (NLP) that aims to understand and generate natural language including the contextual information inherent in language. The breakthrough nature of LLMs lies in several key developments starting with the focus on NLP development via statistical tools such as deep neural networks. These models are trained on vast amounts of textual information (e.g., books, websites, scholarly articles, etc.) and subsequently a large degree of human reinforcement training (i.e., fine tuning).

The seminal paper by Vaswani et al. (2017) introduced the specific neural architecture that serves as the backbone for many of today's most popular and powerful models. This transformer-based model introduced the idea of attention mechanisms (self-attention) which enable the model to selectively focus on different parts of the input from hierarchal levels of representation. Using this transformer architecture, the Generative Pre-Trained Transformer (GPT) family of models have since become amongst the most impactful in the development of LLMs. The importance of these attention mechanisms is that they enable the model to make the most out of its training to predict the next word in a set of text by varying weights on the importance of different words in a sequence rather than simply their position. For each predicted word, the entire section of the previous text is used in order to progressively build out responses with some level of randomness added to produce diverse outputs rather than purely deterministic patterns.

At present, the most popular AI chatbot is ChatGPT, which was developed by Open AI and uses variations of GPT models. A major breakthrough in these models was also that the sheer amount of data used for training led to monumental changes in the output and capabilities of these models—an idea that been referred to as "quantity has a quality all on its own" (Srivastava, 2022). This has led to a data arms race to feed as much data into these models as possible. The number of parameters (akin to the number of weights and biases within the network) in these models is often used as a heuristic for the power or strength of these models. The scale of these training sets and parameters is enormous. For example, estimates suggest GPT-3 was trained on 175 billion parameters while GPT-4 was trained on 1.76 trillion parameters (Wei et al., 2023). The basic costs to train and fit GPT-4 is estimated to cost a staggering US $3 billion (Metz, 2023).

It is important to note that while the architecture for these models is somewhat well understood and documented (e.g., for in-depth descriptions see Wolfram, 2023), there are many emergent properties of these models that are far less understood. In essence, while scaling-up models have been predictably shown to improve performance on a variety of outcomes, there are many emergent properties that larger models exhibit that are absent from smaller models (Wei et al., 2023). There is no smooth linear or quadratic relationship between the number of parameters and subsequent outcomes; rather, models' performance often show spontaneous or enigmatic jumps in performance with parameter increases. Understanding how these models work is an ongoing area of research (e.g., Hadi et al., 2023).

History of Chatbots

While the sophistication of LLMs and their chatbot applications are certainly new, the history of chatbots goes back to the earliest days of computing and is detailed in a number of

sources (e.g., Adamopoulou & Moussiades, 2020; Černý, 2022). The idea of developing chatbots capable of human-like reasoning dates back to Alan Turing (1950) and was proposed in his work detailing the Imitation Game (known today as the Turing test). The first breakthrough in a human-like chatbot was ELIZA, which was designed to respond to human inputs in the form of response outputs modelled after Rogerian psychotherapists (Weizenbaum, 1966). ELIZA was built on the basis of decomposition rules which are triggered by keywords in the input. These keywords then form responses via reassembly rules. One of the most remarkable findings about ELIZA was that despite the relative simplicity of these reassembly rules, its responses were often surprisingly human-like. Many users of ELIZA even struggled to be convinced that ELIZA was not a human responder (Weizenbaum, 1966).

A number of chatbots were soon developed heavily inspired by the success of ELIZA (e.g., PARRY; Schank & Colby, 1973). Some of these next-generation chatbots aimed to imitate human players in video games such as Chatterbot (Mauldin, 1994) with all these chatbots employing more complex pattern-matching capabilities. With the advent of the Internet, online chatbots were developed soon after often with direct inspiration from earlier programs like ELIZA. Some of the notable chatbots include ALICE (Heller et al., 2005) or SmarterChild, which was made available by AOL and Microsoft's MSN in 2001. These chatbots were a noticeable departure from earlier chatbots, not only by virtue of their capabilities, but also in their specific aims in interactions with people. These chatbots were not simply built to test previous thought experiments, but instead, they sought to provide additional information for users such as information about movie times, the weather, and up-to-date information about the news (Adamopoulou & Moussiades, 2020). This trend in the early 2000s for more pragmatic chatbots continues to this day.

After the success of some of the earlier chatbots and growing capabilities in AI, the first mainstream widespread chatbots came to life with the smartphone era and personal assistants built into the devices. These assistants such as Apple's Siri, Google Assistant, and Amazon's Alexa, were also marked by their integration with the rise of the Internet of Things (IoT) as they became commonplace in homes and work environments in the early 2010s. While these chat applications provided unprecedented levels of utility compared with prior generations of chatbots, there are several limitations associated with these assistants. For example, they were limited in their ability to take into consideration context from prior inputs, unlike the next generation of chatbots built using transformers that can pass loops of knowledge from one part of the network to the next (Vaswani et al., 2017). This characteristic of being able to maintain context in a human-like way became a significant breakthrough for modern transformer-based LLMs such as the GPT models, which are the core of the most advanced chatbots available today.

The AI-Centered Approach

Defining AI-Centered

The AI-Centered approach to psychological research on LLMs is characterized by focusing on LLMs in a way analogous to the way psychologists focus on testing humans. It is an approach that treats the LLM as if it were an idiosyncratic human and principally seeks to explore the extent to which LLMs appear to manifest cognitive abilities. In this way, it is very similar to how cognitive psychologists or clinical neuropsychologists would use different tests to probe aspects of one's cognitive abilities. That is, there is a common practice of comparing subject performance with healthy controls on a wide variety of psychological instruments. Thus, the AI-Centered approach can be described as an approach that borrows tools from cognitive psychology. Consequently, many of the research methods and tools stemming from the AI-Centered approach overlap with common assessments done during routine psychological tests or examinations such as the Wisconsin Card Sorting Task, the Go/No Go task, the Stroop task, and many others. While it may seem paradoxical or nonsensical to say that the aim is to 'probe the psychology of LLMs', it does serve as a useful metaphor. The idea of treating LLMs as a 'participant in a psychology experiment' has been previously mentioned (e.g., Binz & Schulz, 2023; Ritter et al., 2017; Shiffrin & Mitchell, 2023) but the idea of classifying a broader range of studies under this aim (even if it is not explicit in the studies themselves) is novel.

Benchmarks

The capabilities of different AI and computing technologies have a long history of comparisons with classic benchmarks. The Turing test (or Loebner Prize) is an example of an

early but long-standing benchmark. Some benchmarks have taken the form of goals on specific tasks such as the ability to outperform humans on certain games like chess, checkers, or Go. Some benchmarks have taken the form of specific challenges such as the ImageNet database for object recognition (Krizhevsky et al., 2012). Benchmark comparisons are also routinely touted with the debut of new LLMs. OpenAI frequently publishes papers establishing the performance of their models on a variety of tasks such as math Olympia problems or more general mechanisms like sample efficiency and generalization (Cobbe et al., 2021). Some of these benchmarks also take the form of popular simulated exams such as the LSAT, SAT, or AP subject exams (Open AI, 2023). There are also larger batteries of task performance such as the BIG-bench (Beyond the Imitation Game benchmarks) which consists of at least 204 tasks from 450 authors across 132 institutions (Srivastava et al., 2022). Most of these benchmarks overlap with different types of cognitive testing that would fall under the AI-Centered approach. Nonetheless, it is worth noting that these types of benchmarks differ from those found in the psychology literature which tend to use more specific instruments for measuring individual constructs. These differences should become clear as the research summarized below focuses on more specific psychological features.

A Note of Caution in Anthropomorphizing LLMs

Before describing the initial findings of many of these AI-Centered studies, it is important to consider a note of caution regarding the anthropomorphizing of LLMs. Anthropomorphism is defined as the transference of human qualities, both mental and non-mental, to nonhumans (Epley et al., 2007, 2008; Waytz et al., 2010a). Some research from the AI-Centered approach can be viewed as a form of anthropomorphism. However, much of this research takes a

conscious, direct decision in this aim, and that it need not be an example of reification. Rather, it is a deliberate use of humans as a model for understanding LLMs, both in the attempt to better understand LLMs and their emergent properties, and an attempt to better understand human psychology as well (see Trott et al. (2023) for a neat example). The challenges associated with this approach are discussed more thoroughly below.

Psychological Constructs Studied in LLMs

Intelligence

Research on machine intelligence has the longest and most extensive history of all the cognitive capabilities studied in machines and AI systems. Many key benchmarks as previously described are directly related to intelligence. For example, Open AI's published benchmarks on a variety of knowledge tests all resemble a sort of intelligence testing with some having a clearer focus on ability or knowledge, and others focusing more on aptitude or general problem-solving ability.

Many of the efforts by psychologists to understand the intelligent-like behaviour seen in LLMs have been in the pursuit of examining the degree to which general problem-solving can be observed with these models. A program that can regurgitate information from a dictionary could be seen as containing a depth of knowledge, but for obvious reasons, is not of interest to psychologists and broader communities, except as a convenient reference. As such, many of the most prominent studies have aimed to understand the ability of LLMs to exhibit critical thinking, decision-making, and reasoning features.

Binz and Schulz (2023) sought to investigate the causal reasoning abilities of GPT-3 using a series of vignettes from popular psychology literature along with popular decision-making tasks, information search problems, and deliberation problems. Some of these vignettes are very well known, such as the 'Linda problem', which illustrates the conjunction fallacy, or the 'cab problem', which illustrates base rate neglect (see Tversky & Kahneman, 1983). Binz and Schulz (2023) were aware of the issue of popular vignettes being in the training data that GPT-3 was trained on—an issue sometimes referred to as *leakage* (Dillion et al., 2023). Nonetheless, the overall ability of GPT-3 to solve reasoning problems correctly, or sometimes incorrectly as humans do, came as a surprise and a pertinent marker of the amount of progress LLMs have made.

There have been many other research efforts with similar aims as above, often reaching analogous conclusions. For example, Orrù et al. (2023) found that ChatGPT performed comparably with human subjects on a set of practice and transfer word problems. Another example includes Loconte et al. (2023), who used a battery of cognitive tests to assess ChatGPT. They found that in comparison to humans, ChatGPT scored relatively normal on verbal reasoning tasks, exceptional on a sentence completion task, and remarkably poorly on the Tower of London task which involves planning capability. Loconte et al.'s (2023) findings supported other research (e.g., Srivastava et al. 2022) that the verbal performance of ChatGPT is very high but more specific behavioural tasks highlight that there is a discontinuous profile of the intelligence capabilities of ChatGPT, such that high performance in one subcategory or measure of intelligence is often quite different from performance on another.

The initial research exploring intelligence in LLMs has thus revealed a number of key findings. First, the idea of machine intelligence imitating human intelligence is not exclusive to

science fiction. Rather, LLMs are currently capable of performing similarly to humans on a range of intelligence-related tasks, including more elusive reasoning tasks that historically have been more challenging for AI to master. Second, the intelligent-like behaviour exhibited by LLMs remains remarkably inconsistent and somewhat unpredictable. Strong performance in some domains of intelligence when compared with humans, like inhibition, are less strongly related to performance on other types of executive functions like planning.

Simply Mimicking Intelligence or Is There Something More?

An enormous elephant in the room when discussing intelligent behaviour in machines is the large debate regarding whether performance on these tasks can equate to intelligence. Once again, this introduces the philosophical controversy of whether or not a machine can have real intelligence which harkens back to ELIZA, Turing machines, and the famous Chinese-room thought experiment by John Searle. In short, Searle's thought experiment (1980) proposed the case of a situation in which an English-speaking monolingual person is enclosed in a room being fed Chinese characters through a slit in the wall. Given the right tools in the room (e.g., pen, paper, and a translation dictionary), the English-speaking person would be able to pass interpretations of the Chinese text out of the room. From the outside, it would appear that the person inside the room was fluent in Chinese despite their inability to understand Chinese. This situation can be said to mirror intelligence perceived in AI in that the problem with interpreting intelligence in AI is that it might simply be capable of manipulating symbols with no underlying understanding of the human-like systems being imitated. The parsimonious account of AI intelligence could thus be that AI systems are simply manipulating symbols—there is no underlying knowledge or understanding akin to humans—a rejection of the physical symbol

system hypothesis positing that a physical symbol system has all the necessary components for general intelligent behaviour (Newell, 1980).

It should be noted that there are eminent criticisms of the Chinese room (e.g., Chalmers, 1992; Dennett; 1991). Chalmers (1992) for example has stated that it is clearly consciousness at the heart of the matter in the Chinese room experiment. And in that vein, Dennett has stated that the Chinese room is simply a reformulation of the *other minds* problem constituting a user illusion. How do we ever come to believe that anyone else is conscious without having direct access to their mind? AI can similarly be *the other mind* and, in that sense, there is nothing special about the absence of direct access to its behaviour. Thus, just as we do with other humans, the parsimonious approach would then be—if the AI is simulating consciousness or intelligence to such a high standard—we should rightly call it intelligent (Nilsson, 2007).

The modern debate regarding this issue is often framed as questioning the degree to which LLMs like ChatGPT are simply *stochastic parrots* or fancy autocorrectors (Bender et al., 2021). There is similarly no clear consensus. Some researchers hold firm that the design of LLMs and their outputs are simply a probabilistic pattern of the most likely word to occur next (with some randomness—or temperature), meaning that they can not have any true understanding (e.g., Bender et al., 2021; Borji, 2023). Others, like Geoffery Hinton, the most influential scientist in the developments of NLP leading to LLMs, have come to the opposite conclusion that much of the rationality and critical thinking that emerges from these models contain all the hallmarks of the same rational and critical thinking that exists in humans (Chalmers, 2023; Rothman, 2023). This latter camp is quick to quip—are we all that different from being a stochastic parrot ourselves, and are human minds not black boxes too?

Dissimilarities with Human Cognition

Given that LLMs are capable of performing well on different sorts of cognitive tasks and intelligence testing, it makes sense to compare them with human minds even without trying to answer whether or not AI can truly be described as intelligent. Then, one can compare how the apparent intelligence differs from that of humans more broadly. There are, after all, critical differences in how LLMs operate or appear intelligent compared to humans.

First, humans are continuous learners—LLMs do not learn past their training period. Humans are also much more efficient in their learning. Children in the US hear about 10-30 million spoken words by the age of 2 which pales in comparison to the number of words fed into GPT3.5's—which is estimated at 114 billion (Demszky et al., 2023). Granted, infants are fed much more contextual information alongside words as fundamentally social creatures (e.g., intonation, gestures, etc.). This discrepancy in language exposure suggests that sufficient amounts of language exposure do not translate into LLMs performing similarly to humans on many abilities, such as theory of mind (Trott et al., 2023).

A second major difference is that human brains are significantly more energy efficient. Some estimates of energy consumption for adult brains runs between 10-20 watts of power (Jorgensen, 2022; Song, 2023) while estimates for ChatGPT's training are around 10 gigawatt-hours, and its daily use of around 1 gigawatt hour, which is equivalent to 333 000 and 33 000 average US household's daily energy use (McQuate, 2023). Of course, these comparisons are an oversimplification and there is no direct comparability of energy cost differences between brains and LLMs. Nonetheless, creating AI that can achieve human-like intelligence without the vast discrepancies in energy requirements would require an entirely new approach over the current

NLP methods and has been discussed as an ongoing area of research for some influential scientists (Rothman, 2023).

There are also substantial differences in the goals between humans and LLMs. Humans are motivated to meet their metabolic needs and our cognition can be understood as a means to interact and navigate the world around us. The latter point highlights that human minds are embodied, unlike LLMs (Chemero, 2023). For decades, there has been a growing recognition that cognition is deeply rooted in bodily interactions with the environment (Varela et al., 1993; Wilson, 2002) and sensorimotor experiences are what grounds cognition (Barsalou, 2008; O'Regan & Noë, 2001). This movement evolved out of a rejection of previous dualist thought and the rise of computationalism as a means for understanding mental features (Chemero, 2013). Critically, this movement has long criticized the physical systems hypothesis (De Vega et al., 2008). This criticism of disembodied LLMs also overlaps with Searle's initial interpretation that a physical body in which the symbol manipulation occurs was the difference between the room and a person, and that even an embodied robot that had 'eyes' would not be truly seeing what comes into its eyes—an appeal to qualia or the ineffable feelings associated with experiencing. Again, these debates have deep philosophical implications that extend far beyond this paper (Avramides, 2023; Cole, 2024). Nevertheless, it is important to state that the embodied nature of human cognition is a major dissimilarity with the disembodied nature of LLMs. This difference may be short-lived as researchers and developers increasingly look toward connecting sophisticated robots to LLMs (Addlesee et al., 2024; Mishra et al., 2024).

Personality

The literature aiming to measure and interpret the personality traits exhibited by LLMs is diverse and growing (Jiang et al., 2023; Miotto et al., 2022; Pan & Zeng, 2023; Rao et al., 2023; Rutinowski et al., 2023; Serapio-García et al., 2023). This line of research follows the goal of chatbot developers to build chatbots capable of exhibiting unique personalities—personality being defined as a set of characteristics that form an individual's pattern of thoughts, behaviours, and traits (Roberts & Yoon, 2022).

Some of the initial research applying personality frameworks to LLMs includes the work by Miotto and colleagues (2022) who used the Honesty-Humility, Emotionality, eXtraversion, Agreeableness, Conscientiousness, and Openness (HEXACO) personality framework to better understand the personality profiles of GPT-3. They found through sampling a wide array of temperature settings that the HEXACO scores of GPT-3 were highly dependent on the specific temperature values but there was notable consistency in personality scores within temperature settings. They also found that in general, GPT-3 scored slightly higher in honesty-humility and slightly lower in emotionality facets compared to human samples.

Another direction in investigating personality in LLMs is the ability of LLMs to take on various personas with a range of personalities, a belief that Jiang et al. (2023) sought to examine explicitly. Jiang et al. (2023) used the 44-item Big Five Inventory (BFI) to assess GPT-3.5's ability to create content based on curated personality profiles. These LLM personas were asked to write an 800-word childhood story to see if the language reflects the assigned attributes (Jiang et al., 2023). For example, more conscientious people are more likely to use words related to achievement and money, and more neurotic people are more likely to include negative emotion words and words related to mental health. These findings were reflected in the LLM-generated

stories suggesting that the personas created by the LLM are able to consistently express an assigned personality profile.

A third major finding from investigating personality traits in LLMs has been that the specific prompts used with LLMs can greatly influence the desired dimensions to mimic specific personality profiles (Safdari et al., 2022). This ability to impersonate different personality profiles is also strongly related to the specific fine-tuning of the model in question with more powerful models with large training sets providing greater reliability in simulating personality profiles (Safdari et al., 2022). In summary, there is a growing body of evidence suggesting that LLMs are capable of imitating a variety of personality profiles measured with some well-validated personality frameworks such as the Big 5 or HEXACO. Moreover, this research has demonstrated that the ability to imitate different personality profiles depends on the strength of the model, the specific prompts, and temperature settings.

Creativity

While the body of psychological research involving creativity is comparably much smaller than some other topics like personality and intelligence; creativity is a particularly interesting construct for investigation in LLMs as people's kneejerk response is to reject the idea that machines are truly capable of displaying creativity. Creativity has been described as a fundamental feature of human intelligence and an inescapable challenge for AI (Boden, 1998). It is rarely afforded in anthropomorphic statements even with a range of animals (Kaufman & Kaufman, 2004). Furthermore, even with the popularity and global recognition of LLMs, there is a seeming reluctance to describe programs like ChatGPT as creative. However, the conventional definition of creativity as being a skill of bringing about something new and valuable (Runco &

Jaeger, 2012; Young, 1985) suggests that machines can be creative, as AI's founders have discussed (Boden, 2009).

There are a number of popular methods for assessing creativity in humans. Some of these include the Torrance Tests of Creative Thinking (TTCT: Lissitz & Willhoft, 1985), the Runco Creativity Assessment Battery (RCAB; Runco & Jaeger, 2012), and the Alternative Uses Test (AUT: Guilford, 1967). In a recent study, Stevenson et al. (2022) used the AUT to investigate the creativity of GPT-3 against university undergraduates for responses to "Think of many creative uses for…" substituting the final word for a book, a fork, and a tin can. Blind judges were used to assess each response on originality, utility, and surprise using a 5-point Likert scale. Hierarchal regression models found that humans scored higher on originality and surprise but lower on utility. The authors concluded that humans still outperformed GPT-3 on the creativity measures that were assessed but the relative gap suggested that it was only a matter of time before GPT catches up and likely surpasses average human performance.

Interestingly enough, a short time later GPT-4 was investigated by Guzik et al. (2023) using the TCTT. The TTCT test subdivides creativity into several facets pertaining to fluency, flexibility, and originality scores. Performance between GPT-4 was again compared with a sample of undergraduates. This time GPT-4 scored in the top 1% for originality and fluency while displaying above average scores for flexibility. As Stevenson et al. (2022) predicted, it was only a matter of time before GPT's performance matched and exceeded human-level performance.

Limitations of the AI-Centered Approach and Fundamental Challenges

As with all the approaches to psychological research into LLMs, there are a number of limitations and challenges associated with the specific approach. As mentioned, the inherent anthropomorphism with the AI-Centered approach is a double-edged sword where it is best to avoid ascribing too much humanness to LLMs even if they appear to respond with similar or even identical answers to humans on a variety of psychological tasks. There are several famous examples illustrating this point such as the story of Clever Hans. In short, observer-expectancy effects are critical to consider for assessing the internal validity of studies, particularly in the AI-Centered approach.

Another analogous idea to the issue of overprescribing complex behaviour is the tendency to equate human and LLMs when they deliver similar responses. Even when behaviours are identical, there is often no reason to think the underlying processes are at all similar. We might save or store a friend's birthday in our head, but it would be naïve to believe that is the same underlying process as saving a date on a computer. Again, we must be careful not to believe that similar performance of LLMs to that in human experiments is an indicator that the processes underlying the outcome are at all similar. This is the fallacy of equivocation.

In general, many of the limitations of the AI-Centered approach are measurement and translation-related. The different models, temperature settings, and specific prompts add complexity to the already complex array of different tasks and tests used to assess features of human psychology in LLMs. On the other hand, the AI-Centered approach of examining LLMs through a psychological lens provides new perspectives on well-established psychology theories and can challenge their underlying assumptions.

The Human-Centered Approach

Defining Human-Centered

The Human-Centered approach is characterized by its focus on individuals and their interactions with LLMs. In this sense, it easily slots into the field of human-computer interaction (HCI) and offers psychologists one of the best opportunities for applying their domain knowledge given that the subject of study is routine. The Human-Centered approach encompasses the vast number of influences that LLMs can have on human behaviour. These can include effects specific to LLMs such as people's perceptions and attitudes toward these systems. They can also extend to include a wider range of influences such as how LLMs may or may not influence social cognition. In comparison to the AI-Centered approach, there is far less history and ongoing research with LLMs aligned with the Human-Centered approach, as the approach is less intimately tied to the history of AI research and requires human interaction with such programs. Nonetheless, we can draw from past research on more limited AI systems. Below we discuss some promising areas of research starting with subtopics with larger roots in past HCI work and moving toward more exploratory work specifically with LLMs.

Anthropomorphism

The study of how people come to anthropomorphize nonhumans, including robots, machines, and AI systems is, in itself, a literature with a wide scope and a long history (e.g., see Waytz et al., 2010b). One method of studying anthropomorphism toward robots and AI systems with widespread use is the mind perception framework (e.g., Stafford et al., 2014; Wiese et al., 2017). Mind perception refers to the degree to which people attribute capabilities or capacities of mind to a wide variety of agents, both human and nonhuman (e.g., animals, robots, AI systems). It is typically assessed using attitude scales comparing several different agents (i.e., the Mind

survey; Gray et al., 2007; Waytz et al., 2010b). Although there is some disagreement about the optimal number of factors (Malle, 2019; Tamir & Thornton, 2018; Tzelios et al., 2022; Weisman et al., 2017), the original 2-factor model developed by Gray et al., (2007) remains in popular use. This 2-factor model, derived from factor analysis, found two principal components: agency, encapsulating the ability to think, do, and act morally; and experience, encapsulating the ability to feel emotions, pain, pleasure, and drives.

Much of the past work involving mind perception has found that several contextual factors influence the degree to which people anthropomorphize, or attribute agency and experience to robots and AI systems. For example, robots that appear more human-like tend to elicit greater anthropomorphism (Müller et al., 2021) and robots that tend to behave more human-like similarly elicit more anthropomorphism (Sacino et al., 2022). Essentially, more capable or more human-like machines elicit greater attributions of agency and experience (Waytz et al., 2010b). Given their recency, there is far less research into contextual factors affecting mind perception toward LLMs compared to the literature on anthropomorphizing social robots.

A recent exception was a study by Jacobs et al. (2023) that specifically examined mind perception and LLMs using Gray et al.'s (2007) framework. They found that people tend to attribute greater agency and experience to ChatGPT after being exposed to it. Other research has supported this result, finding that interaction with LLMs can influence anthropomorphic attitudes (Laban et al., 2024b; Wang et al., 2024) and attributions of phenomenal consciousness in particular (Colombatto & Fleming, 2024). In summary, the initial research concerning anthropomorphism and LLMs has suggested similar patterns to human interactions with social robots but the factors driving anthropomorphic tendencies toward LLMs remain a direction for future research.

Trust

With the growing capabilities of LLMs, and AI more generally, there has been widespread effort to align human goals with these newly developed technologies. For example, the European Union has announced initiatives for safely developing LLMs (European Commission, 2024), and the UK and Canadian governments have invested large sums into research programs for studying the safety and ethical considerations of AI applications (Government of Canada, 2024; UK Government, 2023). Chief among these considerations is the degree to which people can trust AI systems and the companies producing these technologies. In comparison to other subtopics in the Human-Centered approach, there is a large body of empirical research investigating trust toward AI, and some with LLMs specifically.

There have been a number of different tools developed to measure trust, and different camps exist regarding how trust should be conceptualized. Some have argued that trust can be measured by a single underlying factor that encompasses trust and distrust as opposite ends on the same spectrum (Jian et al., 2000). Meanwhile, others have argued that distrust represents a unique construct that can co-exist with varying degrees of trust (e.g., one can trust a technology to produce a reasonable response while fearing that it may not be acting in your best interest) (Govier, 1994; Lewicki et al., 1998). One popular scale that measures trust as a unidimensional construct includes the Trust between People and Automation (TPA) scale (Jian et al, 2020). This scale, however, was designed without the specific goal of measuring trust in AI systems, which has subsequently been addressed in several ways. The simplest way has been to re-word items on the scale to include the word AI, but this has been shown to influence the underlying trust perceived in the system in question (Langer et al., 2022). A more recent unidimensional scale called the Trust eXplainbility Accountability and Invasiveness inventory (TXAI) was constructed

with the specific aim of measuring trust in AI (Hoffman et al., 2023). This scale has also been adopted to measure trust specifically toward LLMs (Perrig et al., 2023). However, it remains to be seen whether trust toward LLMs will be operationalized using this scale or some future method.

Despite the inconsistent operationalization of trust, and issues regarding the psychometric properties of measurements, there has been recent work investigating user trust toward ChatGPT. For example, Choudhury and Shamszare (2023) surveyed frequent users of ChatGPT using their own scale for trust, with many items adopted from the TXAI. The researchers found that ChatGPT use was positively correlated with trust. Other studies have taken a less direct approach, and instead focused on the repercussions of trust in LLMs. A study by Ye et al. (2023) demonstrated that incorporating ChatGPT into a physical robot led to a more natural and intuitive human-robot interaction, and subsequently increased trust in the competency of the robot. Baek and Him (2023) demonstrated that increased task efficiency and personalization led to increased perceptions of trust.

One interesting takeaway from recent work examining user trust in LLMs is that users tend to exhibit high levels of trust despite the widespread knowledge that LLMs can be quite unreliable (Shen et al., 2023). LLMs are infamous for their tendencies to ‘hallucinate’. That is, they often generate responses that appear coherent but have no sensical or factual basis. Nonetheless, users tend to trust the output of LLMs which, when unfettered, can be of concern. Particularly, since the trust research that has been conducted so far (Choudhury & Shamszare, 2023), appears to show that trust in a variety of applications is highly related. In other words, the unidimensional nature of trust suggests that users trust ChatGPT similarly regardless of the

consequences of the information that they are seeking. The potential risks this has for certain issues, like users seeking legal, medical or financial advice, are profound.

Individual Factors

One popular subject in the Human-Centered approach is the focus on individual factors. For researchers interested in examining attitudes toward AI, and LLMs specifically, a natural question emerges: What are the psychological factors that drive these attitudes? This was an explicit question asked in a recent review (De Freitas et al., 2023) that put forth five key interrelated factors regarding AI. Although this review took a more a more AI-Centered approach, the factors identified were how opaque, emotionless, inflexible, autonomous, and nonhuman AI is. For example, the autonomous factor relates to the degree to which people can control their interactions with AI systems. People prefer the freedom to dictate their own choices and find tasks with more choices more enjoyable (e.g., self-determination theory; Ryan & Deci, 2006). This in turn would suggest that attitudes toward LLMs will be influenced by the degree of control users have in shaping LLM responses.

Other research has focused on individual factors through a more Human-Centered approach by examining internal traits that differ in the population. Returning to the example of anthropomorphism, a seminal paper by Waytz et al. (2007) sought to provide a theory for understanding which individuals are more likely to ‘see human in nonhumans.’ Waytz and Cacioppo (2007) put forth the idea that people are more likely to anthropomorphize when anthropocentric information is available and applicable, when motivated to be social agents, and when lacking social connection to other humans. These factors in turn lead to notable individual differences, as people vary in the degree to which these factors shape their behaviour. For

example, there are large differences in how much one might feel socially connected. A scale was later developed by Waytz et al. (2010a) called the Individual Differences in Anthropomorphism Questionnaire (IDAQ). This has been applied to a number of mind perception targets including LLMs. Jacobs et al (2023) found that higher IDAQ scores predicted a greater effect of exposure on subsequent anthropomorphism toward ChatGPT. Again, research is still in an early stage but these substantial inter-individual differences suggest a promising future research direction.

Limitations of the Human-Centered Approach and Fundamental Challenges

Many of the challenges associated with taking a Human-Centered approach to psychology research with LLMs are again shared with other approaches. For example, the specific model, and the specific tasks or contexts in which users are being situated, will influence any measures of user attitudes. However, some of the issues concerning the AI-Centered approach do not extend to the Human-Centered approach because the Human-Centered approach is principally focused on probing folk psychology and making descriptive, rather than normative claims. This focus on measuring folk psychological attitudes circumvents philosophical issues such as what it means for a computer or AI system to display intelligence or other apparent human features. In other words, the Human-Centered approach is not concerned with the validity of people's attitudes toward LLMs; rather, the approach seeks to empirically document peoples' attitudes and the contextual or individual factors that influence observed attitudes.

There are, however, certain challenges more unique to the Human-Centered approach. User attitudes and behaviour are particularly influenced by the snapshot in time and place of any study. Potential history effects may make it hard to compare between different studies and it remains to be seen how attitudes and perceptions will change over time. Despite the immense

rise in the popularity of programs like ChatGPT, it is still in its early days and the effect of long-term exposure to these programs is an outstanding issue.

A major difference between the AI- and Human-Centered approaches is also the range of tools and measures that can be employed. While the AI-Centered approach involves many different psychological tasks, they are typically constrained to inputting text into a LLM and then analyzing its textual output. The Human-Centered approach, meanwhile, has the major benefit of measuring human responses, which offers researchers the possibility of collecting a much greater range of behavioural data compared to purely textual feedback (e.g., reaction times, physiological responses, brain activations). Moreover, the most common constructs assessed in the Human-Centered approach are attitudes or perceptions, which can be assessed via self-report. Self-report has some major advantages including that a large sample of participants can be easily collected via crowdsourcing tools that are popular in the social sciences (e.g., Prolific or Mechanical Turk). While it is currently a challenge to conduct experiments remotely with people interacting with LLMs, new tools are being developed that offer a promising future direction for conducting interactive experiments with LLMs (e.g., Laban et al., 2024b) which will further diversify the types of experiments researchers can conduct in the Human-Centered approach.

The Tool-Centered Approach

Defining Tool-Centered

The Tool-Centered approach to psychology research with LLMs is defined by its focus on the use of applications of LLMs for psychology-relevant outcomes. In this sense, this approach is the widest and most interdisciplinary. From the outset, LLMs were designed to be tools for a wide variety of pursuits (many unintentional; Srivastava et al., 2022). In the following sections, we delve into several areas that are emerging areas of research focused on the efficacy of LLMs for a variety of psychological purposes. In many of these contexts, the cited research raises the question of whether or not LLMs can be an effective tool before attempting to empirically answer it. Again, the below examples are not an exhaustive list of all the psychological applications being considered, rather, they are notable or quintessential examples.

Healthcare

The application of LLMs to various healthcare pursuits is one of the most important ongoing areas of research in the Tool-Centered approach. Past research has already indicated that AI-based chatbots have the potential to revolutionize patient care with their easy-to-access, affordable, and anonymous assistance (Laranjo et al., 2018; Miner et al., 2016; Webster, 2023). Some applications in healthcare however are only tangentially related to psychology and thus will not be covered (see for example, biomedical applications concerning radiology (X. Zhang et al., 2023), genetics (Duong & Solomon, 2024), drug safety (Chen et al., 2023), or summarizing medical summaries (Tang et al., 2023; Thirunavukarasu et al., 2023)). Instead, we opt to focus on healthcare-related applications more closely related to topics usually investigated in health or clinical psychology.

One of these applications include using LLMs to generate personalized health advice as a means for augmenting or replacing professional advice. Currently, ChatGPT has restrictions in place to prevent providing professional healthcare advice but it is easy to bypass these guidelines. ChatGPT is able to answer all sorts of queries related to different medications or medical issues. It is already quite common to use the Internet for self-diagnoses with websites such as WebMD. Google searches for phrases like 'how do I know if a mole is cancerous' are common. A recent study by Shasavar and Choudhury (2023) examining ChatGPT user intentions found that nearly 80% of respondents were willing to use ChatGPT for a variety of self-diagnoses. Predictably, the degree to which users believed in the performance and the relative risk-reward of the situation influenced the likelihood of using ChatGPT. Although the study was largely exploratory, it highlights that many people will use LLMs for self-diagnoses, indicating that understanding user behaviour with LLMs and healthcare advice is an important topic for health psychology.

An interesting test case for the usability of LLMs to provide behavioural nudges toward better health decisions concerns pro-vaccination messages. A recent set of studies by Karinshak et al. (2023) found that GPT-3 generated pro-vaccination messages were perceived as more effective, stronger arguments, and evoked more positive responses than human-authored messages from the Centers for Disease Control Prevention (CDC). Interestingly, when participants were not blind to the source of the messages (i.e., human vs AI-generated), participants did express more distaste for AI-generated messages indicating that the source of the messages was still critical for nudging people toward accepting pro-vaccination attitudes. A study by Sallam et al. (2023) also sought to investigate ChatGPT's responses regarding pro-vaccination and Covid-19 conspiracy theories by administering a Vaccine Conspiracy Beliefs

scale. Although the study was largely exploratory, the results indicated that ChatGPT tended to be dismissive of conspiratorial ideas and instead provided context for vaccinations in a correct, clear, and concise manner.

It is worth noting that many applications of ChatGPT in healthcare interventions have not been as positive as the aforementioned studies. For example, when ChatGPT was asked to evaluate vignettes describing patients with varying levels of burdensomeness and thwarted belongingness, ChatGPT's assessments consistently rated the risk of suicidal attempts lower than mental health professionals despite ChatGPT's relative underestimations of mental resilience in the patients (Elyoseph & Levkovich, 2023). If mental health professionals were to solely rely on ChatGPT for evaluating suicide risk, they would be at a severe risk of underestimating actual suicide risk. These results highlight that in many applications of LLMs, they are best used as tools for augmenting current approaches rather than replacing them and they require deep scrutiny. Overall, the efficacy of using LLMs to make health-related decisions is tenuous and too inconsistent to make broad statements about whether LLMs are effective tools. It would seem to depend greatly on how LLMs are implemented and the specifics of the target to which the tool is being applied. Nonetheless, it is evident that there is a great interest in the scientific and lay community to explore the use of LLMs for numerous applications related to health psychology, and there will likely be a great number of studies investigating the ways in which LLMs can be used to replace, or augment, current approaches.

Education

The use of LLM tools in educational settings is a practical application with great interest. The dialogue nature of applications like ChatGPT fits well with the heavy focus on dialogue-based learning found in evidenced-based pedagogy (Černý, 2022; Tack & Piech, 2022).

Furthermore, LLMs provide specific opportunities for teaching such as their potential use for lesson planning, language learning, professional development (e.g., enabling teachers to stay-up-to-date on best practices), and test development (Kasneci et al., 2023). As one example, Dijkstra et al. (2022.) used GPT-3 to generate multiple-choice questions demonstrating that LLMs can help ease the substantial burden of constantly having to generate novel test questions—thus avoiding frequent reuse of past questions. In return, students can generate their own interactive study tools including practice quizzes and flashcards based on learning objectives in a course (Gabajiwala et al., 2022).

Another major benefit of LLMs for education lies in their ability to address certain inequalities. A free, personalized tutor that is easily accessible through the Internet could be beneficial for students who may otherwise struggle to afford tutoring. LLMs can also be used to benefit students with disabilities. For example, the text-to-speech and speech-to-text capabilities can be effectively used by students with visual impairments. The translation abilities of LLMs can also offer an effective means to improve language abilities in students who might be struggling with English as a second language or have foreign language learning anxiety (Bao, 2019). These examples are far from exhaustive but do illustrate some of the ways LLMs will interact with various inequalities.

The efficacy of LLM applications for education, similar to many of its other applications, largely depends on tailoring its use for certain contexts and toward different populations. The use of LLMs will differ substantially for elementary students compared with high school or university students (Kasneci et al., 2023). ChatGPT can describe ideas and theories at a variety of reading levels without complex prompting. The specific subjects LLMs are being used for will also differ substantially with some areas seeing more clear benefit than others. For example,

LLMs are highly effective at providing comprehensive annotation and descriptions for code development (MacNeil et al., 2022) and coding-related applications such as data science (Bhat et al., 2022).

In summary, there have been several promising ways to improve practices with LLM tools suggested by educators. This optimism for its use is balanced with concerns regarding its applications. These concerns principally echo ethical considerations that have yet to be fully explored, and which share many similarities with other tool-based applications and research. And, whereas education has a long history of integrating technology into practice (Firmin & Genesi, 2013; Yildirim et al., 2018), concerns are particularly important in light of the issues regarding a lack of knowledge or expertise expressed by educators and institutions when it comes to AI (Redecker, 2017).

Psychological Methods

One direction in the Tool-Centered approach to LLMs has been to examine the efficacy and usability of LLMs for aiding methodologies in psychology research and science at large. Some ideas concern particular challenges such as participant recruitment, or other more general challenges, such as extracting meaning from open-ended questions. By some accounts, LLMs are already being seen as a game changer in research practices and publishing by shortening time-to-publication, improving writing fluency, and making science more equitable (Van Dis et al., 2023).

Some of the more concrete means by which LLMs can be used as potential upside to methodological challenges in psychology were outlined in a paper by Salah et al. (2023), with a

special focus on social psychology. Among the ideas that were discussed is the ability of LLMs to rapidly and efficiently process and analyze large swathes of textual data from sources like social media, beyond what would be feasible in more manual methods. In essence, LLMs can be used to extract or uncover patterns in language use, sentiment, or emotional expression (Aydın & Karaarslan, 2022; Haluza & Jungwirth, 2023). The use of LLMs for coding behaviours (i.e., systematically recording specific actions) may soon be able to move beyond textual inputs—current GPT models are now able to provide real-time coding of video. For example, a recent demo showcased the ability of LLMs to commentate real-time sports events (Open AI, 2023).

Apart from the advantages with coding, Salah et al. (2023) highlight that LLMs have been successfully used to generate a number of online dialogues, which can be effective for simulating and modeling social interactions. This can be useful for studying a range of social influences such as examining how individuals adhere to group norms, and how minority opinion can sway majority decisions (Aydin & Karaarslan, 2022).

Another way that LLMs seem to be uprooting the workflow of psychological researchers concerns data analysis. Programs like ChatGPT are highly effective at generating code for data analysis, especially for programming languages like Python and R. LLM applications can also read files uploaded by users. It is highly effective at being able to interpret relatively messy datasets. For example, you can upload a CSV file and ask it to create a bar graph comparing the means and standard deviations of a treatment group compared to a control group. Though these tools are not perfect, it is remarkable that one can now perform many types of data analysis (including most types of inferential statistics) simply by asking an LLM to run whichever test is desired. And with a LLM's ability to contextualize, it is capable of identifying more complicated nuances like the specific test that would be best suited for the analysis, and which types of

assumptions correspond with the test in question. While purely speculative, it is easy to imagine a not-too-distant future where the vast majority of data analysis requires no coding whatsoever and can be done entirely with the use of an LLM. One considerable benefit of more automated data analysis is that it reduces the number of statistical reporting errors—a previously identified issue of concern such that almost half of all published psychology papers contain at least one *p*-value that is inconsistent with its test statistic and degrees of freedom (Nuijten et al., 2016).

Limitations of the Tool-Centered Approach and Fundamental Challenges

The range of ways LLM tools can be applied to psychological research is extremely broad, which makes sweeping claims harder to substantiate regarding limitations and fundamental challenges. The efficacy of any given tool is heavily constrained by the context around its application including human perceptions of the tool—highlighting some overlap between the Human-Centered and Tool-Centered approaches. Again, we are only in the beginning stage of LLMs, and as such it is hard to make predictions about which tools and applications will benefit psychology the most. This is not a problem particular to the Tool-Centered approach—it has been very challenging to predict how LLMs will be used by the public or in professional settings. Nonetheless, we can make inroads by highlighting some of the ethical considerations associated with the Tool-Centered approach. Ethical considerations are tied most strongly to the Tool-Centered approach as it is the area that is the most likely to incur consequential risk. In comparison, ethical considerations regarding the Human and AI-Centered approaches are typically minimal or minor. Whenever a tool is being considered for healthcare or educational purposes one ought to take more care in considering any repercussions.

There have, fortunately, been a number of studies providing insight into challenges with LLM tools, specifically as they apply to psychological research (e.g., Demszky et al. 2023; Salah et al., 2023; Van Dis et al. 2023). One of the notable ideas includes recognizing the inherent bias in training data. Most of the training data in LLMs stem from the Web, which is clearly not free from discriminatory content despite attempts at filtering it. Furthermore, most of the data is scraped from English sources and fits a typical WEIRD profile in origin (Western, Educated, Industrialized, Rich, and Democratic). Some of the more prescriptive ideas for ethical considerations regarding the use of LLMs in psychology research include mitigating the biases in training data via concerted efforts to detect bias in the data with regular monitoring and updating. Continual evaluation, and ensuring algorithmic transparency and interpretability, are essential. Handling sensitive topics with care, prioritizing informed consent and privacy, and developing rules for accountability are all broad initiatives already endeavored in psychology (e.g., Van Dis et al., 2023), and are particularly important regarding the implementation of LLM tools.

There are also ongoing debates about the ethics of accountability in the context of human-AI collaborations, such as with artistic creation and authorship (Oh et al., 2018; Wu et al., 2021). Regarding the latter, scientific journals and publishers have been quick to provide policies about the use of LLMs in manuscripts (e.g., Porsdam Mann et al., 2024), and concerns have become more pronounced as the prose output by the most advanced LLMs becomes increasingly difficult to differentiate from human-generated text. One clever idea for detecting AI-generated text is the use of *tortured phrases,* or phrases that are incorrectly used, instead of conventional phrases to describe popular ideas (Cabanac et al., 2021). For example, simple translations can lead to awry results such as replacing artificial intelligence with ‘counterfeit consciousness,’ or ‘profound neural organization’ rather than deep neural network. Whereas Cabanac et al.’s work

(2021) was released prior to the publication of more advanced LLMs such as GPT-3 or GPT-4, the method of searching for tortured phrases revealed hundreds of published papers in journals across a range of physical and social sciences suggesting that the proliferation of publicized 'junk' science can be exaggerated by AI tools. This strategy has become somewhat obsolete as recent LLMs do not output egregious tortured phrases. Currently, there is a poor ability to detect AI-generated text without generating substantial false alarms or misses in the balance between sensitivity and specificity (Elkhatat et al., 2023). There is, however, some work to suggest that inroads can be made toward better programs for detecting AI-generated text (e.g., Bozza et al., 2023; Tian, 2024.). Still, as it stands, the risks of AI-generated research papers polluting journals are a problem for the scientific community (Elali & Rachid, 2023).

In conclusion, addressing limitations and challenges in the Tool-Centered approach necessitates a focus on ethical considerations. Specifically, it requires focusing on issues regarding bias and accountability in order to best capitalize on the transformative potential of LLMs.

Summary and Conclusions

The recent developments in artificial intelligence (AI), and large language models (LLMs) more specifically, have generated a widespread recognition that these technologies have profound implications for psychology (Binz & Schultz, 2023; Demszky et al., 2023; Kosinski, 2024; Rahwan et al., 2019; Shiffrin & Mitchell, 2023; Van Dis et al., 2023). LLMs are massive statistical language models trained on billions (potentially trillions) of words with the aim of generating human-like text, contextual awareness, and reasoning abilities (Brown et al. 2020; Open AI; 2023). The sheer capabilities of the most powerful language models, like the GPT

models underlying ChatGPT or Anthropic's Claude, are astounding. For example, ChatGPT can be used to write comprehensive essays, presentations, scientific papers, as well as write computer code and run statistical analyses among many more possibilities. The societal implications of such technology require a more interdisciplinary approach with psychologists playing a special role in bringing their expertise in probing behaviour (Rahwan et al., 2019).

The purpose of the present paper is to provide an organizing framework for distilling the variety of ways in which psychologists have or ought to provide their expertise to the subject of LLMs. The benefits of such a framework are multifold. First, introducing this taxonomy provides a new layer of clarification and order. It identifies the major questions, aims, and challenges associated with each of the major approaches. It provides a bridge to past work—for example, the AI-Centered approach is most heavily influenced by the history of computing. It also allows visiting the possible theoretical implications of conceptual integration and conceptual fractionation for the emerging areas of psychological research with LLMs.

The conceptual integration value may be in the unifying of the diverse applications psychologists are applying to LLMs into the three categories discussed throughout. And therein, the recognition of similar aims, questions of importance, and ongoing challenges clustered in each category. The value of conceptual integration has widely been demonstrated in historical examples of unifying constructs and movements in science, and in psychology in particular (e.g., general intelligence or embodied cognition).

The conceptual fractionation value lies in the analysis, or the breaking apart, of topics being investigated. Since the application of psychological theory and methods to LLMs is still in its infancy, the degree of separation between the different categories explored in this review may

become clearer over time. Whereas the means by which psychologists work with LLMs may fall under a larger umbrella, this work can be delineated into distinct approaches as presented here. This may in turn follow a pattern common to many constructs studied throughout psychology wherein a seemingly unitary construct or topic of research (e.g., attention or memory) becomes broken up into distinct subtopics (e.g., covert and overt attention; episodic and semantic memory). By presenting the categories outlined here, we aim to set the stage for a clearer more nuanced approach to understanding the landscape of psychology research with LLMs.

The three approaches outlined are denoted as the AI-Centered, the Human-Centered, and the Tool-Centered approaches. Each of these approaches is principally divided by the primary question of interest and the target of measurement focus: LLM responses for AI-Centered, human users for Human-Centered, and tools or applications for the Tool-Centered approach. Notably, each of these approaches is also differentiated based on the most prototypical methods (cognitive and verbal tests, surveys, and mixed methods for each approach respectively). Likewise, these approaches can be separated by their main subdiscipline of interest in psychology (cognitive, social and personality, and mixed including health and educational) and the degree to which ethics is a crucial factor (minimal, minor, and major). Finally, each of these approaches was discussed along with quintessential subtopics amongst each of the categories: AI-Centered included intelligence and creativity, Human-Centered included anthropomorphism and trust, and Tool-Centered included explorations in healthcare and education. These introductions to the different subtopics covered by each of the approaches were not meant to be exhaustive summaries of all the work in that subtopic or approach more broadly. Rather, they were intended to illustrate the variety of different ways psychological research with LLMs is being conducted and reflect the range of possibilities for psychological researchers to garner

insights. For psychological researchers interested in beginning work related to LLMs, this paper provides a means for highlighting the range of ways expertise can be applied within this developing field of research.

Chapter 3

In Chapter 2, I laid out a framework for ongoing psychological research with LLMs. This framework makes clear that my thesis is focused on the Human-Centered approach. That is, the following empirical studies concerns people's attitudes and perceptions toward AI (or other people) in the context of social robots and LLMs. Chapter 3 contains two empirical studies (five experiments) concerning mental attribution to a wide range of robots (Study 1) and LLMs specifically (Study 2). A version of Study 1 is published in the *International Journal of Social Robotics* and a version of Study 2 is in review.

Study 1: Mind the robot! Variation in attributions of mind to a wide set of real and fictional robots

Introduction

In the 1950s, the cognitive revolution led to a convergence of psychology, neuroscience, anthropology, linguistics, and computer science in the study of human cognition and its processes (Chomsky & Schaff, 1997; Miller, 2003). Cognitive scientists of the era began to view the human mind as an information processor that takes in environmental input, processes the data, and generates an output (McCorduck, 2018). Fueled by the rapid rise of computing power by way of Moore's law, researchers have since sought to develop intelligent computing systems that function in a similar manner to human beings. Whereas most researchers targeted rational tasks such as logic and inference, a subset took an interest in developing affective computers that can recognize, process, and simulate human emotions (Picard, 2000). This quest was facilitated by the nascence of multi-signal processing of different contextual consequences that lead to learning and development (Cai, 2006).

Historically, many researchers have used the Turing test to determine whether machines have developed sufficient intelligence to be considered as having acquired thought (Saygin et al., 2000). If a human participant was unable to determine whether they were interacting with another human or a machine, then the machine was said to be intelligent (Saygin et al., 2000). However, criticisms of this approach quickly emerged after early chatbots like ELIZA (Weizenbaum, 1966) began to pass the Turing test by tricking humans. Essentially, these programs were capable of manipulating symbols to appear as though they have understanding—without actually having any underlying understanding. Highlighting this point, John Searle argued that a computing system could not be shown to have a mind even if it demonstrates seemingly intelligent or human-like behavior (Searle, 1980).

While philosophical debates relating to AI continues, and criticisms of John Searle's idea have emerged (Chalmers, 1992; Dennett, 1991), researchers have since agreed that people often perceive intelligent or human-like qualities in AI regardless of whether that computing system has any real understanding of its behavior (Epley et al., 2007). Despite individual differences like personality (Tharp et al., 2017) and psychopathology (Gray et al., 2011) playing a role, anthropomorphism of nonhumans, including machines, is relatively common (Waytz et al., 2010a). Indeed, one study examining the activation of the mirror neuron system found that the same neural circuitry underlies the perception of human-like behavior in an anthropomorphized robot as the perception of an actual human's behavior (Gazzola et al., 2007).

One major advantage of studying these perceptions of human-like behavior in computing systems is that rather than trying to quantify a computer's understanding of its behavior, one can assess people's perceptions of how a computer might imitate intelligent or human-like behavior. Additionally, researchers have been empirically investigating anthropomorphic attributions for

almost 80 years (for example, see Heider & Simmel (1944)), leaving a broad and solid literature upon which to ground theory. As a result, several empirical paradigms have emerged dedicated to 'seeing human' in nonhumans, such as what physical features (e.g., hands, eyes) result in robots being seen as more or less humanlike (Phillips et al., 2018), some, with widespread support (Epley et al., 2007; Gray et al., 2007; Malle, 2019).

One of the methods that has become popular for seeing human in nonhumans is called the mind perception approach (or framework). The mind perception approach allows researchers to directly probe people's self-reported perceptions. A major advantage of this approach over other paradigms is its focus on the hallmark of humanity—the mind (Gray & Wegner, 2012). This provides a promising means of investigating human qualities like emotion and empathy in digital agents because those traits are considered an essential part of the human mind. The mind perception framework was introduced in the seminal work of Gray et al. (2007). They used pairwise comparisons and factor analysis to examine the critical components of mind perception in humans and nonhumans. They found that there were two principal components: agency and experience. Agency refers to a collection of different capacities including self-control, morality, memory, emotion recognition, planning, communication, and thought. Experience refers to a collection of other capacities including hunger, fear, pain, pleasure, rage, desire, personality, consciousness, pride, embarrassment, and joy. These two separable components of agency and experience map respectively onto Aristotle's distinction between moral agents (agents that morally can do right or wrong) and moral patients (patients that can have moral right or wrong done to them). In the initial Gray et al. (2007) study, they included a social humanoid robot and found that while it scored in the midrange for agency, people assigned it very little experience. While the social humanoid scored higher on agency than a dog, baby, and a dead person, the

dead person and every other character other than God scored higher on ratings of experience. However, Gray et al. only included one type of social humanoid robot (and a rather crude one at that), limiting their ability to observe if the mind perception framework could be used to compare different types of computing systems or machines in terms of their abilities to display evidence of agency and experience.

Subsequent work has demonstrated that the mind perception framework can add to the growing body of research that suggests anthropomorphism can facilitate the development of human–computer interactions, including affective computing systems (Eyssel et al., 2011; K. Gray & Wegner, 2012; Moore, 2016). Indeed, the mind perception framework has been validated for use with social robots by demonstrating that a robot can elicit both agency and experience attributions in a variety of ways. For example, Gray and Wegner (2012) found that manipulating a visual perspective of a robot can produce the same result as manipulating written descriptions associated with a robot which emphasize its capabilities. Essentially, if experimenters emphasize that a robot is more humanlike, either visually or verbally, participants will tend to increase the amount of agency and experience they perceive in it. In a comparable vein, de Graaf and Malle (de Graaf & Malle, 2018) found that written descriptions of basic properties of behaviour can be matched to elicit similar attributions of mind, regardless of whether the agent performing the behaviour is a human or computer. Using these, they were able to develop a pool of robust stimulus behaviours whose properties are matched between a human and robot on ratings of intentionality, surprisingness, and desirability (a three-factor model of mind perception).

These previous applications of the mind perception framework to human-computer interaction have also demonstrated that the framework can predict real-world behavioural differences. For instance, Stafford et al. (Stafford et al., 2014) found positive attitudes towards

robots and perceptions of agency predicted social humanoid robot use and acceptance in residents of a retirement home.

However, a major caveat in previous validations of the mind perception framework is that it is unclear if the framework can be used to compare between a range of different robots. Previous work has only compared a range of behaviours (de Graaf & Malle, 2018), human features (Kamide et al., 2013), or a range of manipulations to the same robot or its description (Gray & Wegner, 2012; Thellman et al., 2017). In particular, it is unclear if social robots will still elicit attributions of mind when a wide range of real-world robots are compared. After all, hypothetical future conscious robots should, by definition, elicit attributions of mind unlike real-world robots, given that any written description of a future conscious robot strongly emphasizes its humanness by describing it as conscious. As such, many questions remain. Can robots, ever, let alone today, compare with living beings on both agency, and more doubtfully, experience? Is there enough perceived variation between robots in the mind perception framework to suggest that people may perceive a mind in a robot, both in terms of a robot's agency and, crucially, its experience?

Present Study

To address this issue, the present study applies the mind perception framework to compare a large group of affective computing systems ranging from fictional characters like Wall-E and R2D2, to a real-world social humanoid robot (Sophia), to other more mundane non-social robots or agents like Siri, Sphero, and a robotic vacuum (see supplementary materials for descriptions of each character https://osf.io/n7xf4/). If we observe significant variation amongst the real and fictional robots on agency, and most critically, experience, then it would suggest that

the mind perception framework can be used as a means of assessing and comparing different affective systems.

Note, that while our primary focus is on the perception of real robots in terms of experience and agency, the inclusion of fictional robots allows us to explore the boundaries on people's willingness to perceive robots on these two dimensions (i.e., we expect the mind perception of fictional robots to exceed that of real robots). However, if there is little variation in the ratings of the robots in terms of agency and experience, especially so for the real robots, then it would suggest that developing real computing systems that will be perceived as being able to do and feel will be a most challenging (and dubious) exercise both presently and in the foreseeable future. As the key measurement is how others perceive robots that are drawn from the past, present, and future; and past work indicates considerable individual differences in anthropomorphism, we thought it prudent to examine if common moderators (age and sex) affect attributions of agency and experience towards robots.

Methods

Participants

There were 123 participants that took part in the study. Participants were recruited via Amazon Mechanical Turk (MTurk). MTurk is an Amazon hosted network for contracting short term work for tasks like surveys. MTurk's robustness has been extensively researched and results support that MTurk is a valid means of collecting online data for the behavioural sciences and is sometimes even superior to in-person collection (Casler et al., 2013; Hauser & Schwarz, 2016). Still, as a precaution, only MTurk experts were recruited, those who have successfully completed 1000 other tasks with a success rate over 99%. Participants' mean age was 41.47 (SD = 11.29). There were 55 females and 68 males. All participants completed the study in the United States.

Material

The study was formatted as a Qualtrics survey. Each survey question was comprised of an image, a one-line description of the image, and two rating scales (see online resources). In total, there were 24 images of human and nonhuman characters, each paired with two 0-100 rating scales for agency and experience. The characters were selected from four different categories similar to the Gray et al. (2007), in order to situate people's ratings and provide a manipulation check. The four categories were humans, animals, inanimate objects, and computational beings. The human category included a woman, a dead person, a baby, a fetus, a child, and a person in persistent vegetative state (PVS). The animal category included a dog, a frog, a chimpanzee, and bacteria. The inanimate category included a rock, a teddy bear, and a shovel. The machine category included Sophia, Sphero, R2-D2, Siri, Roomba, Maslo, Atlas, Beam, future conscious robot (FCR), Alexa, and Wall-E.

Procedure

The survey was distributed through MTurk and Qualtrics. First, participants signed a consent form and filled in a questionnaire. It was a standard demographic questionnaire querying the participant's age and sex. Next, the participants were presented with the image of an adult woman along with a short description. The participants rated the woman on agency and experience on a 0–100 scale. The participants then repeated the process with an image of a rock in order to familiarize them with the range of characters to be presented. After these first two characters (the adult woman and the rock), the participants rated the rest of the characters, which were presented in a randomized order, on agency and experience in the same manner. The definitions of agency and experience and their component capacities were included alongside the rating scales during each character presentation should a participant need to remind themselves of one or both definitions.

Results

The analyses are split into 2 components with each comparing how agency and experience attributions vary across characters. The first section considers all the characters and pertains more to manipulation checks. The second section examines only robots and specifically investigates if there is significant variation in the attributions made to all the robots (real and fictitious), and to the real robots alone.

All Characters

The purpose of this initial set of analyses is to ground our results in the broader literature and provide manipulation checks. Two one-way Greenhouse Geisser corrected within-subjects ANOVAs revealed significant differences amongst the 24 characters on agency, $F(7.38, 900.96) = 163.22$, MSE = 1624.27, $p < .001$, and experience, $F(6.92, 844.69) = 365.88$, MSE = 1317.36, $p < .001$. We did not include age or sex in these analyses given that it would not add to our manipulation checks. In terms of these checks, it appeared that people accurately understood the purpose of the task given the consistently low scores of agency and experience for inanimate characters like the rock, shovel, and dead person. Furthermore, characters that would be expected to score high on agency and experience (e.g., an adult woman) did do so. See Fig. 1.1 below for a visualization of mean agency and experience attributions by character.

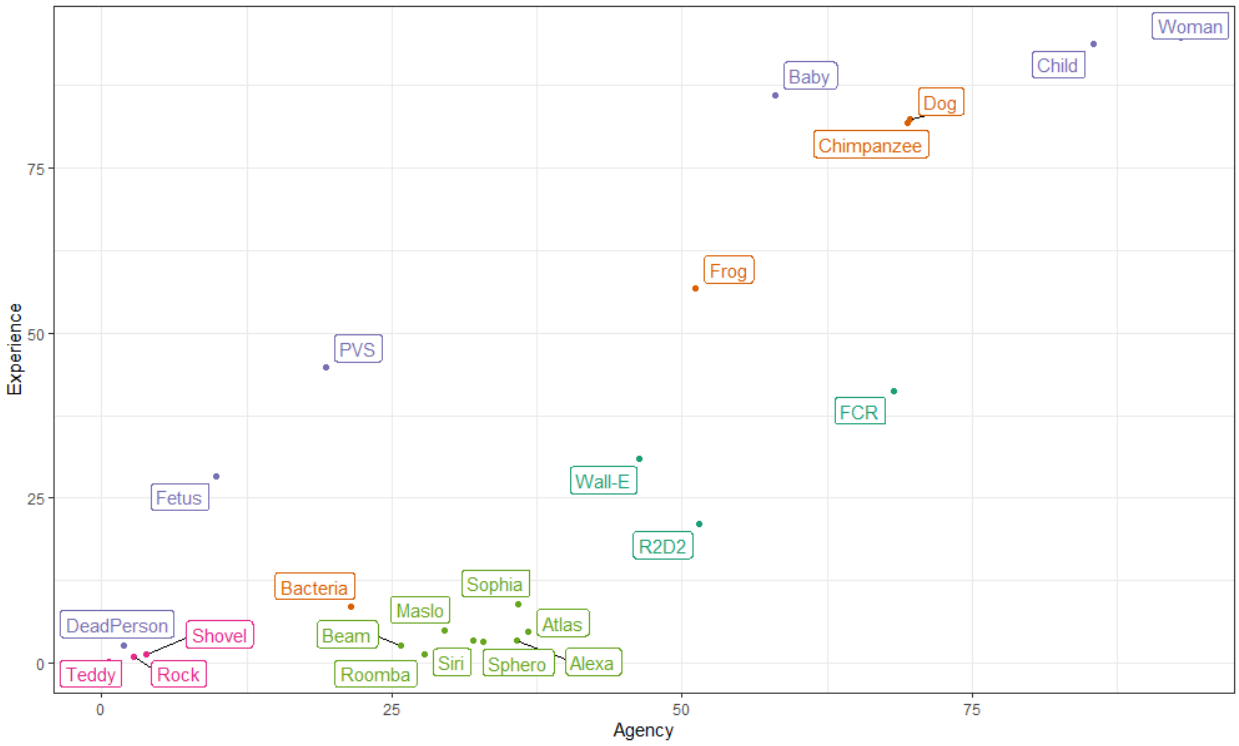

Figure 1.1: Mean attributions of agency and experience by character. Colour indicates character category (blue = humans; orange = animals; pink = inanimate things; light green = robots; dark green = fictional robots).

Robots

The purpose of the second set of analyses is to examine the agency and experience attributions that were made for the robots, with and without the inclusion of fictional robots. In order to more precisely measure the difference between robots, we include sex (self-reported) and age. For age, we perform a median split of the data with those 38 and under in one group and those over 39 in the other. First, an 11 (all robots) × 2 (participant sex) × 2 (participant age split) Greenhouse Geisser corrected mixed model ANOVA was conducted for agency and experience attributions. For the main effect of robot, there were significant differences for attributions of agency, $F(4.64, 552.29) = 45.34$, MSE = 818.79, $p < .001$, and experience, $F(2.81, 334.70) =$

63.86, MSE = 1151.94, $p < .001$. As shown in Table 1.1, this reflects the fact that fictional robots tended to differ from the other robots and from one another. There was no main effect of sex for either agency or experience (p's > 0.18) but there was a main effect of age for both agency, $F(1119) = 7.52$, MSE = 6157.81, p = .007, and experience, $F(1119) = 4.75$, MSE = 1767.14, $p = .031$. There were also significant interactions between robot and sex for agency, F(4.64, 552.29) = 4.10, MSE = 818.79, $p = .002$, and experience, $F(2.81, 334.70) = 2.75$, MSE = 1151.94, $p = .046$. No other significant interactions between any factors or the combination of the 3 factors were observed (p's > 0.05). In sum, there was significant variation between robots both for agency and experience. Sex was not a significant factor by itself, though there were sex differences amongst attributions towards some robots. Age was a significant factor with the younger participants giving higher agency and experience scores for robots

Next an 8 (real robots) x 2 (participant sex) x 2 (participant age split) Greenhouse Geisser adjusted mixed model ANOVA was conducted having omitted fictional robots (FCR, R2D2, Wall-E). For the main effect of robot, there were significant differences for attributions of agency, $F(5.50, 654.47) = 8.95$, MSE = 255.32, $p < .001$, and experience, $F(2.99, 356.12) = 12.41$, MSE = 119.94, $p < .001$. Post hoc pairwise Scheffé tests were conducted for agency and experience scores and are presented in Tables 1.1 and 1.2 below. Note that the nature of these differences varies as a function of experience and agency. Maslo and Roomba differed in terms of experience, and all robots differed from Sophia. Whereas for agency, the difference between the robots is more distributed and idiosyncratic: Sophia only differs from Beam and Roomba; Roomba differs from Alexa and Atlas; and Atlas differs from Beam and Maslo. There was no main effect of sex for either agency or experience (p's > 0.18) but there was a main effect of age for both agency, $F(1,119) = 7.66$, MSE = 5,152.79, $p = .007$, and experience, $F(1,119) = 4.69$,

MSE = 936.25, p = .032. There were no significant interactions between factors or the combination of the 3 factors (p's > 0.07).

In sum, there was still significant variance amongst the real robots in terms of their ratings for agency and experience, and this variation itself changed as a function of the dimension of mind consistent with Gray et al.'s (2007) finding that these are qualitatively different representations of mind. Sex was not a significant factor, but age was, with the younger participants again giving higher scores for robots. The results of these ANOVAs are presented collapsing across sex in Figures 1.3 and 1.4 below. See Fig. 1.2 for a general visualization of mind perception ratings amongst real robots.

Table 1.1: Scheffé pairwise comparisons amongst all robots. Experience presented in red (top) and agency in black (bottom). Significance is indicated by asterisks (***: p < .001, **: p < .01, *: p < .05).

	Maslo	Sophia	Siri	Roomba	Sphero	Beam	Atlas	Alexa	FCR	Wall-E	R2D2
Maslo	•	-1.76	0.72	1.55	0.68	1.00	0.09	0.67	**-15.17*** **	**-10.90*** **	**-6.50*** **
Sophia	2.55	•	2.49	3.31	2.44	2.77	1.85	2.43	**-13.40*** **	**-9.14*** **	**-4.73** **
Siri	1.01	-1.54	•	0.83	-0.05	0.28	-0.64	-0.05	**-15.89*** **	**-11.63*** **	**-7.22*** **
Roomba	-0.67	-3.22	-1.68	•	-0.87	-0.55	-1.47	-0.88	**-16.72*** **	**-12.45*** **	**-8.05*** **
Sphero	1.48	-1.07	0.46	2.15	•	0.33	-0.59	0.00	**-15.84*** **	**-11.58*** **	**-7.17*** **
Beam	-1.19	-3.73	-2.20	-0.52	-2.66	•	-0.92	-0.33	**-16.17*** **	**-11.90*** **	**-7.50*** **
Atlas	2.91	0.36	1.90	3.58	1.43	4.10	•	0.59	**-15.25*** **	**-10.99*** **	**-6.58*** **
Alexa	2.31	-0.23	1.30	2.98	0.84	3.50	-0.60	•	**-15.84*** **	**11.57*** **	**-7.17*** **
FCR	**15.00*** **	**12.46*** **	**14.00*** **	**15.68*** **	**13.53*** **	**16.20*** **	**12.10*** **	**12.70*** **	•	4.27	**8.67*** **
Wall-E	**6.31*** **	3.77	**5.31** **	**6.99*** **	**4.84*** **	**7.50*** **	3.41	**4.00**	**-8.70*** **	•	**4.40***
R2D2	**8.16*** **	**5.62*** **	**7.16*** **	**8.84*** **	**6.69*** **	**9.35*** **	**5.26** **	**5.85*** **	**-6.85*** **	1.85	•

Table 1.2: Scheffé pairwise comparisons amongst real robots. Experience presented in red (top) and agency in black (bottom). Significance is indicated by asterisks (***: p < .001, **: p < .01, *: p < .05).

	Maslo	Sophia	Siri	Roomba	Sphero	Beam	Atlas	Alexa
Maslo	•	**-4.43****	1.82	**3.90***	1.70	2.52	0.22	1.69
Sophia	3.50	•	**6.25*****	**8.33*****	**6.13*****	**6.95*****	**4.64****	**6.12*****
Siri	1.39	-2.12	•	2.08	-0.12	0.70	-1.61	-0.13
Roomba	-0.92	**-4.43****	-2.31	•	-2.20	-1.38	-3.68	-2.21
Sphero	2.03	-1.47	0.64	2.95	•	0.82	-1.49	-0.01
Beam	-1.64	**-5.14*****	-3.02	-0.71	-3.67	•	-2.31	-0.83
Atlas	**4.00***	0.50	2.61	**4.93****	1.97	**5.63*****	•	1.48
Alexa	3.18	0.32	1.79	**4.11***	1.15	**4.82****	-0.82	•

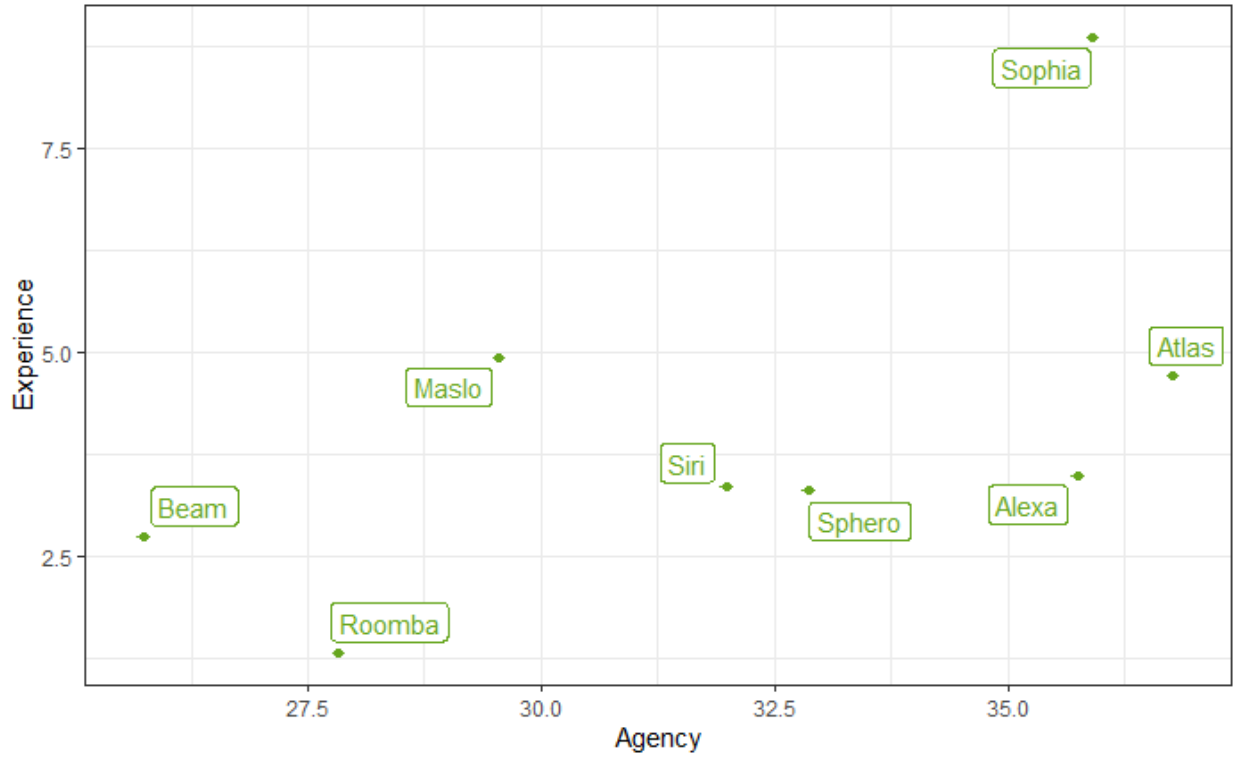

Figure 1.2: Attributions of mind perception for real robots.

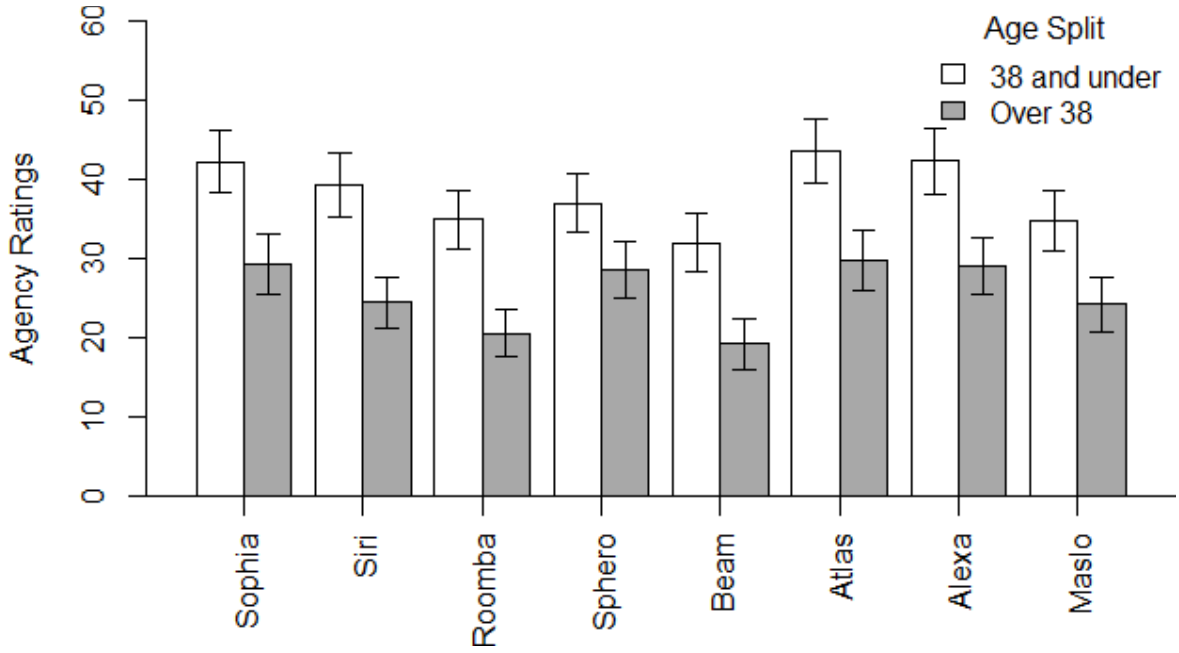

Figure 1.3: Mean agency ratings for real robots split by age. Error bars represent standard error.

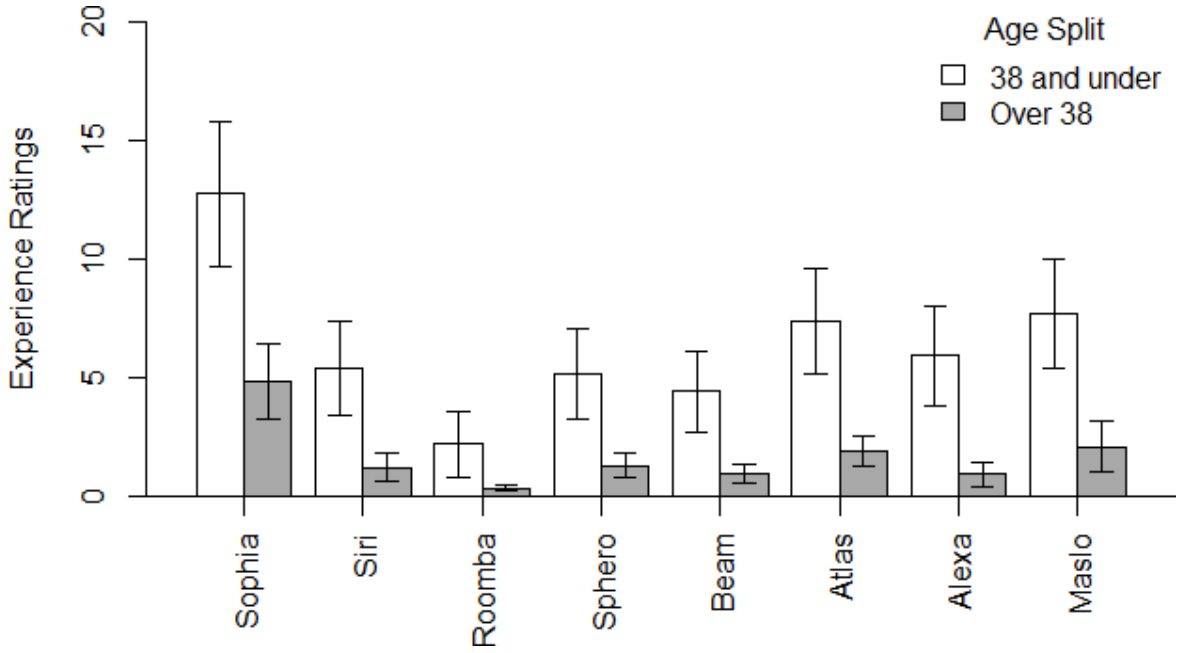

Figure 1.3: Mean experience ratings for real robots split by age. Error bars represent standard error.

Discussion

People have long been interested in developing machines that are capable of emulating key aspects of the human mind, including traits such as empathy and morality. A major technical challenge, however, has been quantitatively assessing different machines in their ability to imitate complex human phenomena. However, recent advances have introduced powerful new research tools for investigating how people view machines in terms of their ability to think, do, and feel, thus opening to the door to answering outstanding questions. Previous work using these new frameworks have established that a particular robot can elicit substantial attributions of mind by manipulating descriptions, perspective taking (Gray & Wegner, 2012), human features

(Kamide et al., 2013), or behaviours associated with the agent (de Graaf & Malle, 2018; Thellman et al., 2017). However, an outstanding question remains as to whether or not a wide range of robots will elicit different attributions both on agency, and critically, experience? And crucially, when only real robots are compared amongst themselves, is there still significant variation in people's attributions? If differences only emerge for fictional robots, it would represent a potential boundary condition for the actual development of real-world robots that are perceived as possessing minds. Nevertheless, it would be far less remarkable to find that fictional robots can be perceived as having minds because, by design, those robots are free to be endowed with human qualities (including consciousness) as they are not constrained by the same limitations as real robots. Thus, we were especially interested in any variation in the perceptions of mind for real robots, particularly in terms of experience (the ability to feel).

The mind perception framework was used to address this issue by comparing a large group of affective computing systems ranging from fictional characters like Wall-E and R2D2, to a real-world social humanoid robot, to other more mundane digital agents like Siri, Sphero, and a robotic vacuum. In general, the pattern of results was clear. The results replicated the Gray et al. (2007) key findings, with a rock and human representing the two extremes of mind, with a relative decline emerging as one moved from an adult to a child to a fetus to a dead person.

Against this robust backdrop our results revealed that robots, both real and fictional, were perceived as having significantly weaker qualities of mind in comparison to humans, including a child or baby. As a group and individually (including fictional robots), robots scored consistently lower on agency, and particularly, on experience (see Figure 1.1).

Nevertheless, our results demonstrated that people perceive significant variation of mind in robots both in terms of agency and experience. In terms of agency, we found differences

between digital agents, especially between more limited systems like Siri and Alexa compared to more complex robots like Sophia or Atlas. This is perhaps reasonable given that the latter two can move around rooms while at best, in terms of embodiment, Siri and Alexa are confined to a speaker. Intuitively, Sophia and Atlas can do much more than Siri and Alexa. In terms of experience, however, it is intriguing that a more complex robot like Sophia differed so profoundly in how much experience people attributed to it compared to more limited systems. After all, on first blush it seems less intuitive to think that Sophia or Atlas can feel much more than Siri or Alexa can. We can observe what Siri or Sophia can do and note differences, but how do we assess what they can feel? Given that a hallmark of the mind, and what separates humans from nonhumans is related more to experiential capabilities than agency (Gray et al., 2007), it is particularly noteworthy to find differences on experiential capabilities because this would suggest that robots are becoming significantly more human-like in the most important way possible.

In sum, the present study indicates that people are open to perceiving qualities of mind in robots—in both agency and experience—and that the mind perception framework can be used as a means of assessing and comparing different robots on both their agency and affective capabilities. Furthermore, these results persist when fictional robots are excluded, indicating that people perceive these capabilities in robots that can be found today. This suggests that a landscape of robots differing in their agency and affective capabilities is not part of a distant future; rather it exists today. Indeed, if all the robots had scored low on experience, or if robots only differed in agency, it would have suggested that building robots capable of emulating key aspects of the human mind would have been a most challenging exercise. Instead, the present

results support the notion that building robots to emulate key hallmarks of the human mind, like empathy and morality, while still a difficult if not daunting task, is a fruitful avenue to pursue.

Our study also revealed that age may be a critical factor to mind perception in robots. Older adults (> 38 years) gave lower scores for real robots on both agency and experience compared to their younger (< 39 years) counterparts. These data suggest that there are generational differences in how open people are to perceiving both agency and experience in robots. It is our speculation that these data may be emblematic of a gradual societal change in how people perceive robots as these agents become increasingly more sophisticated and intertwined with everyday living. This also supports the idea that people are becoming more willing to perceive a range of capacities of agency and experience in robots.

In conclusion, the present investigation discovered that there is significant variation in the mind perception of a range of robots, not only in terms of what they can do (agency), but crucially, with regard to what they can feel (experience). This challenges the idea that developing computing systems that will be perceived as being able to do and feel is a distant, possibly unattainable future. Moreover, the mind perception framework can continue to facilitate our understanding of human–computer interaction, even when investigating a range of real-world robots because people do perceive them as being capable of both doing and feeling as shown here. The generational difference we found in how people attribute agency and experience to robots also suggests that people can change how they perceive minds in robots, highlighting the potential influence of exposure and education. Collectively these data provide substantial support for the notion that future robots will be perceived as possessing closer approximations of the human mind—even on the most unique parts of what makes us human.

Study 2: Attributing mind to Large Language Models: The effect of exposure and individual differences

Following the discovery of widespread attributions of agency and experience to a range of real and fictional robots, I turn toward examining mind perception specifically with LLMs. Study 2 contains 4 experiments and is currently under review. The 4 experiments differ in the type of exposure (vignettes or real interaction) and the amount of interaction (3 prompts or an extended open-ended interaction). Experiment 3 in Study 2 also manipulates the origin of the LLM response (ChatGPT, Claude, or LLaMA).

Introduction

The rapid improvements in artificial intelligence (AI) in recent years have led to increasingly sophisticated chatbots and large language models (LLMs) such as ChatGPT (OpenAI, 2022) or Claude (Anthropic, 2024). These chatbots with impressive language understanding capabilities have renewed questions about how people anthropomorphize AI systems as tens of millions of users have begun to use them (Hu, 2023). The act of anthropomorphism, which can range from the tendency to see human-like shapes in the environment (Złotowski et al., 2014), to a broader definition entailing the attribution of human-like qualities to nonhumans, has been studied extensively, providing a rich corpus of research to draw upon in order to better understand people's perceptions of AI and LLMs specifically (Epley et al., 2007; Waytz et al., 2010a).

One method of empirically investigating anthropomorphism is to use the mind perception framework. Mind perception, in general, refers to the degree to which individuals attribute mental capabilities, such as thoughts, feelings, intentions, and drives to others including other people, animals, robots, and crucially, AI systems (Gray et al., 2007; Waytz et al., 2010b). One

popular model of mind perception is the 2-factor model of mind perception developed by Gray, Gray, and Wegner (2007). In their seminal work using factor analysis, Gray et al. (2007) found that mind perception can be separated into two principal factors: agency, which represents the ability to do, think, and act morally; and experience, which represents the capacity to feel emotions, sensations, and drives. This model of mind perception has seen widespread use in psychology, cognitive science, computer science, and other related fields (e.g., Jacobs et al., 2022; Stafford et al., 2014; Wiese et al., 2017).

Previous work investigating anthropomorphism has found that a number of contextual factors influence anthropomorphism, such as a robot's appearance (Müller et al., 2021) or behaviour (Waytz et al., 2010a). AI systems that are designed to look or behave in a human-like manner tend to elicit higher levels of anthropomorphism in different contexts (DiSalvo et al., 2002). Individual differences have also been shown to play a significant role in shaping perceptions of AI (Waytz et al., 2010a). People with a higher propensity to anthropomorphize tend to attribute more mental states to AI and robots (Epley et al., 2007). Moreover, demographic factors including age, gender, and cultural differences have been reported to influence attitudes toward AI and robots (Uysal et al., 2023). For instance, Jacobs et al. (2022) found that younger individuals tend to attribute greater experiential capabilities to AI or social robots. Eyssel and colleagues (2012) discovered that when robots and participants shared the same gender they were more likely to feel psychological closeness (Eyssel et al., 2012) and Syrdal et al. (2020) have noted that some psychometrically assessed attitudes toward robots are culturally specific.

Personality traits have also been found to be associated with different attitudes toward AI and robots (Rossi et al., 2020). Individuals scoring higher on openness to experience, as well as other factors such as trust, have been positively related to the acceptance of social robots (Rossi

et al., 2020). Big 5 personality traits have also been closely linked to mind perception in humans (Tharp et al., 2021). However, the extent to which these personality traits are related to mind perception in the context of advanced AI chatbots like ChatGPT remains unclear due to their nascent nature. Furthermore, despite the existing literature on mind perception of AI, little is known about the impact of advanced AI chatbots, such as ChatGPT, on mind perception, in addition to the potential influence of individual differences in this process.

This absence of knowledge regarding mind perception toward LLMs is a particularly compelling area to study given the deep interconnectedness of mind perception and various physical and mental health conditions such as Autism Spectrum Disorder (ASD) (Gray et al., 2011). As LLMs become more integrated with social robots (e.g., Addlesee et al., 2024; Esteban-Lozano et al., 2024), a greater understanding of how people make mental attributions to LLMs will provide insight into how they can be used for a variety of interventions such as supporting people's emotional health (Laban et al., 2024a) or helping individuals with ASD (Mishra et al., 2024). For example, if people are open to ascribing experiential features (the more socially relevant qualities of mind) to LLMs, it would suggest that LLM integration with social robots may be a more promising avenue for social interventions than previously believed. This integration with social robots may bring the benefit of greater conversational ability compared to traditional social robots without the issue of disembodiment (i.e., voice only) that has been shown to limit healthy behaviours such as self-disclosure in health interventions (Laban et al., 2020).

The present study investigates the impact of exposure to LLMs on mind perception across four experiments differing in the type of exposure (vignettes or real interaction) and the model participants are exposed to (ChatGPT, LLaMA, Claude). Based on previous work

extending individual differences to mind perception and AI, we employ measures of personality (Big 5: McCrae & Costa, 1996), anthropomorphism (IDAQ: Waytz et al., 2010a), autistic traits (AQ-10: Baron-Cohen et al., 2001), and other common moderators including gender, age, and education in the two online experiments with larger samples (Experiments 1 and 3).

Experiment 1

In this first experiment, participants are tasked with judging the perceived mind of ChatGPT before and after being presented with three examples of real responses. We expect that ratings of mind would increase after being exposed to ChatGPT (H1). Additionally, we expect that there will be a split between facets of mind perception such that we hypothesize exposure will influence experience attributions to a lesser extent than agency attributions (H2). We formulate this latter hypothesis (H2) given that people in general are reluctant to ascribe experiential qualities to robots and machines in contrast to agency (e.g., Gray et al., 2007; Thellman et al., 2022). However, some research has provided examples of people differentiating between robots or machines in their levels of ascribed experience (Jacobs et al., 2022; Müller et al., 2021) or relatedly, emotion (Lakatos et al., 2014; Spatola & Wudarczyk, 2021). We also measure a range of individual differences that have been known to influence mind perception (i.e., age, gender, AQ, Big-5, and IDAQ) to assess how these may impact a change in mind perception.

Methods

Participants

There were 148 participants that took part in the study following a G*Power analysis suggesting a sample size of 147 for detecting a small effect (d=.3) between two independent means of matched pairs with 95% power. No data were excluded. There were 92 males and 56 females. The mean age was 39.9 years of age (SD = 11.5). In terms of education, 1 individual

stated ‘Less than high school degree’, 21 stated ‘High school graduate (high school diploma or equivalent including GED)’, 26 stated ‘Some college but no degree’, 16 stated ‘Associate college degree (2-year)’, 63 stated ‘Bachelor’s degree in college (4-year)’, 17 stated ‘Master’s degree’, 2 stated ‘Doctoral degree’, and 2 stated ‘Professional degree (JD, MD)’. This experiment, and all subsequent experiments, was approved by the Behavioural Research Ethics Board of the University of British Columbia (H22-00572).

Procedure

Participants were recruited through Amazon’s Mechanical Turk (MTurk) and took part from IP addresses in the United States. After obtaining written consent to take part in the study, participants were asked to fill out several demographic questions regarding their age, gender, and education. Participants were then asked about their previous exposure to AI chatbots, and ChatGPT specifically, using a 5-point Likert scale ranging from ‘None at all’ to ‘A great deal’. Next participants were asked to rate AI chatbots on their agency and experience capabilities (see below for more details; Gray et al., 2007). Subsequently, participants were shown three real prompts provided to ChatGPT (GPT-4; April 2023) and its responses (see https://osf.io/rxpvm/). The first was answering a question about the nature-nurture debate, the second was a trick question about how many feet fit in a shoe, and the final question asked ChatGPT to create a cover letter for a janitorial position on the moon. These prompts were chosen because they reflect the diversity of responses ChatGPT is capable of providing, from being more educational and informative, to being more creative. After these prompts, participants were asked again to rate their mind perception of ChatGPT. Finally, participants were asked to fill out the AQ-10, TIPI (short form Big-5), and IDAQ (anthropomorphism) scales (see details below).

Materials and Measures

The following scales were administered through Qualtrics after dissemination through MTurk.

Mind Perception. Participants were asked to rate the agency and experience of ChatGPT. Specifically, participants were told “Using your intuition, please rate what you think of AI chatbots on the following scales.”, followed by “To what degree do you think AI chatbots (such as Chat-GPT) exhibit these abilities?”. These instructions were paired with 2 sliding scales for agency and experience with additional definitions placed in parentheses on each sliding scale: “Agency (the ability to do)” and "Experience (the ability to feel)”. These scales ranged from 1-100 and the sliding scale indicator had a start point set in the middle that participants adjusted in order to proceed to the next page. It is worth noting that various mind perception measures exist from different dimensional models to different modifications to the number of items. Single-item sliding scale measures and brief definitions (i.e., defining agency as the ability to do and experience as the ability to feel) have been used in a variety of past studies (Edwards et al., 2024; Will et al., 2021). They have also been shown to be effective at capturing variations between different targets of mind perception including a wide range of robots and AI systems (Jacobs et al., 2022). Additionally, research conducted using both single-item and multi-item scales has produced similar patterns of participant responses in the context of AI (Jacobs et al., 2024).

AQ-10. The Autism Spectrum Quotient (AQ-10) is a brief 10-item self-report questionnaire used to assess autism spectrum traits in adults (Allison et al., 2012). The AQ-10 provides an efficient means of measuring autistic traits, allowing researchers and clinicians to identify individuals who may require further evaluation for autism spectrum disorder (ASD).

The AQ-10 has been validated for clinical and research purposes (Allison et al., 2012; Baron-Cohen et al., 2001; Booth et al., 2013; Forby et al., 2023; Ruzich et al., 2015).

TIPI. The Ten-Item Personality Inventory (TIPI) is a short self-report questionnaire designed to measure the Big Five personality traits: extraversion, agreeableness, conscientiousness, emotional stability, and openness to experience (Gosling et al., 2003). As a brief alternative to longer personality measures, the TIPI offers a time-efficient and user-friendly means of assessing the Big Five traits in various research settings.

IDAQ. The Individual Differences in Anthropomorphism Questionnaire (IDAQ) is a self-report questionnaire designed to assess the degree to which individuals anthropomorphize or attribute human-like characteristics, such as intentions, emotions, and cognition, to nonhuman entities (Waytz et al., 2010a). The IDAQ scale has been shown to have strong psychometric properties in terms of validity and reliability (Waytz et al., 2010a; Epley et al., 2007).

Data Analysis

All data analyses were conducted in R (v4.0.5; R Core Team 2021) using the R packages tidyverse (Wickham et al., 2019), ggplot2 (Wickham, 2016), ez (Lawrence, 2009), and effsize (Torchiano, 2013). Data for this experiment, and all subsequent experiments, are publicly available on OSF at https://osf.io/rxpvm/ along with the vignettes shown to participants.

Results

Influence of Exposure

The influence of exposure to prompts was measured through mind perception questions administered before and after exposing participants to the real-AI-generated responses using GPT-4 (see Figure 2.1). A paired samples t-test was conducted comparing pre- and post-agency ratings and another paired samples t-test was used to compare experience ratings (pre and post).

There was a significant increase in agency scores (the ability to do) after the vignettes (*M* = 75.88, *SD* = 26.31), *t*(147) = 6.95, *p* < .001, d = .29, 95% CI [.21, .38], compared to before the vignettes (*M* = 68.09, *SD* = 27.07). Similarly, there was a significant increase in experience ratings, *t*(147) = -3.57, *p* < .001, *d* = .14, 95% CI [.06, .22], with participants rating ChatGPT's experience (the ability to feel) higher after exposure to vignettes (*M* = 30.13, *SD* = 34.45) compared to before the vignettes (*M* = 25.33, *SD* = 31.08). These results indicate that people increase their mind perception of ChatGPT after being given examples of its capabilities.

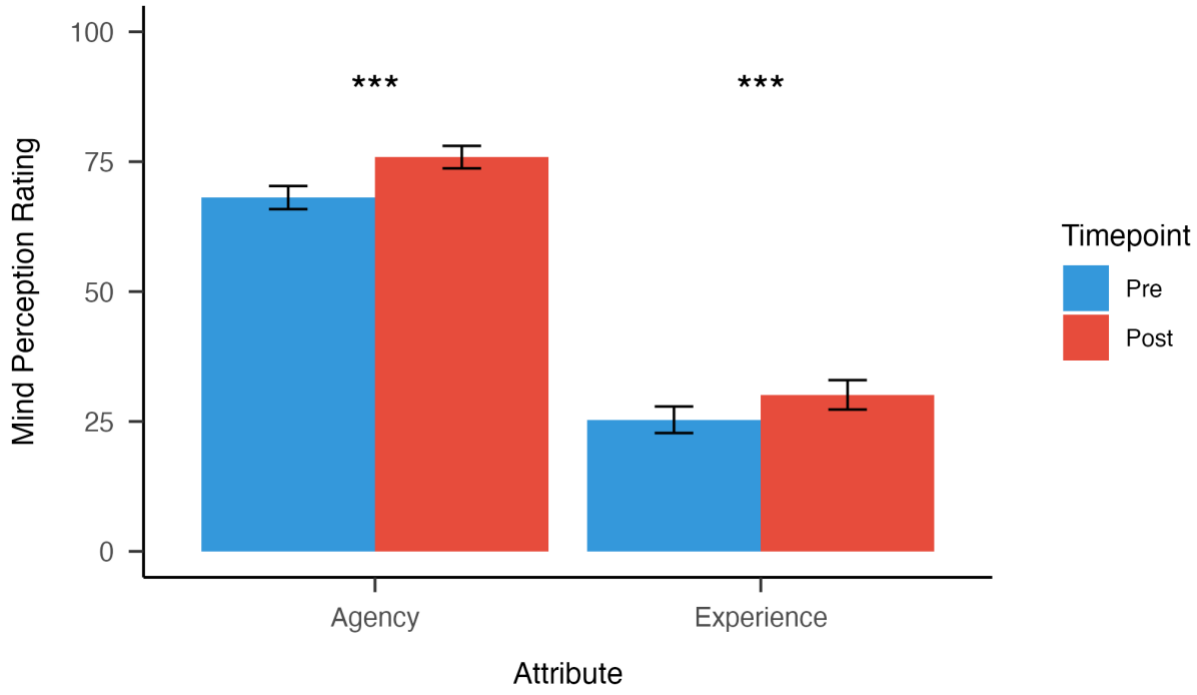

Figure 2.1: Mean agency and experience ratings pre- and post-exposure. Error bars represent standard error and *** indicates significance at p <.001.

Prior Exposure

Prior exposure to AI chatbots, including ChatGPT, was assessed using a 5-point Likert scale item ranging from ‘None at all’ to ‘A great deal’. These responses were dummy-coded and correlated with agency and experience attributions before exposure to the vignettes (see Figure 2.2). These correlations showed that both agency, $r(146) = 0.174$, $p = .034$, 95% CI [0.013, 0.326], and experience, $r(146) = 0.382$, $p < .001$, 95% CI [0.235, 0.512], were significantly correlated with prior exposure, such that attributions of mind increased with prior exposure. These results conceptually mirror, and reinforce, the main manipulation and core finding of the present study—that participants increase attributions of mind after exposure to ChatGPT. Following these findings, an additional ANCOVA was conducted with attribute (agency/experience) and timepoint (pre/post) as factors with prior exposure as a covariate. Main effects were found for attribute and timepoint matching the paired samples t-tests; no interaction was observed, $F(1, 147) = 3.00$, $p = .085$, GES =.000.

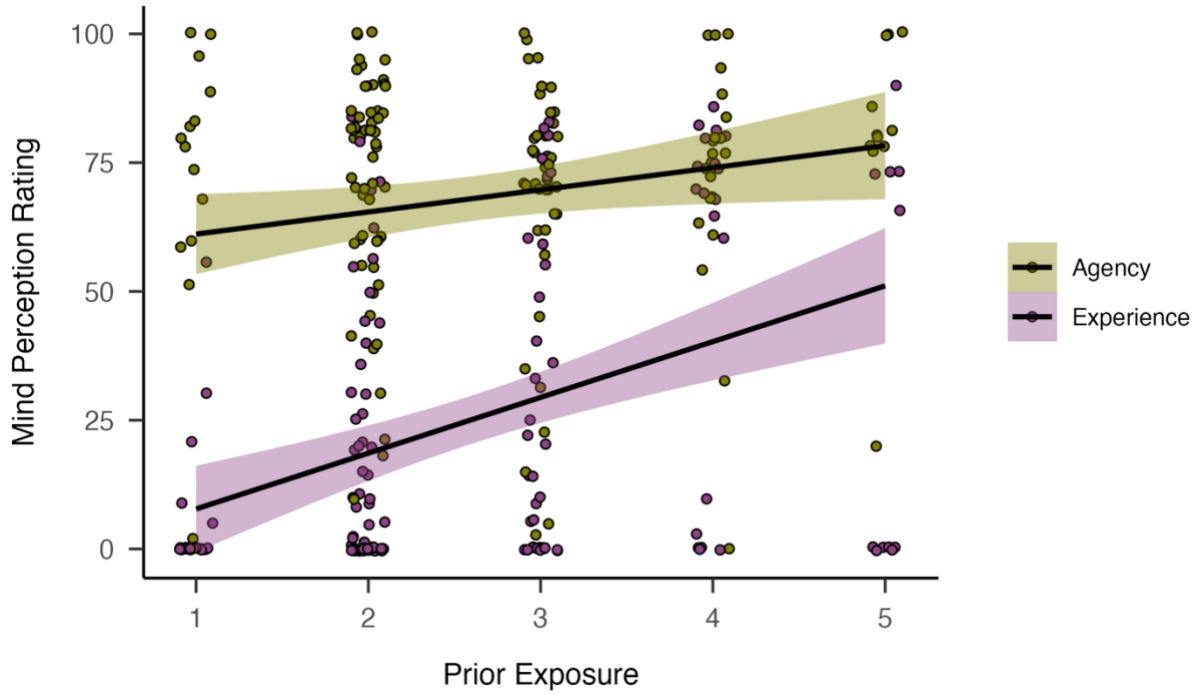

Figure 2.2: Agency and experience ratings by prior levels of exposure. The shaded area represents standard error.

Individual Differences

In addition to prior exposure, a number of other individual difference measures were recorded, including age, education, gender, AQ scores, IDAQ scores, and Big-Five personality traits. While a number of these measures were correlated with pre- and post-exposure assessments of mind perception (see Table 1 in Appendix), most did not relate to *changes* in mind perception following exposure to ChatGPT. The lone exception was the IDAQ (anthropomorphism) score, which was associated positively with a change in perceived experience (the ability to feel), $r(146) = 0.189$, $p = .022$, 95% CI [0.028, 0.340]. However, there was no comparable or significant effect for perceived agency, $r(146) = -0.085$, $p = .305$, 95% CI [-0.243, 0.078]. See Table 2.1.

Table 2.1: Individual differences and their correlations with changes in mind perception (agency and experience). *α* indicates Cronbach's alpha for the corresponding scale.

	Agency Change	Experience Change
Age	0.05	-0.09
Education	-0.03	-0.09
Prior Exposure	-0.01	0.07
AQ-10 Total (α = .66)	0	0.05
IDAQ Total (α = .89)	-0.08	0.19 *
Openness	0.01	0.01
Extraversion	0.07	0.04
Agreeableness	-0.13	0.03
Conscientiousness	0.04	0.16
Emotional Stability	0.01	-0.03

Discussion

Our findings in Experiment 1 indicate that participants significantly increased their mind perception of ChatGPT—both in terms of what it can do (its agency) and what it can feel (its experience) after being given short examples of its capabilities. These results fit our predictions that mind perception would increase post-exposure (H1). Although there was no interaction between attribute (agency/experience) and timepoint, the relative change was numerically greater for agency than experience which agrees with H2 and prior findings comparing agency and experience attributions to machines or AI (e.g., Lakatos et al., 2014; Spatola et al., 2021).

These changes in mind perception are particularly noteworthy for at least two reasons. First, they occurred after only a very brief exposure (i.e., the presentation of 3 examples). Second, participants did not actually interact with ChatGPT themselves, but were simply shown examples of it in operation. Thus, the present findings are likely to be conservative estimates of

how mind perception toward AI systems such as ChatGPT will change as individuals' exposures to, and interactions with, these devices increase.

Convergent with this idea is our finding that individual differences in prior exposure to ChatGPT were associated with higher agency and experience attributions, mirroring the results of the main manipulation of our study. Our study also suggests a relationship between an individual's propensity to anthropomorphize and changes in the degree of mind perception toward AI. Consistent with previous research (Epley et al., 2007; Waytz et al., 2010a), we found that IDAQ scores, a measure of individual propensity to anthropomorphize, were associated with changes in mind perception ratings after exposure and were related specifically to attributions of experience (the ability to feel). This suggests that individuals who are more likely to anthropomorphize are more likely to increase their perception that ChatGPT can feel, even after just a brief exposure. However, this effect did not extend to perceptions of agency (the ability to do), perhaps reflecting that the IDAQ scale captures more of a propensity for individuals to attribute experiential capabilities to nonhumans as opposed to agentic capabilities.

As for our finding that a large number of individual difference measures were not associated with changes in mind perception, although suggestive, we hasten to note that our study was purposely designed to be a conservative test of the effect that brief exposure to ChatGPT would have on mind perception. While it is possible that our findings will prove to be robust, it is also possible that different personality traits will respond differently to sustained exposure to and/or interactions with AI in the future. In general, however, the present experiment's conservative approach provides compelling evidence that different propensities to anthropomorphize play a significant role in the mind perception of AI.

Experiment 2

In order to better understand the generalizability of the findings from Experiment 1, we chose to conduct a second study aimed at understanding if the vignettes used in Experiment 1 provide a similar exposure effect as users generating their own prompts, as they would typically do. To this end, participants were recruited from the University of British Columbia (UBC) to go through the same mind perception pre-post exposure manipulation, with the key difference being that participants were tasked with generating and asking their own 3 questions in a ChatGPT window. Individual differences were not assessed due to the smaller sample associated with running the in-person design. We hypothesize (H3) that the changes in mind perception would be similar to Experiment 1 but will be relatively larger given that the exposure in Experiment 2 stems from real-time interaction with ChatGPT, rather than vignettes, which should be more demonstrative of the capabilities of ChatGPT.

Methods

Participants

A total of 55 undergraduate participants were recruited from the Human Subject Pool at the University of British Columbia following a G*Power analysis suggesting a sample size of 54 for detecting a medium-sized effect (d=.5) between two independent means of matched pairs with 95% power. A number of participants, however, were excluded. These included 14 for failing to follow instructions to enter the correct number of prompts. Participant exclusions made no difference to what effects were significant but the following analyses contain only the 41 remaining participants. There were 10 males and 31 females. The mean age was 20.8 years of age (SD = 2.0).

Materials and Measures

The same materials and measures were used as in Experiment 1.

Procedure

Participants came into the lab and after providing written consent were told that they will be completing two surveys. The first survey was similar to the pre-survey in Experiment 1 where they were asked to provide demographic details before rating ChatGPT on the mind perception measures and their prior exposure to LLMs. After completing these questions, a research assistant told them they would be interacting with ChatGPT on a local computer. They were given the task of entering 3 prompts after which participants were asked to rate ChatGPT again on the mind perception measures. Participants were then thanked, debriefed, and compensated for their efforts.

Results

Mind Perception Attributions

Paired samples t-tests were conducted to compare the mind perception scores before and after entering prompts into ChatGPT. There was a significant increase in agency scores (the ability to do) after the prompts (M = 75.85, SD = 20.48), $t(40) = -3.79$, $p < .001$, d = .27, 95% CI [0.13, 0.42], compared to before the prompts (M = 70.32, SD = 19.74). There was no significant difference in experience attributions from before (M = 22.68, SD = 19.53), and after the prompts (M = 24.95, SD = 24.76), $t(40) = -0.82$, p =.41, d =.10, 95% CI [-0.14, 0.34].

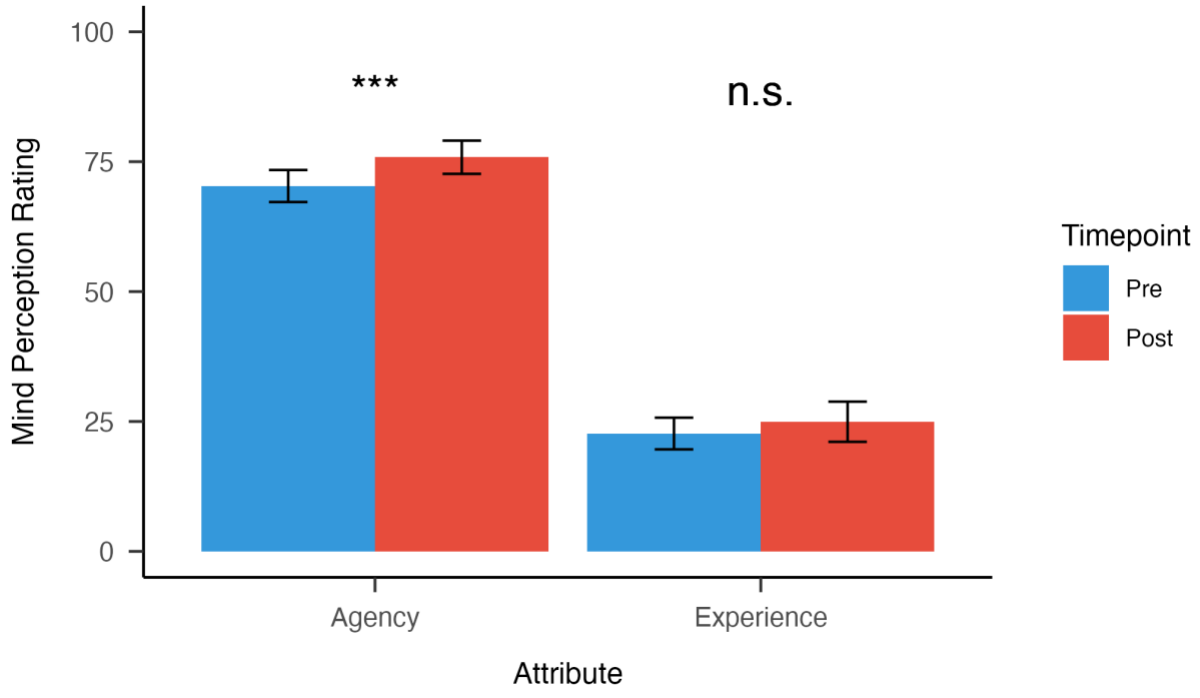

Figure 2.3: Experiment 2 mean agency and experience ratings pre- and post-exposure. Error bars represent standard error and *** indicates significance at p <.001.

Discussion

Our findings from Experiment 2 highlight several key ideas. First, the generative method in Study 2 replicates the effects of exposure using the vignette method in Study 1. Participants attributed greater agency to ChatGPT after exposure. However, we failed to find an effect of exposure on attributions of experience to ChatGPT, although post-exposure experience ratings were higher than pre-exposure. This leads to the second, and perhaps more interesting finding, which is that the presumption that the pre-generated responses were a conservative form of exposure was not supported. Contrary to our hypothesis (H3), the self-generative method in Study 2 did not lead to larger exposure effects. One possible reason for this is that the vignettes

used in Experiment 1 were generally more creative in nature than the average set of questions that users in Study 2 came up with (e.g., many prompts in Experiment 2 were basic factual questions like “How old is Elon Musk?”). In other words, the pre-generated responses may have been a better demonstration of ChatGPT's attributes of mind vis-a-vis agency and experience than the questions users generated in Experiment 2. In both studies, however, the overall exposure effect remained conservative in the quantity of exposure as both only contained just 3 responses from ChatGPT.

Experiment 3

Following the results of Experiments 1 and 2, a third experiment replicating Experiment 1 with other LLMs became desirable for improving the generalizability of the present findings with a new sample. Recent research has suggested that different LLMs can vary significantly in the degree to which they display human biases or task strategies (Binz & Schulz, 2023; Chang et al., 2024) suggesting that different LLMs could also significantly differ in the amount of mental attribution they elicit from users. To this end, participants were recruited using the same methodology as Experiment 1, except this time they were randomly assigned to view 3 responses from either ChatGPT (identical to Experiment 1), LLaMA (Meta, 2024), or Claude (Anthropic, 2024). The responses were generated using the same prompts between ChatGPT, LLaMA, and Claude. We hypothesized that the general pattern of results would be similar to Experiment 1 (H4).

Methods

Participants

A new power analysis was conducted for Experiment 3 to reflect the between-subjects design (i.e., participants were randomly assigned to ChatGPT, Llama, or Claude). The G*Power

power analysis for a 2 (within) x 2 (within) x 3 (between) ANOVA for a within-between interaction (*f*=.1, power = 95%) suggested a sample of 264 participants. In total, 265 participants (133 female; Mean age = 34.6, SD age = 11.8) took part after one participant was excluded for failing a basic attention check. Participants reported their highest education: 28 stated 'High school graduate (high school diploma or equivalent including GED)', 55 stated 'Some college but no degree', 29 stated 'Associate college degree (2-year)', 94 stated 'Bachelor's degree in college (4-year)', 41 stated 'Master's degree', 10 stated 'Doctoral degree', and 5 stated 'Professional degree (JD, MD)'.

Materials and Measures

All measures were the same as Experiment 1. New vignettes (3 per model) were created using real responses from LLaMA (v3.2 70B; Aug, 2024) and Claude (v3.5 Sonnet, Aug, 2024) based on the same prompts used to generate the responses in Experiment 1 with ChatGPT (see https://osf.io/rxpvm/).

Procedure

The procedure closely matched Experiment 1. Participants were recruited through Prolific. After obtaining consent, participants were asked to fill out demographic details (age, gender, education) before being asked to rate the amount of agency and experience they attribute to LLMs. Then, participants were randomly assigned to view the vignettes from ChatGPT, LLaMA, or Claude before being asked to answer the same mind perception questions. Participants were then asked to fill out the individual differences measures also collected in Experiment 1. After completing the study, participants were thanked and compensated for their efforts.

Results

Mind Perception Attributions

A 3-way ANOVA with attribute (agency/experience), timepoint (pre/post), and model (ChatGPT/LLaMA/Claude) was conducted. For the main effects, timepoint, $F(1,262)= 35.69$, $p <001$, $\hat{\eta}_G^2= .006$, and attribute, $F(2,262)= 747.33$, $p <001$, $\hat{\eta}_G^2 = .506$, were significant but model was not, $F(2,262)= 0.57$, $p =.567$, $\hat{\eta}_G^2 = .002$. There was also a significant attribute by timepoint interaction, $F(1,262)= 11.81$, $p <001$, $\hat{\eta}_G^2= .002$, and a significant model by timepoint interaction, $F(2,262)= 3.90$, $p = .021$, $\hat{\eta}_G^2 = .001$. The other two-way interactions and the three-way interaction were not significant, p's > .101. An ANCOVA model with the same factors but including prior exposure as a factor yielded the same pattern of results. Follow-up pairwise comparisons on the ANOVA with Tukey corrections revealed that agency, $t(262) = -7.17$, $p < .001$, and experience, $t(262) = -2.00$, $p = .047$, were significantly higher post-exposure, agency attributions were greater than experience attributions (Pre: $t(262) = 25.85$, $p < .001$; Post: $t(262) = 26.39$, $p < .001$), and attributions were independently higher post-exposure for each model (ChatGPT: $t(262) = -2.50$, $p = .013$; LLaMA: $t(262) = -2.12$, $p = .035$; Claude: $t(262) = -5.74$, $p < .001$). These findings are visualized in Figures 2.4 and 2.5.

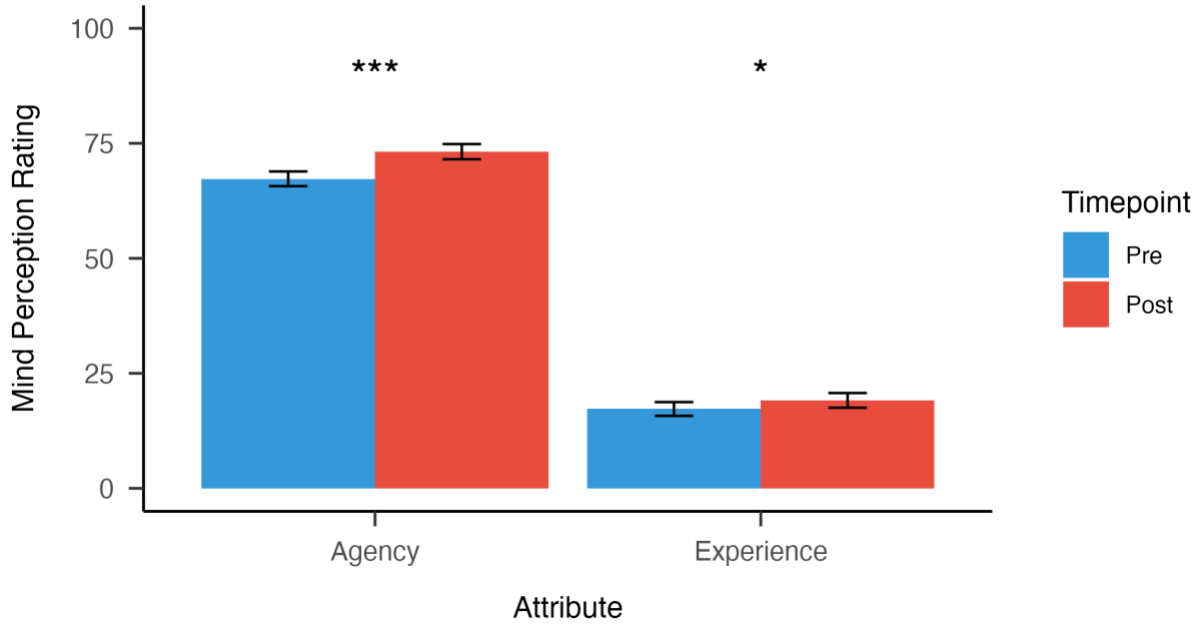

Figure 2.4: Experiment 3: Mean agency and experience attributions pre-and post-exposure. Error bars indicate standard error and asterisks indicate level of significance. *** = $p < .001$, * = $p < .05$.

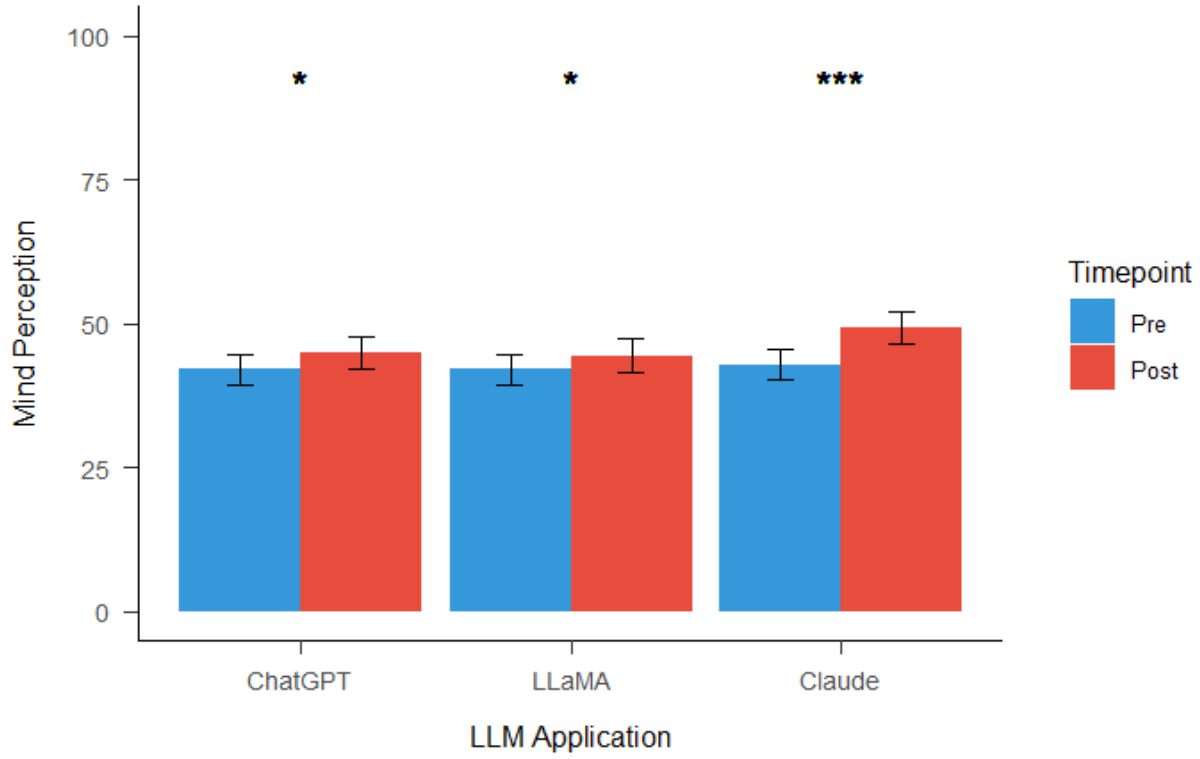

Figure 2.5: Experiment 3: Mean mind perception attributions pre-and post-exposure by model. Error bars indicate standard error and asterisks indicate level of significance. *** = p <.001, * = p <.05.

Individual Differences

Similar to Experiment 1, a number of other individual difference measures were collected from participants, including age, education, gender, AQ scores, IDAQ scores, and Big-Five personality traits. While a number of these measures were correlated with pre- and post-exposure assessments of mind perception again (see Table 2 in Appendix), most did not relate to *changes* in mind perception following exposure to ChatGPT. For changes in agency, however, age was negatively correlated, $r(263) = -0.120$, $p = .050$, 95% CI [0-.238, -0.000], as was prior exposure, $r(263) = -0.165$, $p = .007$, 95% CI [-0.280, .-0.045]. For changes in experience, agreeableness was positively correlated, $r(263) = 0.147$, $p = .016$, 95% CI [0.027, 0.263].

Table 2.2: Experiment 3: Individual differences and their correlations with changes in mind perception (agency and experience). *α* indicates Cronbach's alpha for the corresponding scale.

	Agency Change	Experience Change
Age	-0.12*	0.07
Education	-0.11	0.00
Prior Exposure	-0.16*	-0.04
AQ-10 Total (α = .59)	0.05	-0.01
IDAQ Total (α = .87)	-0.03	0.03
Openness	-0.02	0.01
Extraversion	-0.10	-0.02
Agreeableness	-0.07	0.15*
Conscientiousness	-0.04	0.02
Emotional Stability	-0.10	0.02

Discussion

Using a new sample of individuals, Experiment 3 replicated the main findings of Experiment 1—that participants increase their attributions of mind after being given short vignettes demonstrating the capabilities of LLMs, with the effect of agency exceeding that of experience overall, and showing the greatest change across time. This pattern was observed similarly for all three models—ChatGPT, LLaMA, and Claude, though Claude was found to yield the greatest change across time in perceived mind (see Figure 2.5). These findings suggest that the exposure effects leading to greater mind perception are robust and generalize across different LLMs. Regarding individual differences, many relationships for pre- and post-exposure ratings replicated Experiment 1 (e.g., prior exposure, IDAQ scores, see S1 & S2). Again, however, individual measures generally failed to predict the strength of the exposure effect on mind perception ratings. Notably, unlike Experiment 1, IDAQ scores did not predict greater

increases in experience attributions post-exposure, which might be related to sample differences between experiments or a more complex relationship between IDAQ scores and prior exposure.

Experiment 4

Experiment 3 replicated the key findings of Experiment 1—that mind perception increases for agency and experience following exposure to LLMs, with the effect being larger for agency than experience. Experiment 3 also extended these findings to other LLM models.

Experiment 2 also found an effect of exposure for agency, but in contrast to Experiments 1 and 3, it failed to find an effect for experience. We had speculated that this failure to find an effect of experience may have been due to participants' limited number of real-time interactions with ChatGPT or due to the tendency of participants to prompt for factual information rather than more demonstrative, creative responses.

In Experiment 4, participants were once again recruited to interact in real-time with ChatGPT but this time without the constraint of entering only 3 prompts. Participants were given 15 minutes to enter as many prompts as they desired after they filled out the pre-survey questions. We hypothesized that with the longer duration of real-time interaction with ChatGPT, participants would show an exposure effect for both agency and experience (H5).

Methods

Participants

The same power analysis as Experiment 2 suggested 52 participants take part. 52 students (43 female; MAge = 21.1, SDAge = 4.77) were recruited and none were excluded from analyses.

Materials and Measures

All materials and measures were the same as Experiment 2.

Procedure

The procedure for Experiment 4 closely matched Experiment 2. After consenting to take part in the experiment, the participants were asked to fill out a pre-survey. The pre-survey included the same mind perception measures as the previous experiments. Then, participants were instructed to converse freely with ChatGPT and that they could discuss whatever topics they wished. After 15 minutes, participants were asked to fill out the post-survey that matched Experiment 2. Participants were thanked and given course credit. The number of prompts entered by each participant was tallied and added to their survey data by the researchers.

Results

Mind Perception Attributions

Paired samples t-tests were conducted to compare agency and experience attributions pre- and post-exposure. Agency from pre-exposure (M = 64.50, SD = 23.55) to post-exposure (M = 65.87, SD = 24.92) did not differ significantly, $t(51) = -0.79$, $p = .044$, $d = .06$, 95% CI [-0.09, 0.20]. Experience from pre-exposure (M = 8.73, SD = 16.35) to post-exposure (M = 10.63, SD = 18.93) also did not differ significantly, $t(51) = -1.42$, $p = .0162$, $d = .10$, 95% CI [-0.04, 0.25] . These data are visualized in Figure 2.6.

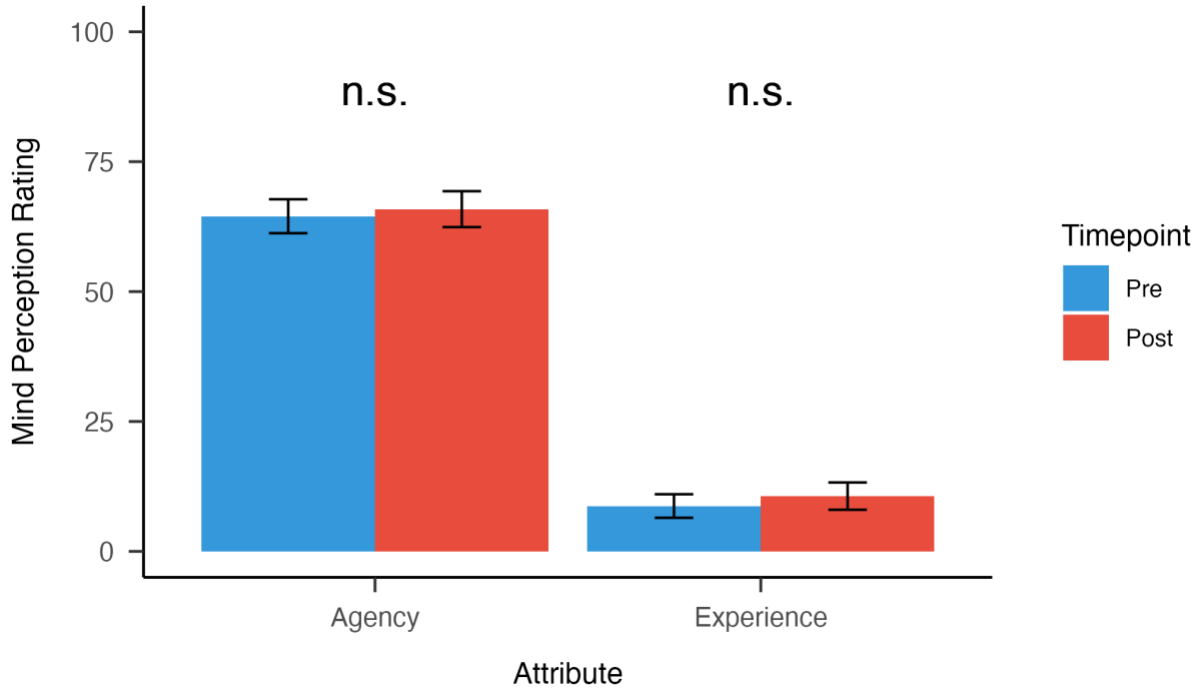

Figure 2.6: Experiment 4: Pre-and post-exposure mind perception ratings.

Prompt Frequency

Overall, participants entered an average of 21.12 number prompts (SD = 12.63). The number of prompts entered by each participant was used to predict each individual's change in agency and experience via regression models. For agency (β = -.16, SE = .14, $t(50) = -1.17$, $p = .248$), the number of prompts participants entered did not predict the difference between pre-and post-attributions. This was also the case for experience attributions (β = -.00, SE = .11, $t(50) = 0.72$, $p = .998$) indicating that participants who entered more prompts were not more likely to show an exposure effect.

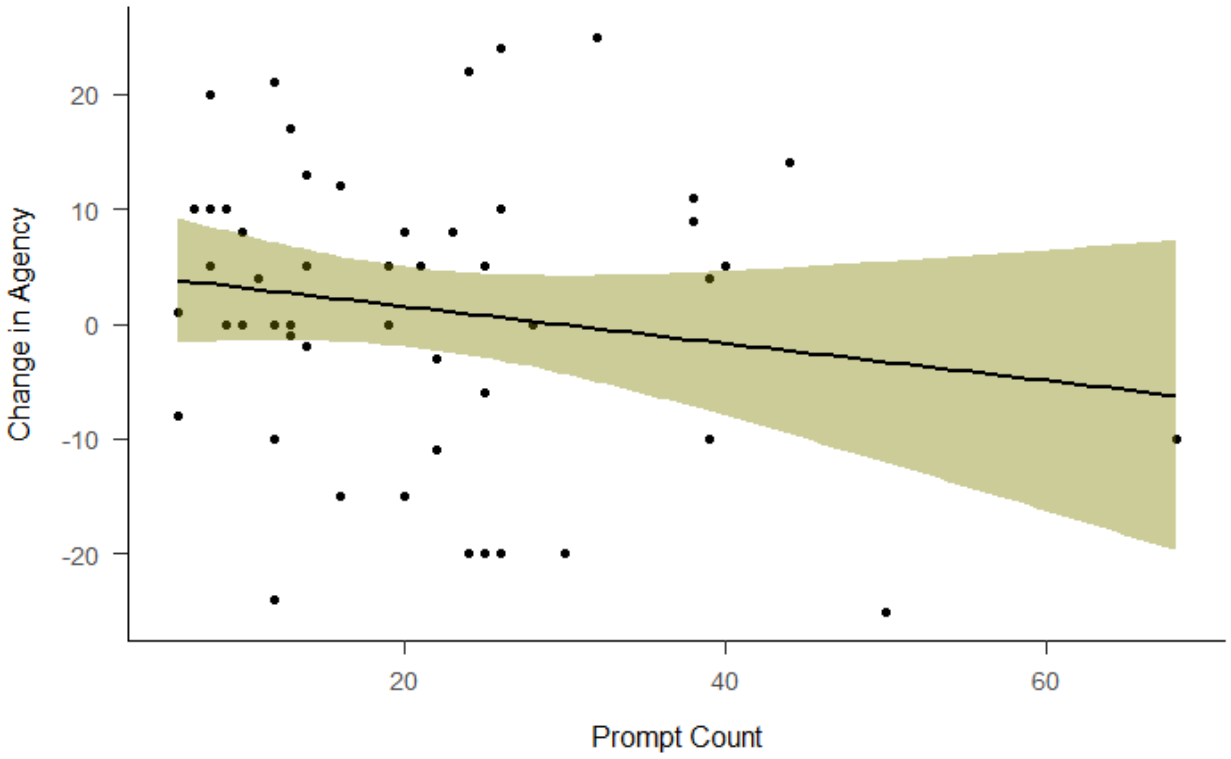

Figure 2.6: Experiment 4: Change in agency in relation to prompt count. The shaded area represents standard error.

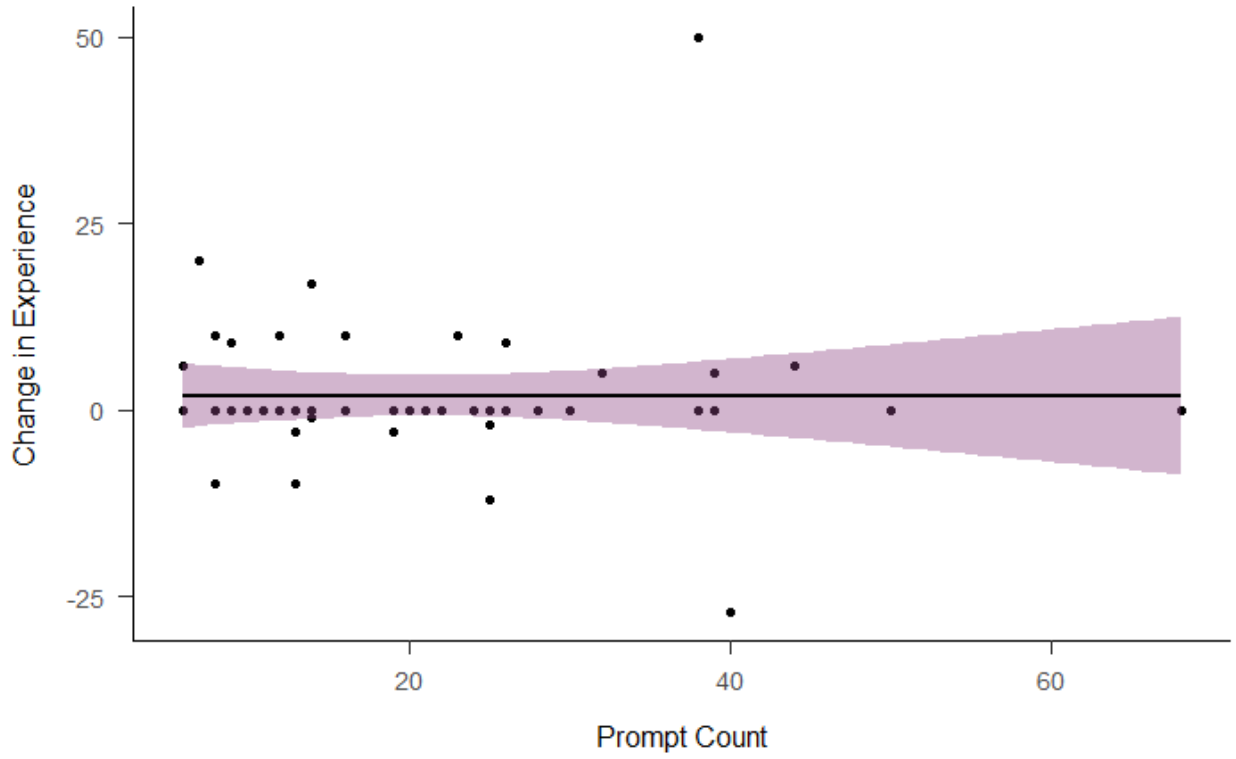

Figure 2.7: Experiment 4: Change in experience in relation to prompt count. The shaded area represents standard error.

Discussion

When participants were limited to 3 real-time interactions with ChatGPT in Experiment 2, their perception of its agency increased, but not their perception of its experience. In Experiment 4, their opportunity to interact with ChatGPT was increased, but now there was no change for agency in addition to experience. In short, it appears that greater real-time interaction with ChatGPT had no impact on people's perception of what it can feel, and negates any initial heightened perception of what it can do (i.e., the agency attribution). This latter conclusion is consistent with Experiment 4's elimination of the perceived change in agency observed in

Experiment 2, and with the negative slope in Figure 2.7, suggesting that the perception of agency in ChatGPT declined as the number of real-time interactions with it increased.

These results support the idea raised at the end of Experiment 2 that self-generated prompts do not lead to larger exposure effects. Even though participants did interact with ChatGPT for a longer duration in Experiment 4, many of the interactions were similar to Experiment 2's interactions which consisted of more *tool-oriented* or *utility-oriented* interactions rather than more complex conversational dynamics. For example, one participant asked: "what is the population of Vancouver", while another asked "why does the east coast have such bad weather". Perhaps not surprisingly, engaging in these utility-oriented interactions for a longer period of time merely serves to re-confirm participants' pre-existing beliefs about what an LLM can do, resulting in no perceived change in its agency. And of course, engaging in utility-oriented interactions fails again to move participants' perception of its experience. In summary, the results of Experiment 4 indicate that exposure to LLMs need not lead to greater mental attribution toward LLMs.

General Discussion

The emergence of highly capable and highly accessible large language models (LLMs) has underscored the need to better understand how people perceive minds in AI systems, especially as researchers seek to examine their use for implementation in social robots for physical and health interventions. The systematic study of how people perceive minds (Gray et al., 2007) and anthropomorphize nonhumans (Epley et al., 2007) has grown considerably in past decades and fits with growing calls to conduct psychological investigations to examine how people perceive complex AI systems (e.g., Binz & Schulz, 2023; Kosinski, 2024). The present study investigated the impact of brief exposure to ChatGPT on mind perception across four

experiments that differed in the nature of the exposure (vignettes vs. real-time interaction). With the larger sample size afforded by Experiments 1 and 3, we also examined the influence of individual differences on the effect of exposure (age, gender, prior exposure, Big-5 personality, autistic traits, and individual propensities to anthropomorphize).

Our findings in Experiment 1 showed that when participants were shown 3 vignettes demonstrating the range of ChatGPT's capabilities, participants significantly increased their mind perception of ChatGPT—both in terms of what it can do (its agency) and what it can feel (its experience). In Experiment 2, participants engaged in 3 real-time interactions with ChatGPT and we again found an exposure effect for agency but failed to find differences for attributions of experience. This divergence in experience attributions was particularly intriguing given that we expected to find larger exposure effects when ChatGPT exposure was interactive and non-scripted, unlike the vignettes in Experiment 1. Our working hypothesis is that the vignettes used in Experiment 1 were actually more demonstrative of ChatGPT's capabilities than the fact-based questions asked by many participants in Experiment 2. This tendency of participants to ask more fact-based questions may reflect pre-existing assumptions that ChatGPT is a computer program akin to Siri or Alexa, and as such, it can only handle questions related to facts and figures (e.g., "What is the capital of Hungary?"). We refer to this as the users taking a more *utility-orientated* approach to conversing with LLMs. In contrast, the vignettes that were selected for Experiment 1 were designed to demonstrate some of the creativity that ChatGPT is capable of (e.g., the vignette about a cover letter for a job on the moon). Thus, one interesting takeaway is that self-generated prompts seemed to reveal what users expect from a chatbot. In other words, the apparent lack of creativity in self-generated prompts could be reflecting people's general

intuition that AI systems are much more capable of agentic abilities (i.e., knowledge-based) rather than experiential or emotional abilities that require more abstract reasoning.

In Experiment 3, we sought to replicate Experiment 1 but this time also included vignettes containing responses generated by different models (ChatGPT, LLaMA, and Claude). The results of Experiment 3 supported the findings of Experiment 1 but also showed that the exposure effect can be modulated by the specific model or vignettes shown to participants. We found that Claude elicited the greatest exposure effect on mind perception (see Figure 2.6). The exact reason for this is unknown, but this could be related to the first-person language Claude used, unlike ChatGPT or LLaMA.

Experiment 4 sought to probe if the diverging effects of Experiment 2 were a function of the very brief number of prompts participants were allowed to self-generate which was done to match the number of vignettes shown to participants in Experiment 1 (and Experiment 3). In Experiment 4, participants were given 15 minutes to generate as many real-time interactions with ChatGPT on whatever topic they wished. The results were that neither agency nor experience attributions increased post-exposure. The style of prompts entered by participants in Experiment 4 were very similar to Experiment 2. Participants often took the more utility-oriented approach to interacting with ChatGPT (e.g., asking fact-based questions), rather than having a more complex, genuine social conversation. These results reinforce the idea that simply interacting with LLMs does not necessitate a change in mind perception. If users believe that LLMs only have agentic capabilities—there may be a confirmatory bias that emerges from users only probing LLMs for more fact-based, agency-oriented responses. In other words, exposure to LLMs can increase mind perception attributions but the nature of the exposure matters; more creative, or demonstrative types of exposure will lead to greater changes in mind perception even at the cost

of being non-interactive exposure. Understanding how finer-tuned changes to LLMs can elicit different degrees of mental attribution to LLMs is a promising future direction especially as recent developments in research toolkits (e.g., Laban et al., 2024b) are making it easier for researchers to explicitly manipulate some of these variables (e.g., spontaneity or creativity in responses) that may affect interactions with LLMs and subsequent mental attributions.

Despite differences between the interactive (Experiments 2, 4) and non-interactive (Experiments 1, 3) investigations, the main finding across experiments supports the hypothesis (H1) that mind perception toward LLMs is influenced by exposure. This was notable in that changes in mind perception occurred after only a brief exposure (i.e., the presentation/interaction of 3 occurrences). The diverging results also suggest that mere exposure does not necessarily lead to greater mental attribution toward LLMs—the type of exposure matters. This fits with recent research findings that mind perception toward LLMs can be influenced by the instructions given to an LLM (Laban et al., 2024b), which would change the nature of the exposure. It is also noteworthy that participants attributed greater agency than experience and that the change in attributions as a function of exposure was larger for agency than experience, reflecting the general finding that people are more reluctant to ascribe experiential qualities to AI than agentic qualities (Gray et al., 2007; Waytz et al., 2010a; Malle et al., 2017).

A key complement to the main finding that exposure increased mind perception toward ChatGPT was that individual differences in prior exposure to ChatGPT were associated with higher agency and experience attributions (Experiments 1 and 3, see Supplementaries 1 and 2): in effect, mirroring key results of our study. The results align with some prior research suggesting that exposure to AI and robots can influence the degree of anthropomorphism attributed to these entities (Waytz et al., 2010a). It also suggests that as people become more

familiar with LLMs they may attribute more mind to LLMs. However, it is hard to predict how AI perception will change in the future. For example, one could equally imagine that longer-term exposure toward AI would actually decrease mind perception as people could over time learn to shift explanations of behaviour from *intentional* to *designed* (Dennet, 1988), with the latter being less associated with mind perception. Nonetheless, the present results suggest, at least initially, that mind perception afforded to LLMs like ChatGPT increases with exposure.

Experiments 1 and 3 also revealed a relationship between an individual's propensity to anthropomorphize and the degree of mind perception afforded toward AI. Consistent with previous research (Epley et al., 2007; Waytz et al., 2010a), we found that IDAQ scores, a measure of individual propensities to anthropomorphize, were associated with mind perception ratings before and after exposure. Experiment 1 also found that IDAQ scores were associated with greater increases in mind perception from pre-to post-exposure. This suggests that individuals who are more likely to anthropomorphize are also more likely to increase their perception that ChatGPT can feel, even after just a brief exposure. However, this effect did not extend to perceptions of agency (the ability to do), perhaps illustrating that the IDAQ scale captures more of a propensity for individuals to attribute experiential capabilities to nonhumans as opposed to agentic capabilities. Moreover, it is worth noting that experiential capabilities are considered to be the more human-like factor of mind perception (Gray et al., 2007) both because it captures greater variance in a variety of mind perception attributions, and because it is considered more quintessential to what makes humans unique (Jacobs et al., 2024). Additionally, this relationship between IDAQ scores and the exposure manipulation did not occur in Experiment 3. This could be related to a more complex interaction involving prior exposure and IDAQ scores. Participants in Experiment 3 had greater prior exposure to LLMs (in part because

of the one-year time period between experiments). Perhaps the increase in mind perception for individuals with higher IDAQ scores becomes more leveled out after a certain amount of exposure. Nonetheless, the observed relationships between individual tendencies to anthropomorphize and mind perception specifically toward LLMs, build on existing literature regarding individual differences and attitudes toward AI by expanding the scope to which interactions with AI are affected by anthropomorphism (Rossi et al., 2020; Złotowski et al., 2014). The degree to which individual differences are associated with mind perception to ChatGPT for more generative, real-time interactions remains a topic for future work but should also include measures related to prior exposure.

In conclusion, our research demonstrates that exposure to ChatGPT can increase mind perception toward LLMs, with people's likelihood to anthropomorphize being associated with the magnitude of this effect. These results emerged after only a brief descriptive or real-time interactive exposure to ChatGPT. Critically, across the four experiments, the pattern of results also indicated that the nature of exposure influences how mind perception can change with exposure to LLMs (and even negate changes in mind perception (Experiment 4)). In sum, the present results suggest that the merest of exposure to AI systems may significantly increase people's perceptions of mind, that it depends on the type of exposure, and that the effect of exposure is stronger for agentic rather than experiential attributions. This is an encouraging indicator of the possibility that social robots that incorporate LLMs may have a positive and immediate impact in private and public institutions where establishing a quick social connection is crucial (e.g., customer service, health care agencies, and the like) but future research will need to investigate the longer-term effects of exposure and mind perception. Finally, we suggest that by continuing to consider individual differences in mind perception, AI developers can create

more effective and user-friendly systems that cater to a diverse range of users with varying backgrounds and propensities to anthropomorphize.

Chapter 4:

Chapter 3 found that people frequently ascribe agentic and experiential qualities to both social robots and LLMs. Chapter 4 turns to focus on a particular individual difference known to be related to anthropomorphism—loneliness or social isolation. Chapter 4 contains one empirical study which is a version of a manuscript submitted for publication.

Study 3: Perceiving AI Minds: Loneliness increases attributions of feelings, but not agency, to Large Language Models

Introduction

The stark prevalence of loneliness, the negative emotional response to perceived social isolation, has been identified as a significant crisis and growing public health problem (Cacioppo & Cacioppo, 2018; Gerst-Emerson & Jayawardhana, 2015; Mann et al., 2022). Loneliness is associated with a range of poor health outcomes such as an increased risk of stress levels (Adam et al., 2006; Lee & Goldstein, 2016), depression (Heinrich & Gullone, 2006; Mann et al., 2022) and suicide (McClelland et al., 2020). Researchers have explored the tendency to anthropomorphize, that is to attribute human qualities to nonhumans, as a means of coping with social isolation (Epley et al., 2007). Epley et al. (2007, 2008a) argue that the need for social connection is a critical predictor of when and why people anthropomorphize and a variety of research has shown a relationship between loneliness and anthropomorphism (Bartz et al., 2016; Eyssel & Reich, 2013; Shin & Kim, 2020). For example, Epley et al. (2008b) demonstrated through both non-experimental and experimental methods that greater loneliness was associated with greater anthropomorphism toward gadgets, gods, and animals.

This connection between a higher need for social interaction and greater anthropomorphism has also been demonstrated in autonomous cars (Waytz et al., 2014), and

more relevantly, in the case of chatbots (Sheehan et al., 2020). However, this connection has yet to be made in the case of large language model (LLM) chatbots which differ considerably from previous generations of chatbots both in terms of their ability to imitate human language and their real-world applications. Previous chatbots typically were relegated to niche roles such as facilitating customer service interactions. Unlike this previous generation of chatbots, ChatGPT has been endorsed and used by thousands of individuals for more complex interactions such as aiding one's mental health often as a substitute for a real therapist.

While the novelty of LLM applications means that there is still little scientific work investigating how LLMs interact with people's loneliness, there has been some work investigating the degree to which people anthropomorphize LLMs. Mind perception is one common method of investigating anthropomorphism as it can capture the degree to which people attribute qualities of minds to humans, animals, robots, and inanimate things. One popular method of measuring mind perception was developed by Gray et al. (2007) in which they distilled mind perception into 2 key factors, coined agency and experience. These two factors map, respectively, onto the cognitive and emotional elements of mind. Jacobs et al. (2023) revealed that people increase the degree to which they anthropomorphize ChatGPT, even after brief exposure, and in addition to their baseline levels of prior exposure. Although the need for social connection is considered to be a key factor in anthropomorphism (Epley et al., 2008b), it has not yet been discovered the degree to which the need for social connection influences the relationship between anthropomorphism and LLMs.

The present study investigates the role of loneliness and the need for social interaction as a predictor for tendencies to anthropomorphize LLMs. Greater clarity in this relationship would provide more insight into the kinds of people that may be most susceptible to blurring the lines

between speaking with an LLM and speaking with a real person. The prescient need for this line of work is altogether clearer given the popularity of ChatGPT being used similarly to a mental health therapist and the ongoing loneliness health crisis.

Participants were recruited online to assess self-reported loneliness and their associated ratings of the anthropomorphic qualities of ChatGPT. In addition, participants' prior exposure with ChatGPT was measured in order to investigate its role as a potential moderator given past work finding an association between anthropomorphism and prior exposure.

Methods

Participants

Sample size was determined using an a priori power analysis in G*Power for multiple regression. For a small-to-medium sized effect (f=.10), 95% power, and 2 predictors, the ideal sample size was 132 people. In total, 150 participants were recruited through CloudResearch to take part in the study with oversampling occurring due to anticipated data exclusions. 10 participants failed an attention check. As a result, 140 participants were included in the data analysis (M_{Age} = 39.2, SD_{Age} = 11.5; 92 males, 48 females) and took part remotely from IP addresses listed in the United States.

Procedure

Participants were informed of the nature of the study and provided written consent to take part. First, participants were asked to input their age and select their gender and relationship status. Participants were then asked "How much exposure have you had with AI chatbots such as ChatGPT?" which was a 5-point Likert scale ranging from 'None at all' to 'A great deal'. Participants were then presented with a screen stating "Using your intuition, please rate what you think of ChatGPT on the following scales" and "To what degree do you think AI chatbots (such

as ChatGPT) exhibit these abilities?". Two scales were presented which ranged from 0-100 with details stating "0 - indicates none or very little" and "100 – indicates the max or to a very strong degree". The first scale was labeled "Agency (the ability to do things)" and the second scale was labeled "Experience (ability to feel things)". After providing these ratings, participants were tasked with answering 15 items from the Individual Differences in Anthropomorphism Questionnaire (IDAQ). Included was an attention check 2/3rds of the way through stating "Please use the third bubble from the left to answer this question". Finally, participants were asked to fill out the Three-Item Loneliness Scale (Hughes et al., 2004) which has possible responses ranging from 'Hardly ever' to 'Often' and has previously been used in anthropomorphism studies (e.g., Epley et al., 2008a). After participating, participants were thanked and compensated for their efforts.

Data Analysis

All data analyses were conducted in R (v4.0.5; R Core Team 2021) using the R packages dplyr (Wickham et al., 2014) and lm4r (Bates et al., 2003) in addition to core packages.

Results

Pearson correlations were first run to capture the associations between agency, experience, loneliness, IDAQ, and prior exposure measures. The correlations along with their significance are presented in Table 3.1. In summary, agency, experience, loneliness, prior exposure, and general anthropomorphic tendencies (IDAQ) were positively associated with each other. However, the relationship between agency and loneliness scores was not significantly correlated.

Table 3.1. Cross correlations between dependent and individual measures. Asterisks indicate the level of significance. * p <.05, ** p <.01, *** p < .001.

	Agency	Experience	Loneliness	Prior Exposure	IDAQ Total
Agency	1				
Experience	0.22 **	1			
Loneliness	0.02	0.28 ***	1		
Prior Exposure	0.17 *	0.44 ***	0.27 **	1	
IDAQ Total	0.22 *	0.58 ***	0.25 **	0.29 ***	1

Next, two models (a main effects model and an interaction model) containing loneliness and prior exposure levels were regressed onto agency and experience attributions separately. To ensure multicollinearity was not distorting model fits, Variance Inflation Factors (VIFs) were computed for the predictors (all VIFs = 1.07), indicating that multicollinearity was not an issue.

For experience, the main effects model revealed that loneliness (β = 2.79, SE = 1.23, t(137) = 2.26, p = .025) and prior exposure (β = 10.50, SE = 2.07, t(137) = 5.06, p < .001) were significant predictors of attributions of experience. The intercept was also significant (β = 22.73, SE = 2.32, t(137) = 9.79, p < .001). The interaction model again returned main effects for loneliness (β = 3.08, SE = 1.23, t(136) = 2.51, p = .013) and prior exposure (β = 10.40, SE = 2.05, t(136) = 5.08, p < .001) as well as a significant interaction (β = 2.51, SE = 1.16, t(136) = 2.16, p = .033). The intercept was also significant (β = 21.22, SE = 2.40, t(136) = 8.85, p < .001). This model reveals that the effect of loneliness on experience attributions changes depending on the level of prior exposure with greater attributions occurring as loneliness and prior exposure levels increase.

For agency, the main effects model revealed no effect of loneliness (β = -0.31, SE = 1.198, t(137) = -0.27, p = .791) but a main effect for prior exposure (β = 4.08, SE = 1.98, t(137)

= 2.06, p = .041). The intercept was also significant (β = 70.20, SE = 2.21, t(137) = 31.66, p < .001) indicating agency scores were significantly higher than zero when controlling for loneliness and prior exposure. For the interaction model, prior exposure was again significant (β = 4.04, SE = 1.98, t(136) = 2.04, p = .043), as was the intercept (β = 69.67, SE = 2.32, t(136) = 30.00, p < .001), but there was no interaction (β = 0.89, SE = 1.12, t(136) = 0.79, p = .432). In summary, the models showed that only prior exposure predicted agency attributions. The interaction models are visualized below in Figures 3.1 and 3.2.

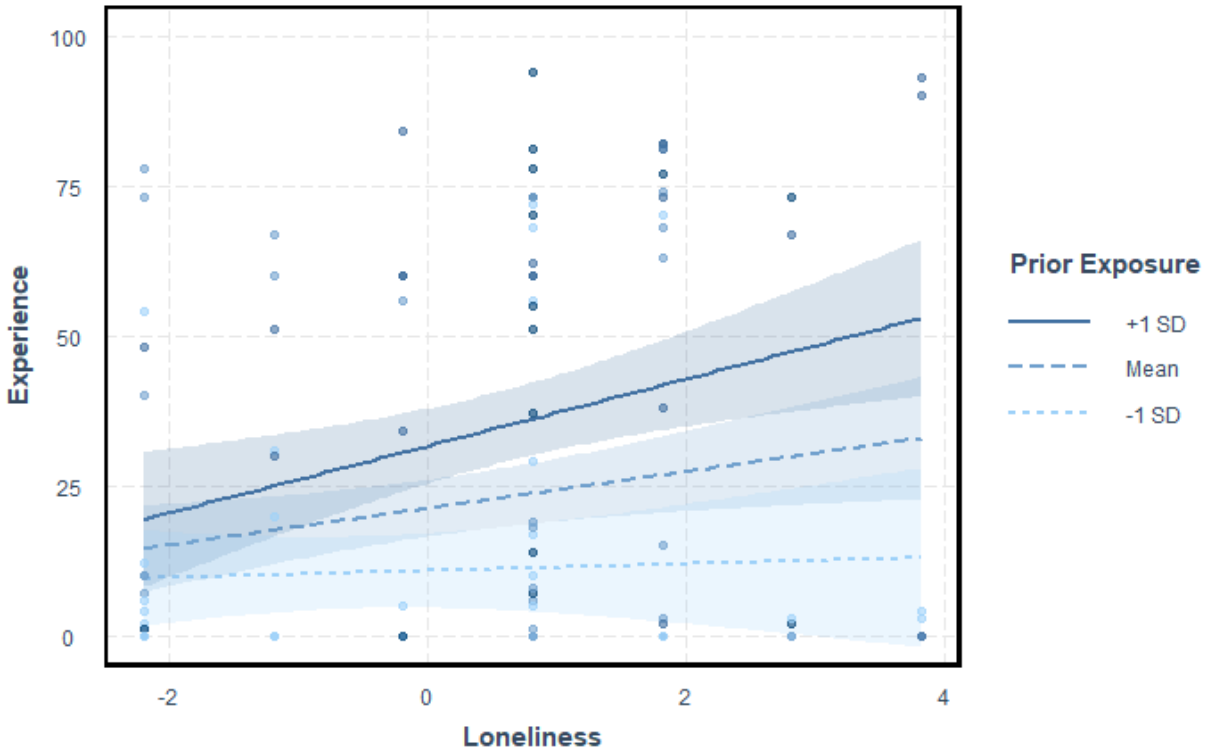

Figure 3.1: Experience values by loneliness and prior exposure. Loneliness and prior exposure levels are expressed in SDs.

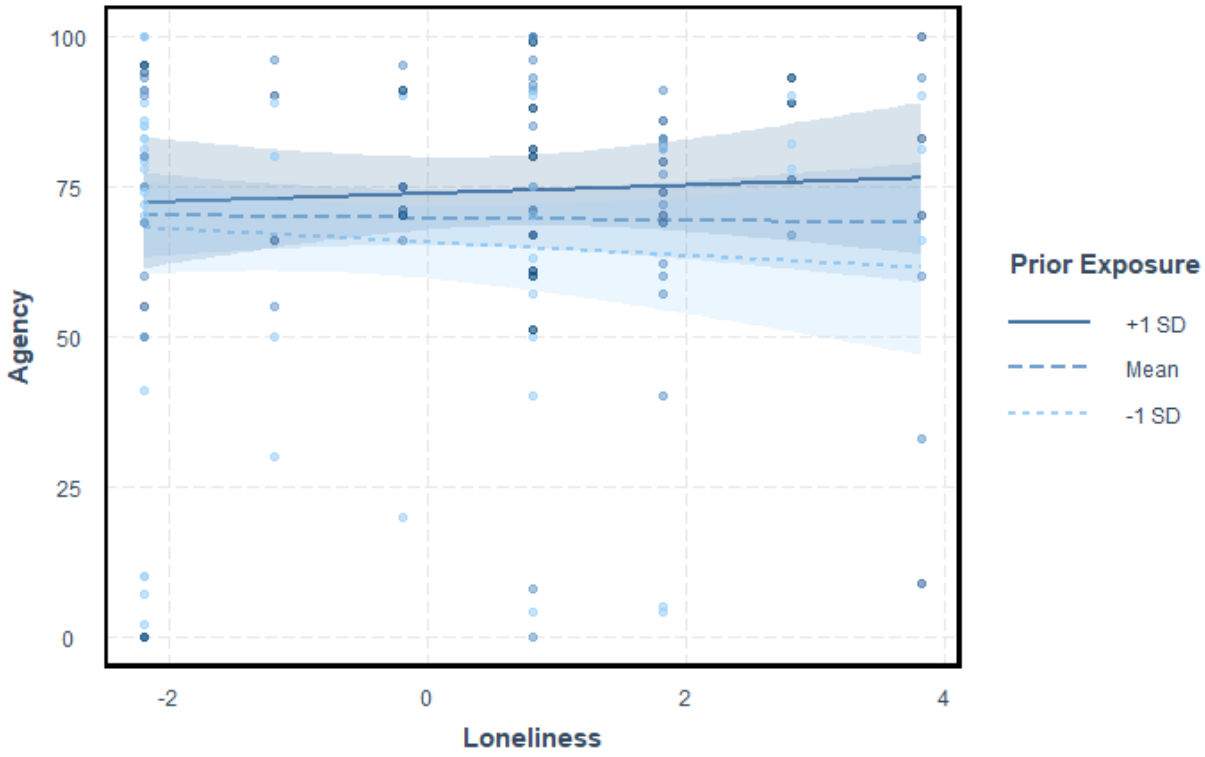

Figure 3.2: Agency values by loneliness and prior exposure. Loneliness and prior exposure levels are expressed in SDs.

Discussion

The growing recognition of loneliness as a serious health epidemic has coincided with a number of technological developments over the past two decades. The rise of Large Language Models (LLMs) has marked a significant increase in the ability of generative AI to roleplay as a social conversational agent—raising questions about whether social isolation may affect AI perception. While past work has explored the relationship between anthropomorphism and loneliness, little research has thus far examined the relationship specifically with LLMs (c.f., Folk et al., 2024). This is a particularly timely space to explore empirically as many members of the public self-report using LLM instances akin to a personal therapist.

The present study sought to examine if trait loneliness predicts anthropomorphizing LLMs. Specifically, the 2-factor model of mind perception (Gray et al., 2007) was used to probe anthropomorphism because of its parsimony in distilling a wide variety of mental features that can be attributed to others. Another major benefit is that the 2 factors, agency and experience, differentiate between mental features that relate more to cognitive, action-orientated capabilities versus emotional, feeling-orientated capabilities respectively. Moreover, given past research demonstrating that prior exposure influences anthropomorphizing LLMs (Jacobs et al., 2024), prior exposure was also examined as a moderating variable.

The first key finding was an association between trait loneliness and the general propensity to anthropomorphize—in alignment with past work suggesting such an association (e.g., Epley et al., 2008a; Eyssel & Reich, 2013; Shin & Kim, 2020; Bartz et al., 2016). Second, and more uniquely, trait loneliness was associated with mind perception attributions specifically toward LLMs. This was only true for experiential attributions but not for agentic attributions. Essentially, participants who reported greater feelings of loneliness were more likely to view

LLMs as human-like agents in terms of their ability to exhibit feelings. This finding fits nicely with hypotheses about why people anthropomorphize—loneliness is more related to a lack of perceived emotional connection and a desire for empathy or emotional understanding. Agentic capabilities are more mechanical, more cognitive, and more goal-orientated, which are less relevant to addressing emotional needs. Moreover, this finding fits with previous hypotheses regarding how differences in anthropomorphism may in part stem from individual differences in needs for emotional or social connection (Epley et al., 2007).

Notably, these latter findings contrast with a recent study that also included a measure of trait loneliness while examining anthropomorphism toward LLMs (Folk et al., 2023). Their study used, as is typically done, a more generalized anthropomorphic scale that did not differentiate between types of mental attributions involved in anthropomorphic tendencies, unlike the work here. The differing results between agency and experience attributions suggest that by collapsing across types of mental attributions, the relationship between loneliness and anthropomorphizing LLMs might be obscured.

Another observation was that prior exposure to LLMs predicted the degree to which LLMs are anthropomorphized both for agentic and experiential attributions. This also mapped onto our predictions following past findings (Jacobs et al., 2023). Critically, the relationship between loneliness and anthropomorphism again differed depending on the mind perception facet. For agency, there was no interaction between loneliness and prior exposure in predicting anthropomorphism whereas there was for experience. Essentially, increased loneliness predicted greater experiential ratings the more prior exposure participants had with LLMs. These results again support the idea that loneliness or a perceived lack of social connection leads individuals to seek connection or assign human-like qualities to nonhuman agents. And further, the type of

mind perceived in LLMs may be conditional based on the specific needs of the individual with experience being more closely linked to the need for social and emotional connections—which LLMs may symbolically fulfill. These findings also align with theories of compensatory control in that one response to lacking control in certain domains is to reestablish structure in the environment by perceiving patterns, even illusory ones, related to affirming the self (Kay et al., 2009; Whitson & Galinsky, 2008).

It is important to note several limitations of the present study. As a preliminary exploration of loneliness and anthropomorphic tendencies toward LLMs, the current data do not provide a comprehensive guide to investigating the nuances that might be involved in the patterns of results. Loneliness was not systemically manipulated and future directions could include a more controlled design using one of the various means of inducing feelings of loneliness in individuals (Zagic et al., 2024). New directions in the behavioural sciences include easing the means by which researchers can expose participants to custom-tailored instances of LLMs (Laban et al., 2024b). This could improve the rigor of future studies examining loneliness and anthropomorphic tendencies to LLMs by allowing researchers to manipulate the type of exposure to LLMs. These future investigations could also provide a means of exploring the causal mechanisms underlying the relationship between loneliness and anthropomorphizing in addition to highlighting any differences between types of LLMs. Future investigations could also explore how LLMs influence social isolation, as opposed to studying how social isolation influences AI perception.

In sum, the present findings offer an initial exploration into the relationship between loneliness and the attribution of human-like qualities to AI. While the current findings warrant caution in overinterpreting the results, they open up promising avenues for research by

demonstrating that the more emotional, socially relevant forms of mental attributions toward AI are connected to the degree to which individuals report being lonely. This line of inquiry also provides a method of probing the social and emotional roles AI may play in human life, particularly for individuals experiencing social isolation.

Chapter 5:

Chapters 3 and 4 revealed that people attribute mind toward social robots and LLMs. Moreover, the extent of these attributions is related to individual differences, such as the general propensity to anthropomorphize, the amount of prior exposure to LLMs, and traits including loneliness. In Chapter 5, I present two empirical studies (three experiments) applying the mind perception framework to evaluations of human minds. Study 4 first aims to validate using the mind perception framework for comparing self- and other-perception. Study 5 then examines if anthropomorphizing LLMs can affect how people view themselves and other human minds more generally. A version of Study 4 is published in *PLOS One* and a version of Study 5 is published in *Consciousness and Cognition.*

Study 4: Self-discrepancies in mind perception for actual, ideal, and ought selves and partners

Introduction

People perceive minds in other people, as well as other animals (e.g., cats and dogs), and even in nonbiological objects (e.g., robots). The field of work that is concerned with understanding these percepts is referred to as mind perception (Waytz et al., 2010). Mind perception has been applied to many different research areas within psychology and computer science (e.g., Wiese et al., 2017). However, surprisingly little, if any, work in mind perception has focused on investigating beliefs about one's own mind.

There has, however, been a considerable amount of research following Rogers' initial work linking psychotherapy outcomes with differences between domains of selves (Rogers,

1959). These included the real/actual self (the self that one actually is), the ideal self (the self that one desires to be), and the ought self (the self that one sees others as believing one should or ought to have). Roger's work was later greatly expanded upon by Higgins' landmark development of self-discrepancy theory which proposed that conflicting cognitive representations of the self result in emotional vulnerabilities and internal conflicts (Higgins et al., 1985). Indeed, differences between these three domains using a variety of scales (Hardin & Lakin, 2009; Higgins, 1987) have been linked with depressed affect (Phillips & Silvia, 2010), suicidal ideation (Cornette et al., 2009), and other specific affective states (Barnett et al., 2017). However, subsequent studies investigating these links have provided inconsistent results, in part due to how self-discrepancies have been operationalized (Ozgul et al., 2003; Tangney et al., 1998) as well as psychometric properties of its assessments (Watson et al., 2010; Watson et al., 2016).

Notably, Higgins used self-questionnaires to measure self-discrepancies between actual, ideal, and ought selves. The Selves questionnaire developed over time (Higgins et al., 1985; Higgins et al., 1986) but was constructed by asking participants to ascribe adjectives and attributes for each of the categories of selves. Far from a perfect measure, this method highlights the difficulty in measuring discrepancies in qualities of mind, a problem that has plagued investigations into the subject even as additional domains of selves have been investigated (Carver et al., 1999). A more recent study (Philips & Silvia, 2010) presented a newer, more robust method for measuring self-discrepancies based on integrating idiographic (individual) and nomothetic (group-based) tools. While it has been a valuable tool for validating predictions of self-discrepancy theory, it too uses the same method of allowing participants to self-generate attributes. The advantage of this method is that the attributes that are important to the self are

included. A fundamental disadvantage, however, is that direct comparisons between different attributes that vary between individuals are not possible.

In sum, the inconsistent operationalization of self-discrepancies and the questioning of its psychometric measures is at least in part a reflection of the difficulty in distilling qualities of mind into meaningful facets for interpretation. One potential remedy to this problem is to employ the mind perception framework developed by Gray et al. (2007) which has been applied successfully in many other contexts (Appel et al., 2020; Gervais, 2013). A major advantage of this approach to understanding how people perceive minds is that the majority of variance in people's perceptions of mind can be distilled into two distinct principal factors. Gray et al.'s framework (2007), derived from factor analysis, was the formation of a 2-factor model with one factor labeled experience (associated with capabilities related to feeling) and one factor labeled agency (associated with capabilities related to doing). In turn, the perception of these two factors, experience and agency, have been linked with meaningful predictors of real-world behaviour each in their unique way such as with personality (Tharp et al., 2017), psychopathology (Gray et al., 2011), and moral attitudes (Gray et al., 2012).

In the original study by Gray et al. (2007), one of the many targets in question were the participants themselves. In other words, participants were asked to rate themselves on a wide variety of capacities of mind. This in effect began to tap into metacognitive beliefs about one's own mind or one's actual self. Gray et al. found that people rated themselves higher on agency and experience than other animals, and they rated themselves higher again than a dead person, a robot, or inanimate objects. Beyond this initial foray, to our knowledge there has been no systematic study investigating perceptions and attitudes towards people's own minds using the mind perception framework. This is important because, as previously noted, the scope of

attributes encompassed by more traditional self-discrepancy measures of mind introduces variability based on individuals and specific contexts. This may mask meaningful relationships and contribute to the inconsistent pattern of findings in previous literature. The mind perception framework has the potential to aid the categorization of qualities of mind into a more precise and consistent measure for understanding self-discrepancies of mind.

It is worthwhile to highlight the subtle distinction between the self and the mind herein. The self typically refers to an individual's identity, personal characteristics, values, and attitudes (Baumeister & Finkel, 2010). It contains both the entirety of the 'hardware' and 'software' that makes us who we are. The mind, however, is conceptually thought of as the 'cognitive machinery' that enables us to come to think, feel, and understand the world around us which is encompassed by the self (Baars & Gage, 2010; Kim, 2018). In the context of self-discrepancy theory, there is a key distinction regarding self-discrepancies of the self, more broadly, versus the mind. The extant literature on self-discrepancies regarding physical features of the self as with body image is itself a large corpus of research relative to the focus on qualities of mind or mental features (Ahadzadeh et al., 2017; Altabe & Thompson, 1996; Bergstrom & Neighbors, 2006). The focus on qualities of mind in self-discrepancy theory is the branch particularly troubled by the inconsistent operationalization and psychometric issues (Ogzul et al., 2003; Watson et al., 2010; Watson et al., 2016) and is the branch in which the mind perception framework can potentially provide immediate value.

The purpose of the present research is to examine—using the mind perception framework—how people attribute values across different domains (actual, ideal, ought) and agents (self versus partner). We do so by first conducting an exploratory study with only actual versus ideal domains before following up with a pre-registered study involving all three domains

(self, ideal, ought). The inclusion of two agents reflects our intuition that people may be motivated to desire different attributes for themselves compared to a relationship partner, in line with some of Higgins' original findings that different emotional vulnerabilities relate to different kinds of self-discrepancies (Higgins et al., 1985). Finally, we examined sex differences to determine to what extent, if any, this factor influences discrepancy perceptions as a number of sex differences have been observed in other types of self-discrepancies, such as with body image (Mintz & Betz, 1986).

Methods

Participants

The number of participants was determined using an a priori power analysis using WebPower in R (Zhang & Mai, 2018). We sought to detect a medium-sized effect (d =.5) with a desired power of 90% for the lowest powered analysis which involved an interaction with sex. This led to a required sample size of 114 for each sex.

In total, there were 265 participants that took part in the survey (147 men and 118 women; M age: 40.71, SD age: 11.44), with 192 participants indicating that they were currently in a relationship. All participants took part using IP addresses from the United States and were sampled using Amazon Mechanical Turk (MTurk) and had approval rates of 99% on prior surveys. The distribution of participants reporting their highest educational attainment were as follows: 27.2% had a high school diploma, 10.6% had a post-secondary diploma, 44.9% had an undergraduate degree, and 17.3% had a postgraduate degree. This study was approved by the ethics board of the University of British Columbia (H10-00527) and all participants provided informed consent to participate.

Material and Procedure

A Qualtrics survey was distributed to participants using the crowd sourcing program MTurk, which is a common tool for online data collection used in behavioural sciences (Hauser & Schwarz, 2016). The data and survey are available online through https://osf.io/gwhc6 . After reviewing an ethics form and consenting to take part, participants were asked to fill in a basic demographic questionnaire with items pertaining to age, sex, and relationship status. Next, participants were given instructions that they would be asked to rate their current self, their ideal self , their current partner (if applicable), and their ideal partner on 6 questions on a 5-point numerical scale for each of the domains/agents (i.e., "For the following questions please answer them according to how you would assess your current self.", "For the following questions please answer them according to how you would assess your ideal self.", "For the following questions please rate your current partner as he or she currently is.", and "For the following questions please imagine an ideal partner. This person would be 'the person of your dreams'."). The 6 questions for each set of ratings were based on an updated version of the Mind Survey (Gray et al., 2007; Tharp et al., 2017). These were: 'How capable of feeling fear are you?', 'How capable of exercising self-control are you?', 'How capable of feeling self-pleasure are you?', 'How capable of remembering are you?', 'How capable of feeling hunger are you?', and 'How capable of acting morally are you?'. For the sections asking participants to rate their ideal self these questions were changed from present to future tense specifically by changing 'are you' to 'would you be'. Similar adjustments were made for the ratings of an ideal partner. The individuals who indicated they are currently in a relationship were presented with another block that included questions about their current partners, e.g., "How capable of remembering is your current partner?". The item for self-pleasure as part of the experience factor was removed for improved internal consistency. The resulting Cronbach's alphas for the self were: actual agency, $\alpha = .74$,

actual experience, α = .60 (formerly .59), ideal agency, α = .83, and ideal experience, α = .73 (formerly .53). The internal consistency for partners was as follows: actual agency, α = .74, actual experience, α = 64 (formerly .53), ideal agency, α = .76, and ideal experience, α = .72 (formerly .53). Participants were also asked an attention check question between the demographic and mind survey questions. After completing all questions, participants were thanked for participating and compensated for their time.

Results and Discussion

A linear mixed model was conducted to investigate the effect of agent (self and partner), domain (actual and ideal), and participant sex (male and female) as fixed factors on agency ratings (Figure 4.1) while participant was included as a random factor. The main effect of domain was significant, $\beta = 0.48$, $SE = 0.03$, $df = 723.76$, $t = 14.98$, $p < .001$; individuals preferred higher agency for ideal agents ($M = 4.42$, $SEM = 0.03$, 95% CI [4.35, 4.50]) than actual ones ($M = 3.95$, $SEM = 0.03$, 95% CI [3.87, 4.02]. No other main effects nor any interactions were significant (all p's > .05).

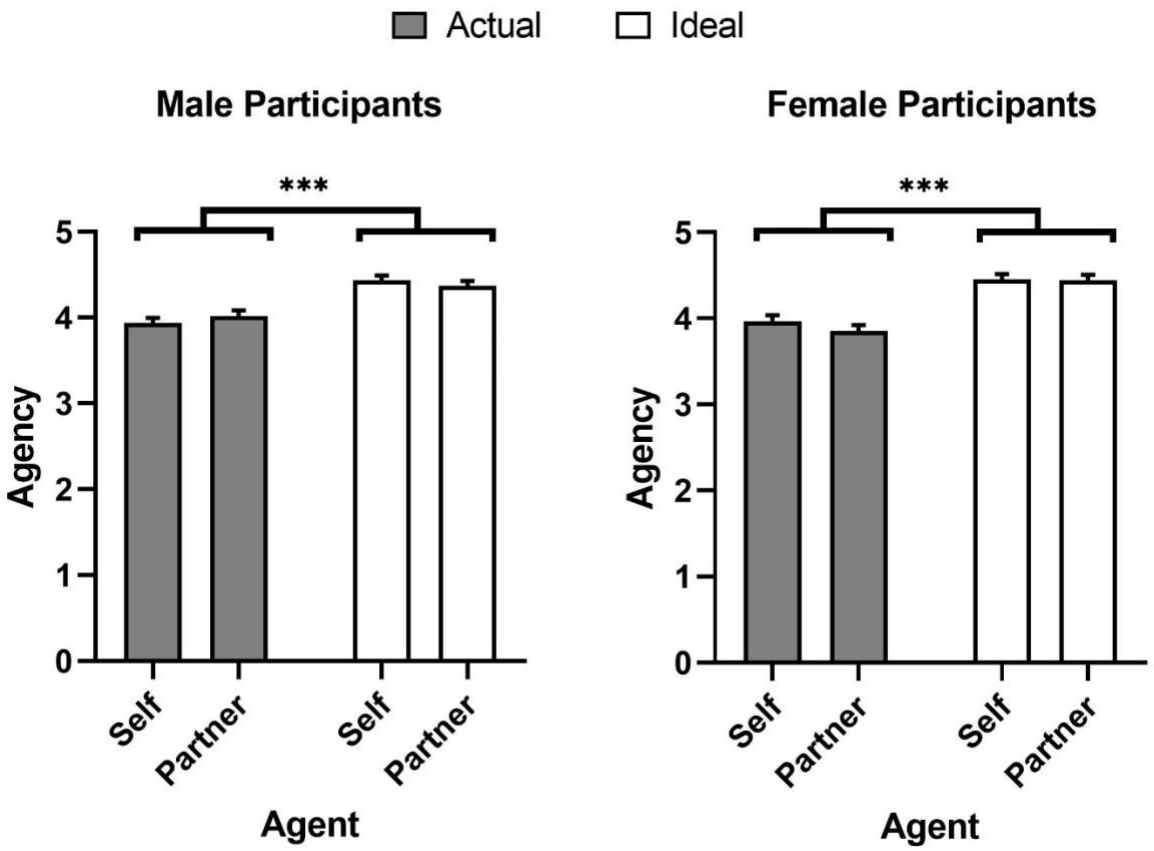

Figure 4.1: Mean and SEM for agency ratings as a function of domain. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Another linear mixed model was conducted predicting experience ratings with agent (self and partner), domain (actual and ideal) and participant sex (male and female) as fixed factors, in addition to participant being included as a random factor (Fig 4.2). The main effect of domain was significant, β = -0.45, SE = 0.04, df = 722.47, t = -11.89, p < .001. Moreover, significant two-way sex × agent and domain × agent interactions were qualified by a significant three-way agent × domain × sex interaction, β = 0.36, SE = 0.15, df = 722.47, t = 2.35, p = .019. Pairwise comparisons with Bonferroni corrections showed that men preferred less experience for their

ideal self (*M* = 3.59, *SEM* = 0.08, 95% CI [3.44, 3.74], $p < .001$) and ideal partners (*M* = 3.67, *SEM* = 0.08, 95% CI [3.52, 3.82] , $p < .001$) than their actual partners (*M* = 4.10, *SEM* = 0.09, 95% CI [3.93, 4.27]). Men also preferred less experience for their ideal self than their actual self (*M* = 4.02, *SEM* = 0.08, 95% CI [3.87, 4.17], $p < .001$). Similarly, women preferred less experience for their ideal self (*M* = 3.44, *SEM* = 0.09, 95% CI [3.27, 3.61]) than their actual self (*M* = 4.09, *SEM* = 0.09, 95% CI [3.92, 4.26], $p < .001$). Moreover, women preferred less experience for their ideal partner (*M* = 3.54, *SEM* = 0.09, 95% CI [3.37, 3.71]) than their actual partner (*M* = 3.84, *SEM* = 0.09, 95% CI [3.66, 4.02], $p = .010$) The main effects of agent and sex, and sex × domain interaction were not significant (all *p*'s > .228).

The results from Study 1 indicate that participants strongly differentiated between agency and experience. Men and women desired greater agency but reduced experience across agents being assessed (self or partner). This split across facets of mind perception reflect similar patterns found in other applications of mind perception that suggest each facet predicts unique clusters of real-world behaviour (Gray et al., 2007; Gray et al., 2011). Finally, men, unlike women, indicated that they perceived more experience in their current partners than ideal selves. See Table 4.1 for the means and standard deviations of each mind perception item.

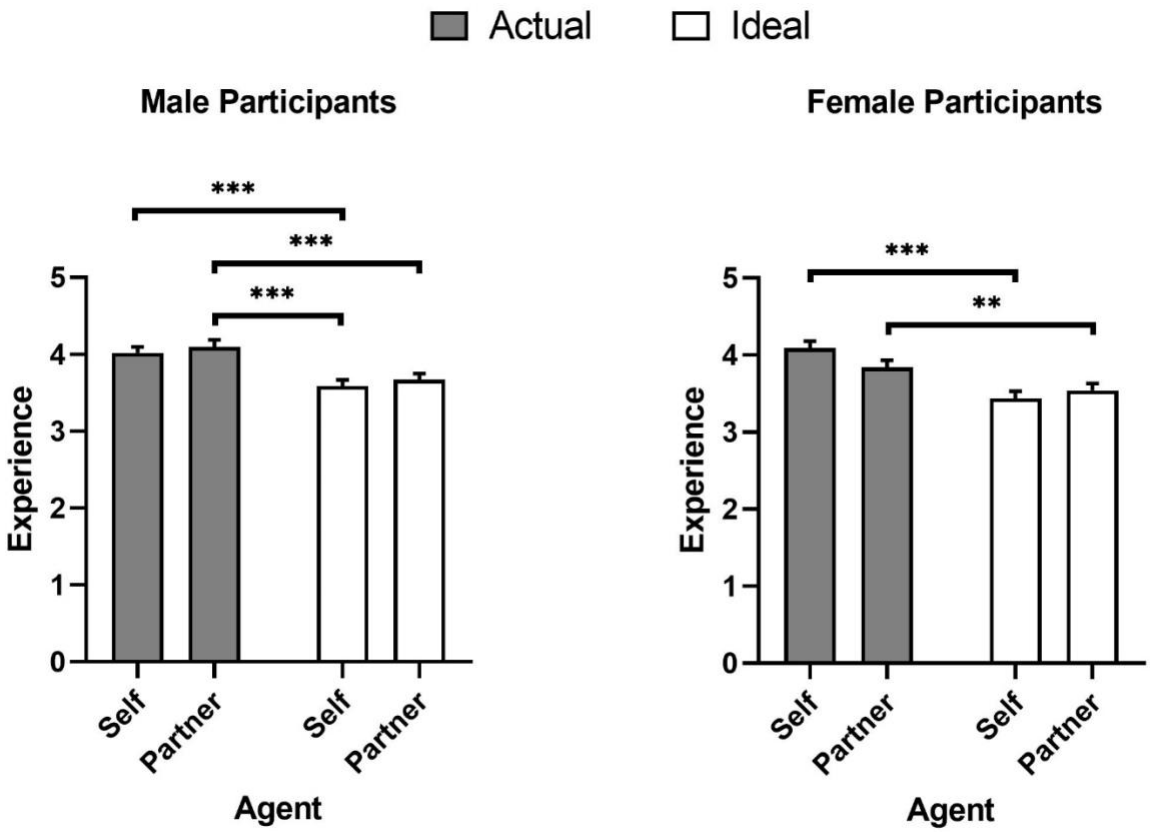

Figure 4.2: Mean and SEM for experience ratings for male and female participants as a function of agent, domain, and participant sex. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 4.1. Means and standard deviations for mind perception factors.

			Experience		Agency		
			Fear	Hunger	Self-control	Remembering	Acting Morally
Self	Mean	Ideal	3.36	3.68	4.35	4.42	4.53
	SD		1.19	1.18	0.89	0.83	0.75
	Mean	Actual	3.91	4.19	3.76	3.83	4.28
	SD		1.05	0.97	1.01	0.96	0.76
Partner	Mean	Ideal	3.44	3.79	4.28	4.32	4.61
	SD		1.12	1.06	0.83	0.74	0.65
	Mean	Actual	3.81	4.22	3.77	3.83	4.29
	SD		1.02	0.89	1.01	0.97	0.78

Experiment 2

Following the results of Experiment 1, an additional study became desirable to examine the robustness of the prior results through replication with a new sample. The secondary aim of Experiment 2 is to also include an additional domain, the ought self, for assessment. This study was pre-registered and is available for viewing on OSF: https://osf.io/gwhc6.

The hypotheses for Experiment 2 follow from the results of Experiment 1. We expect that individuals will desire greater agency and less experience in their ideal selves and ideal partners compared to their actual selves and actual partners. The new hypothesis concerns the addition of the ought self. Specifically, we expect that individuals will express that the ought self and ought partner should possess greater agency and less experience than the actual self and partner. This hypothesis stems from the idea that the ought self is heavily shaped by societal norms and expectations similar to perceptions of the ideal self. In this case, we believe that norms that value competence and leadership will lead to believing others think one ought to have greater agency,

and norms about avoiding vulnerability will lead to believing others think one ought to have less experience.

Methods

Participants

205 participants took part in the survey (94 men and 111 women; M age: 41.63, SD age: 10.97), with all participants indicating that they were currently married or in a relationship (a criterion for eligibility to take part). All participants took part using IP addresses from the United States and were sampled using Amazon Mechanical Turk (MTurk) and had approval rates of 99% on prior surveys. The distribution of participants reporting their highest educational attainment were as follows: 21.0% had a high school diploma, 8.8% had a post-secondary diploma, 44.9% had an undergraduate degree, one indicated elementary school and 24.8% had a postgraduate degree.

Material and Procedure

The materials and procedure closely resembled Experiment 1. After consenting to take part, participants were asked to indicate their age, sex, and education. Qualtrics was used for the dissemination of the survey through CloudResearch. After the demographic questions, participants were presented at random each of the domains (actual, ideal, ought) and agents being assessed (self vs. partner). The instructions were the same as in Experiment 1 (e.g., “For the following questions please answer them according to how you would assess your current self.”) with the exception being the novel instructions for the ought self (“For the following questions, please answer them according to how you would assess what others think you ought to be (should be like)” and ought partner (“For the following questions, consider the attributes that external parties—such as family, friends, or society at large—deem essential or desirable in your

partner."). The same 6 questions from the Mind survey were asked for each of the domains and agents. Again, these were: 'How capable of feeling fear are you?', 'How capable of exercising self-control are you?', 'How capable of feeling self-pleasure are you?', 'How capable of remembering are you?', 'How capable of feeling hunger are you?', and 'How capable of acting morally are you?'. These questions were again modified to grammatically fit the accompanying domain/agent being assessed. The item for self-pleasure as part of the experience factor was removed for improved internal consistency once again. The resulting Cronbach's alphas for the self were: actual agency, α = .65, actual experience, α = .55 (formerly .59), ought agency, α = .76, ought experience, α = .64 (formerly .60), ideal agency, α = .77, and ideal experience, α = .67 (formerly .53). The internal consistency for partners was as follows: actual agency, α = .71, actual experience, α = .54 (formerly .56), ought agency, α = .72, ought experience, α = .65 (formerly .64), ideal agency, α = .80, and ideal experience, α = .61 (formerly .53). Participants were also asked an attention check question among the mind ratings. After completing all questions, participants were thanked for participating and compensated for their time.

Results and Discussion

A linear mixed-effects model was used to examine the influence of domain (actual, ideal, ought), agent (self vs partner), and participant sex on agency ratings with participant as a random effect. The fixed effects omnibus test revealed a main effect of domain, $F(2,1015) = 131.98$, $p < .001$, a main effect of agent, $F(1,1015) = 4.05$, $p = .044$, and an interaction between participant sex and domain, $F(2,1015) = 4.62$, $p = .010$. The main effect of sex, the interaction between sex and agent, and the 3-way interaction were all nonsignificant; all p's > 0.05. The fixed effects parameter estimate for agent revealed that across both sexes, people attributed greater agency to themselves ($M = 4.29$, $SEM = 0.03$, 95% CI [4.17, 4.31]) versus their partner, ($M = 4.24$, $SEM =$

0.03, 95% CI [4.22, 4.36], $t(1015) = 2.011$, $p = 0.044$). Moreover, post-hoc Bonferroni comparisons between domains revealed that participants rated the ideal domain ($M = 4.46$, $SEM = 0.03$, 95% CI [4.38, 4.53]) higher than the ought ($M = 4.38$, $SEM = 0.03$, 95% CI [4.30, 4.45], $p = 0.046$) and actual domains ($M = 3.96$, $SEM = 0.03$, 95% CI [3.88, 4.03], $p < .001$). They also rated the ought domain significantly higher than the actual domain ($p < .001$).

The post-hoc comparisons on the sex by domain interaction revealed that among women, participants desired greater agency for their ideal scores ($M = 4.44$, $SEM = 0.05$, 95% CI [4.34, 4.54]) compared to their actual scores ($M = 3.83$, $SEM = 0.05$, 95% CI [3.74, 3.95], $p < .001$), and greater agency for their ought scores ($M = 4.34$, $SEM = 0.05$, 95% CI [4.24, 4.44]) than actual scores, ($p < .001$). Among men, the same pattern occurred wherein actual scores ($M = 4.07$, $SEM = 0.06$, 95% CI [3.96, 4.18], were significantly lower than ideal ($M = 4.47$, $SEM = 0.06$, 95% CI [4.36, 4.58], $p < .001$) and ought scores ($M = 4.41$, $SEM = 0.06$, 95% CI [4.30, 4.52] , $p < .001$). Between sexes, female actual scores were lower than male ideal ($p < .001$) and ought scores ($p < .001$). Male actual scores were likewise lower than female ideal ($p < .001$) and ought scores ($p < .001$). Finally, female actual scores were lower than male actual scores ($p < .001$).

These results replicate the main finding from Experiment 1, specifically, that across sexes, people desire greater agency in their ideal selves and partners compared to their actual selves and actual partners. Similarly, it showed that people desired greater agency for their ought selves and ought partners. The marginal differences between Experiment 1 and 2 are that both sexes rated their actual selves as having greater agency than their partners and that men rated their actual selves higher than women rated their actual selves.

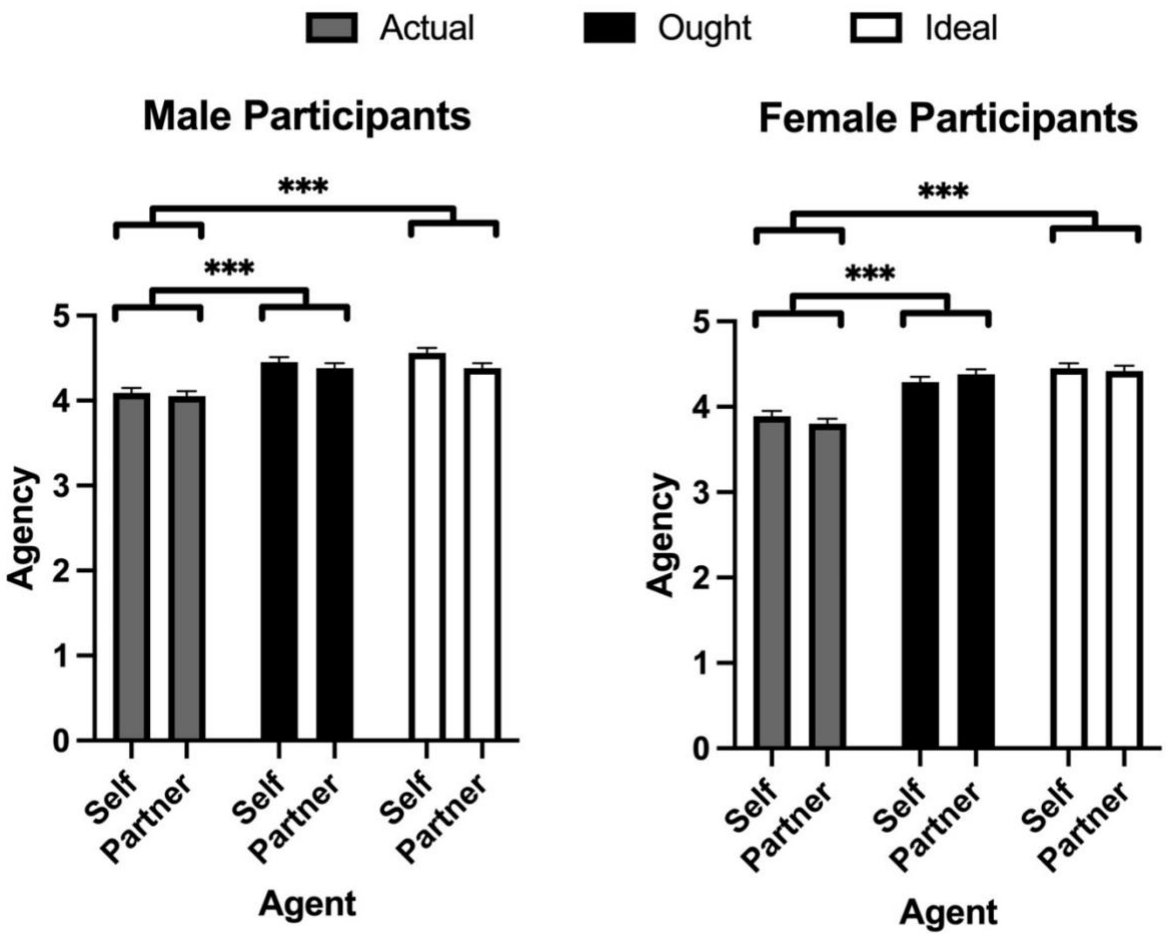

Figure 4.3: Mean and SEM for agency ratings as a function of domain. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

For experience ratings, another linear mixed-effects model was used with domain (actual, ideal, ought), agent (self versus partner), and participant sex as fixed effects and participant as a random effect. The fixed effects omnibus test revealed a main effect of domain, $F(2,1015) = 94.62$, $p < .001$, and an interaction between domain and agent, $F(2,1015) = 11.62$, $p < .001$. Other main effects and interactions were all nonsignificant, p's $> .05$.

Post-hoc comparisons of experience ratings between domains revealed that participants desired less experience in their ideal scores ($M = 3.43$, $SEM = 0.05$, 95% CI [3.32, 3.53]) than

actual scores (M = 3.96, SEM = 0.05, 95% CI [3.58, 4.06], $p < .001$), ought scores (M = 3.53, SEM = 0.05, 95% CI [3.42, 3.63]) over actual scores ($p < .001$), and ideal scores over ought scores ($p = .038$).

Post-hoc comparisons on the interaction between domain and agent were also examined. Participants rated their actual partner (M = 3.88, SEM = 0.06, 95% CI [3.76, 4.00]) as having more experience than their ideal partner (M = 3.55, SEM = 0.06, 95% CI [3.43, 3.67] , $p < .001$), ideal self (M = 3.31, SEM = 0.06, 95% CI [3.19, 3.42], $p <.001$), ought partner (M = 3.56, SEM = 0.06, 95% CI [3.44, 3.68], $p <.001$), and ought self (M = 3.50, SEM = 0.06, 95% CI [3.38, 3.62], $p <.001$). Participants also rated their actual self (M = 4.04, SEM = 0.06, 95% CI [3.92, 4.16]) as having more experience than their ideal self ($p < .001$), ought self ($p < .001$), ideal partner ($p < .001$), and ought partner ($p < .001$). The remaining significant comparisons were that participants attributed more experience to their ideal partner than ideal self ($p < .001$), more experience to their ought self compared to ideal self ($p = .015$), and more experience to their ought partner than ideal self ($p < .001$).

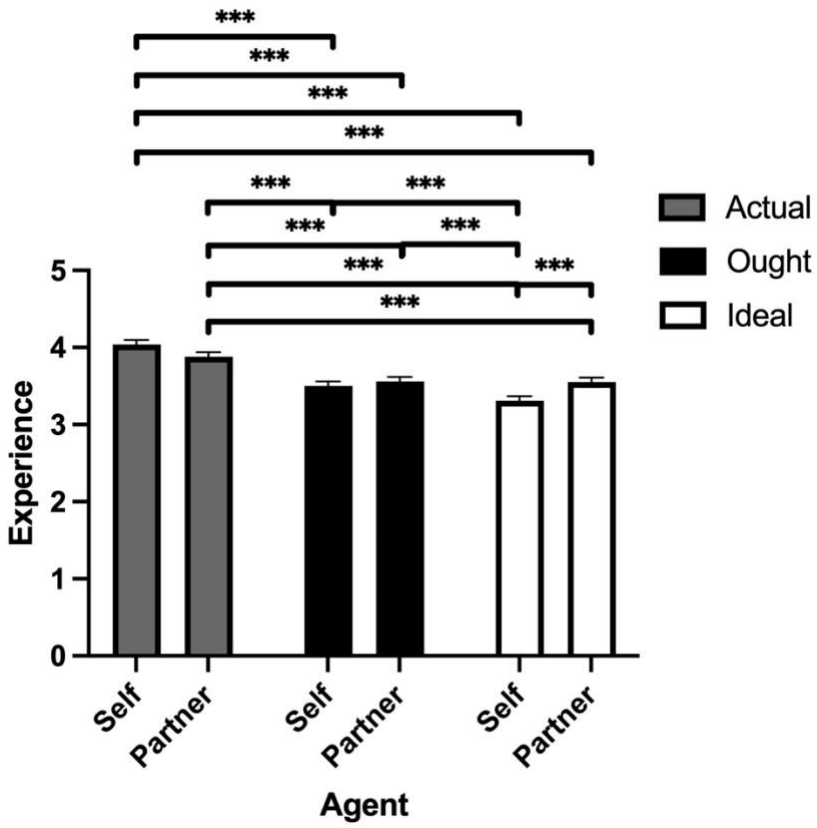

Figure 4.4: Mean and SEM for experience ratings for male and female participants as a function of agent, domain, and participant sex. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

In summary, the main results for experience attributions were very similar to Experiment 1. Participants attributed greater experience to their actual selves and partners than their ideal selves and ideal partners. Participants also attributed greater experience to their actual selves and partners than their ought selves and partners. These gaps between domains appeared larger for self-ratings than partner ratings. This latter finding could reflect that the desire for reduced vulnerability in oneself and partner is stronger for the self both ideally and for what people believe others think they ought to be like.

Table 4.2: Means and standard deviations for mind perception domains.

			Experience		Agency		
			Fear	Hunger	Self-control	Remembering	Acting Morally
Self	Mean	Ideal	3.08	3.52	4.40	4.54	4.58
	SD		1.19	1.17	0.84	0.72	0.72
	Mean	Ought	3.33	3.65	4.30	4.29	4.50
	SD		1.10	1.08	0.84	0.75	0.71
	Mean	Actual	3.90	4.18	3.81	3.96	4.19
	SD		0.93	0.81	0.94	0.89	0.82
Partner	Mean	Ideal	3.37	3.72	4.35	4.37	4.49
	SD		1.01	0.99	0.78	0.72	0.72
	Mean	Ought	3.37	3.73	4.29	4.33	4.53
	SD		1.04	1.02	0.78	0.71	0.67
	Mean	Actual	3.63	4.12	3.79	3.74	4.21
	SD		1.03	0.84	0.96	0.93	0.76

General Discussion

Understanding the psychological consequences of self-discrepancies in domains of the self (i.e., actual, ideal, ought) has been a long, meritorious scientific endeavor. However, discrepancies regarding physical features (i.e., body image) are far better understood than discrepancies in mental qualities. In other words, while the field knows much about the causes and consequences of self-discrepancies between one's actual, ideal, and ought body image (Ahadzadeh et al., 2017; Altabe & Thompson, 1996; Bergstrom & Neighbors, 2006; Mintz & Betz, 1986), research into discrepancies in attributes of mind remain a much less investigated and a far more convoluted issue (Watson et al., 2010, 2016). As noted in the introduction, we take as a working hypothesis that a key reason for the latter situation stems from the complexity in past measurements of mind that could introduce variability and restrict comparisons between individuals. Thus, we took a novel approach to this issue by using a 2-factor model of mind perception (Gray et al., 2007) to probe discrepancies in mind. A critical advantage of this approach is that it provides a succinct and common metric to both measure and interpret a variety of qualities of mind between individuals.

In Experiment 1, 265 participants took part in a CloudResearch distributed mind survey probing participants on their attitudes towards their current mind and their preferences for an ideal mind for themselves and their partners. The results clearly demonstrated that both men and women desire greater agency for themselves and their partners. In contrast, men and women want less experience for themselves and their partners while men, unlike women, also prefer less experience than their current partner.

In Experiment 2, a new sample of 205 participants were recruited using the same methodology as Experiment 1 but with the inclusion of an additional domain of the self—the

ought self. The large discrepancies between actual and ideal ratings for one's self and partner emerged again, and ratings for the ought self and ought partner displayed a similar pattern wherein people thought they ought to have more agency but less experience. In Experiment 2, participants also indicated they viewed themselves as having more agency than their partner regardless of their sex, and that men compared to women rated themselves higher on agency. Participants also stated they preferred less experience than their ideal partners, ought partners, and ought selves.

Across both studies the findings suggest several key ideas. First, mind perception frameworks can be used to elucidate discrepancies between domains of the selves similar to other methods that have been used to examine discrepancies in the past (Barnett et al., 2017; Hardin & Lakin, 2009). Additionally, and more importantly, the 2-factor model of mind perception reveals a distinction between the factors of agency and experience in terms of how people conceptualize their actual versus ideal or ought attributes of mind. People desire more agency in their ideal and ought self along with their partner. It is perhaps intuitive to discover that people desire greater self-agency and that they ought to have more of it. After all, people often lament their inability to do more or to do better. More surprising, is that our results on ratings of experience indicate that people wish and believe they ought to feel less. It would be unusual to hear remarks about wanting to feel less hope or sexual drive, or other forms of feeling—except perhaps for feelings of pain and depression. This finding that people idealize less experience is especially interesting given the fact that between agency and experience, Gray et al. (2007) found that experience to be the more important factor in terms of its ability to capture what makes us uniquely human.

The difference between agent (self versus partner) attributions in Experiment 2 also yielded an interesting pattern of results. The difference between one's actual experience and ideal or ought experience is larger than those perceived in one's partner. In part, this is due to the lower actual perceived experience in one's self in addition to the higher levels of ideal or ought experience in one's partner. If perceived experience is viewed as a form of moral patiency, as has been previously described (Gray et al., 2007), these findings suggest that people view themselves as somewhat more vulnerable than their partners, while ideally, they ought to be less. This fits a narrative that people believe they ought to be supportive and protective of their partners but want to avoid being the vulnerable partner themself.

The attitudes toward experience ratings and their association with moral patiency also may explain a more intriguing finding from Experiment 2. People believed they ought to have less experience than they do in reality, and they ideally would have even less than what they believe they ought to have. In essence people ideally want to avoid being the target of moral right or wrong more than they believe others think it is acceptable to be. Perhaps this is reflecting an avoidance of victim blaming while considering a strong desire to avoid becoming a victim oneself.

In addition to the main findings, there were some intriguing sex differences between men and women. In Study 1, men preferred less sensitivity (i.e., experience) in their idealized self compared to their actual partners, whereas women did not. This finding did not re-emerge in Study 2. Similarly, a key sex difference in Experiment 2—that men rated their actual agency higher relative to women—was not found in Experiment 1. It could be that popular conceptualizations of masculinity being associated with less overt displays of emotion (Walter et al., 2020) may play a part in the sex effects for experience and agency. However, the collective

evidence between experiments does not provide strong evidence that large sex differences exist in self-discrepancies between domains and agents—at least when examined within the mind perception framework—and thus the sex differences we observed are best treated with caution.

Future Directions and Limitations

As an initial pair of studies delving into self-discrepancies between actual, ideal, and ought minds using a mind perception framework, these findings raise many questions for future research. For example, researchers may ask how people consider the minds of family members or friends, compared to strangers or foes? Or in a broader sense, how might individuals attribute mind to ingroup versus outgroup members? Mind perception has already been linked and discussed within the context of moral judgements (Gray et al., 2012; Will et al., 2021). Dehumanization, which involves the denial of universally human attributes (Haslam, 2006), could map on to larger discrepancies between actual and ideal minds in particular outgroups. In other words, the mind perception approach that was applied here could be used to investigate whether dehumanizing percepts in part stem from larger discrepancies between actual and ideal minds reflected in the moral judgements of others. Furthermore, the denial of quintessentially human attributes suggests that self-discrepancies related to dehumanization would be particularly large on the experience factor of mind perception.

It is also worth noting that the present study had limitations and raises questions warranting further inquiry. The internal consistency of the mind perception survey had acceptable reliability only after the removal of the self-pleasure item for the experience factor across both studies, and in general, suggests that modifications to the survey may be required to probe self-discrepancies in mind for other purposes. More notably, the divergence of ideal qualities for agency and experience highlighted that there may be differences in the social

desirability and valence of items in the mind perception survey. There appears to be more positive associations with agency items than experience items; a point surprisingly absent in the mind perception literature. One potential remedy would be to create new questions matched for valence (e.g., switching fear to courage or hunger to satiety) in future work applying the mind perception framework to self-discrepancies. Recent work has also revisited the optimal number of dimensions of mind perception with some advocating for a unidimensional factor structure (S. Lee et al., 2020; Looser & Wheatley, 2010; Tzelios et al., 2022), some advocating 3 factors (Tamir & Thornton, 2018; Weisman et al., 2017) and some even advocating for 5 factors (Malle et al., 2019). Interestingly, Malle et al. (2019) found that a 5-factor approach worked best when examining mind perception in desired robots—perhaps indicating that larger factor solutions may be optimal for examining idealized qualities of mind which future work could extend to humans in a method similar to the present approach.

In conclusion, the present work suggests that self-discrepancies between actual, ideal, and ought attributes of mind can be succinctly distilled using the mind perception framework. Using this framework, it was discovered that people in general idealize and believe they ought to have greater agency and weaker experiential qualities both for themselves and their partners. These findings suggest combining mind perception and self-discrepancy theory can provide new ways of investigating psychological well-being and prosocial behaviours such as moral judgments.

Study 5: Large Language models have divergent effects on self-perceptions of mind and the attributes considered uniquely human

Introduction

The meteoric rise of highly capable AI chatbots such as ChatGPT and other large language model (LLM) applications has spurred calls to use psychological tools to investigate how interactions with these systems influence a number of perceptions (Binz & Schulz, 2023; Kosinski, 2024; Shiffrin & Mitchell, 2023). A variety of tools in psychology, cognitive science, and other related fields have emerged to study perceptions in the context of diverse human-computer interactions, which can be applied to LLMs. While this interest harkens back to an earlier period with chatbots such as ELIZA (Shum et al., 2018; Weizenbaum, 1966), the immense technological developments and widespread popularity of LLMs have reinvigorated the need to understand attitudes toward AI systems.

The novel capabilities of AI offer a particular opportunity to investigate the tendencies of people to anthropomorphize and its subsequent consequences. Anthropomorphism involves recognizing a spectrum of human-like traits, from seeing human-like shapes in the environment (commonly referred to as animacy; Bartneck et al., 2009), to attributing human-like qualities, both non-mental and mental, to nonhuman entities (Epley, Akalis, et al., 2008; Epley et al., 2007; Waytz, Gray, et al., 2010).

The attribution of mental states has been distilled into two principal factors by Gray, Gray, and Wegner (2007): agency—the ability to do, think, and act morally; and experience—the ability to feel emotions, drives, and pleasure or pain. Their mind perception framework has been applied extensively to a wide variety of beings including humans, animals, robots, and AI systems (e.g., Jacobs et al., 2022; Shank et al., 2019; Wiese et al., 2017). Recently, this mind

perception framework has also been applied to ChatGPT, revealing that even the briefest of exposures to ChatGPT can increase agentic (the ability to do) and experiential (the ability to feel) attributions to ChatGPT (Jacobs et al., 2023a). In other words, exposure can increase anthropomorphic attitudes toward LLMs—including those relating to the perception of mind.

In the initial work by Gray et al. (2007) one of the targets of mind perception was the self, such that individuals were asked to rate themselves on a variety of capacities of mind. People predictably rated themselves higher than nonhumans such as a dog, frog, or robot on both agency and experience. However, in their study, as in the mind perception literature in general, the measures of mind have tended to focus predominately on external mental attributions toward *others* rather than internal attributions toward the *self* (i.e., self-perception).

Though most investigations in the field have concentrated on one's perceptions of other minds (c.f., Jacobs et al., 2023b), it is essential to acknowledge that our understanding of our own mind does not develop in isolation—it emerges within a comparative context. We define and perceive ourselves often in comparison to others (e.g., social comparison theory; Festinger, 1954; Gerber et al., 2018). The degree to which we do this with nonhumans is far less studied or conclusive, with past investigations often alluding to, but not directly addressing, self-comparisons with nonhumans (Brette, 2022; Kiesler et al., 2008; Waytz, Gray, et al., 2010). However, if we take as a working hypothesis that the external attributions of mind to nonhuman entities (i.e., anthropomorphism) can in turn affect our self-attributions—where humans not only attribute human-like qualities to nonhuman entities but also re-evaluate their own characteristics as a result—then there is a large gap in our understanding of anthropomorphic tendencies and self-perception. As interactions with increasingly sophisticated AI systems like ChatGPT become more prevalent, investigating this often-overlooked consequence of anthropomorphism becomes

all the more important. Moreover, this interplay between self-perception and other-perception raises intriguing questions. For instance, how might our growing familiarity with LLMs impact our self-perception of mind (i.e., how we view our own minds) and what makes us human?

Approaches to this question have complex philosophical and religious roots that begin with Aristotle's numerous writings on the essence of what it is to be human with his emphasis on rationality (Barker & Stalley, 1995). Perhaps not surprisingly, since that time a range of different viewpoints have emerged. For example, Charles Darwin famously noted that the difference in mind between man and higher animals, although great, is more a matter of degree than kind (Darwin, 1859). In contrast, other scientists such as Michael Gazzainga (2008) have leaned towards the metaphor of a phase shift in evolution, with the result being that it is nearly impossible to think of the minds of humans and animals as having the same constituent parts (Cacioppo & Patrick, 2008). Psychology and neuroscience also have a long history of searching for what makes the human brain unique, and by extension the human mind. A number of hypotheses have been put forth ranging from the high ratio of brain size to body size (Cairó, 2011), to the unusually large development of the cerebral cortex (Molnár et al., 2019), to the plasticity of human brains compared to our closest ancestors (Gómez-Robles et al., 2015).

Critically, the conceptualization of what makes the human mind unique has always been thought of in a comparative context first with animals and now, perhaps increasingly, with machines. Furthermore, a common thread between the disparate ideas that exist is that it is the cognitive abilities of the human mind that separate it from all else. It is our ability to think rationally, perform mathematical calculations, and create works of art that make us uniquely human. However, as AI encroaches on each of these capabilities, it is understandable that people begin to question if these agentic abilities 'to do' are what establishes the human mind as unique.

Rather, people may begin to emphasize our experiential capabilities 'to feel' as crucial to what makes them in particular, and humans in general, unique. In short, it is not our ability to do that makes us special, but the way we feel and experience our world that makes us human. A key distinction is that when attempting to address these questions about self-perception and the degree to which certain qualities of mind are uniquely human, is that there are two approaches one can take. There is the loftier philosophical approach, and there is the empirical approach that seeks to survey folk psychology. The benefit of the mind perception framework is that it can be used for the latter to probe people's attitudes on these important questions (Jacobs et al., 2022).

The present study examines if exposure to LLM responses can influence self-perception of mind and the attributes of mind considered to be uniquely human. We predict that people will place greater emphasis on the experiential components of mind because as AI systems continue to become more sophisticated in their range of capabilities to do various tasks, the experiential capabilities of humans become comparatively larger than any agentic differences. To test this hypothesis, participants were recruited to fill out mind perception scales *rating themselves* and indicating the degree to which agency and experience are uniquely human, before and after brief exposure to ChatGPT prompts.

Methods

Participants

G*Power was used to conduct a power analysis to detect differences among groups for small effects (*d*=.3) with 95% power. This suggested a sample size of at least 134 participants. To protect from participant dropout and exclusion, 150 participants took part via Amazon's Mechanical Turk (MTurk). 17 participants were excluded for rapidly clicking through the survey leaving 133 for the analysis. Of these participants, 75 were men, 57 were women, and 1 individual preferred not to identify. The mean age was 40.7 (SD=12.4) and all participants took part from IP addresses listed in the United States. This study was approved by the Behavioural Research Ethics Board of the University of British Columbia (H22-00572).

Procedure

After agreeing to participate following a brief description of the study that did not mention ChatGPT or LLMs, participants were asked to answer demographic questions related to their age and gender. Next, participants were asked to rate themselves on a modified 5-item Mind Survey (Gray et al., 2007; Jacobs et al., 2023b; Tharp et al., 2017) which had a 5-point numerical scale from 'Not at all' (1) to 'A great deal' (5). These items were: 'How capable of exercising self-control are you?', 'How capable of remembering are you?', 'How capable of feeling fear are you?', 'How capable of feeling self-pleasure are you?', and 'How capable of feeling hunger are you?'. A morality question was unintentionally omitted from the survey. The first 2 questions measured agency with acceptable reliability (α =.74), and the latter 3 questions measured experience with acceptable reliability (α =.71). Participants were then asked to rate themselves on agency (ability to do things) and experience (ability to feel things). These items were measured using a slider scale, with 0 indicating 'Not at all', and 100 indicating 'A great deal'. These questions were included in addition to the Mind survey in order to determine if the Mind

survey questions would differ from simply asking for agency and experience ratings as is occasionally done (Will et al., 2021; Jacobs et al., 2022). Next, participants were asked to answer the question: 'To what degree do you think that agency (the ability to do things) makes humans unique compared to AI?' and "To what degree do you think that experience (the ability to feel things) makes humans unique compared to AI?" using sliding scales, from 0 to 100. After answering the previous questions, participants were presented with three real prompts provided by ChatGPT (GPT-4) along with its responses. The prompts were related to (a) the nature-nurture debate, (b) a tricky question about the number of feet that fit in a shoe, and (c) generating a cover letter for a janitorial position on the moon. These prompts were chosen to highlight the remarkable diversity of responses ChatGPT is capable of providing, from being educational to demonstrating reasoning and creativity. After reading these prompts, participants were once again asked to rate the above-mentioned 9 mind perception questions. The study took about 15 minutes for an individual to complete.

Data Analysis

Data analyses and visualizations were conducted in R (v4.2.1; R Core Team 2021) using the packages tidyverse (Wickham et al., 2019) and papaja (Aust & Barth, 2018). Data are publicly available on OSF at https://osf.io/tye3k/.

Results

Mind Survey Scale

Items from the Mind survey (Gray et al., 2007) were averaged across the items corresponding to agency and the items corresponding to experience. These scores were then analyzed with a 2x2 within-subjects ANOVA with timepoint (pre versus post) and attribute (agency versus experience) as factors (see Figure 5.1). There was a significant main effect of

timepoint, $F(1,132) = 7.84$, $p = .006$, $\hat{\eta}_G^2 = .002$, no main effect of attribute, $F(1,132) = 0.95$, $p = .332$, $\hat{\eta}_G^2 = .002$, and the timepoint by attribute interaction was not significant, $F(1,132) = 1.00$, $p = .320$, $\hat{\eta}_G^2 < .001$. This analysis indicates that agency scores and experience scores were significantly higher to a similar degree post-exposure ($M_{Agency} = 4.27$, $SD_{Agency} = 0.75$, $M_{Experience} = 4.35$, $SD_{Experience} = 0.71$) compared to pre-exposure ($M_{Agency} = 4.22$, $SD_{Agency} = 0.73$, $M_{Experience} = 4.26$, $SD_{Experience} = 0.74$).

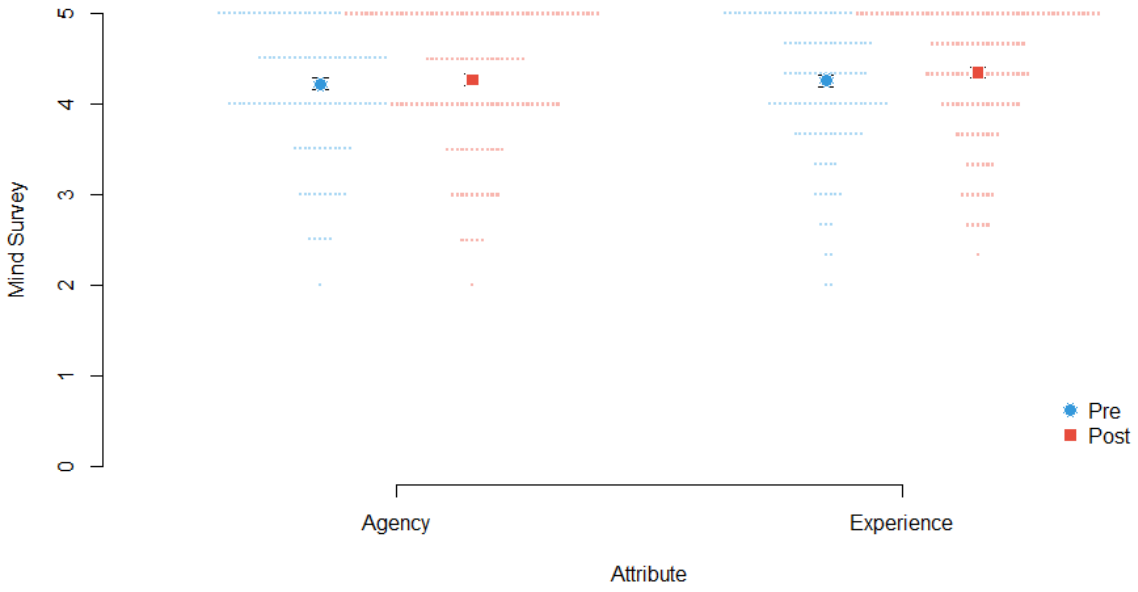

Figure 5.1: Scaled self-ratings of mind perception pre-and post-exposure. Error bars indicate SEM.

Self-Ratings Single Item Measures

Using the single-item mind perception assessments, a 2x2 within-subjects ANOVA was also conducted for the self-perception ratings made before and after exposure to the ChatGPT

prompts. The main effects of attribute, $F(1,132) = 14.37,\ p < .001, \hat{\eta}^2_G = .015$, and timepoint, $F(1,132) = 4.64,\ p = .033, \hat{\eta}^2_G = .001$, were significant, and there was no interaction, $F(1,132) = 0.47,\ p = .500, \hat{\eta}^2_G < .001$. This analysis indicates that self-attributions of experience were significantly higher than that of agency ($M_{Experience} = 89.8$, $SD_{Experience} = 13.9$; $M_{Agency} = 86.3$, $SD_{Agency} = 16.4$), with the increase from pre-exposure ($M_{Experience} = 89.8$, $SD_{Experience} = 13.9$; $M_{Agency} = 86.3$, $SD_{Agency} = 16.4$) to post-exposure ($M_{Experience} = 91.1$, $SD_{Experience} = 13.4$; $M_{Agency} = 87.1$, $SD_{Agency} = 16.4$) being equivalent for both forms of attribution. Correlations between the Mind Survey scores and single-item responses for agency ($r = .47$) and experience ($r = .57$) revealed they were strongly related.

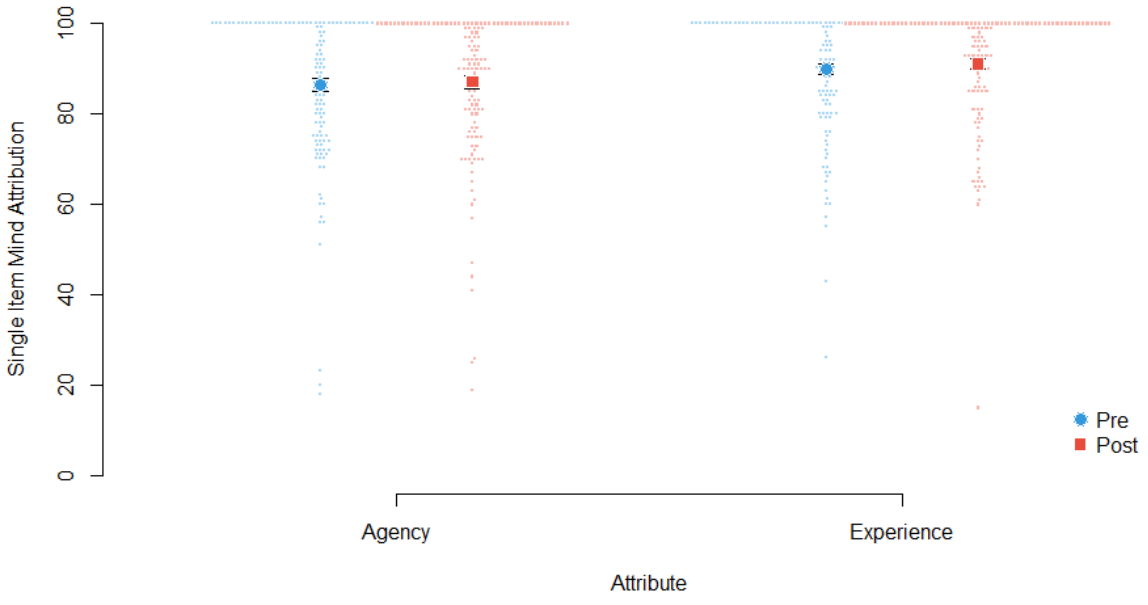

Figure 5.2: Single-item self-ratings of mind perception pre-and post-exposure. Error bars indicate SEM.

Human Uniqueness

A 2x2 within-subject ANOVA was also applied to the agency and experience (single items) attributions measuring the degree to which they are perceived to be unique to humans (see Figure 5.3). As illustrated, experience ratings are higher overall than those for agency ($M_{Experience}$ = 86.9 vs M_{Agency} = 65.0) with both falling a similar degree from pre-exposure ($M_{Experience}$ = 88.7, $SD_{Experience}$ = 19.2; M_{Agency} = 67.1, SD_{Agency} = 28.7) to post-exposure ($M_{Experience}$ = 85.1, $SD_{Experience}$ = 22.7; M_{Agency} = 62.9, SD_{Agency} = 28.7). This was confirmed statistically, with significant main effects for attribution, $F(1,132) = 65.61$, $p < .001$, $\hat{\eta}_G^2 = .161$, and timepoint, $F(1,132) = 7.62$, $p = .007$, $\hat{\eta}_G^2 = .006$, and no interaction, $F(1,132) = 0.03$, $p = .857$, $\hat{\eta}_G^2 < .001$.

These single-item measures indicate that self-perception attributions increase post-exposure, while attributions for human uniqueness decrease post-exposure. This divergence was confirmed statistically by entering question type (self-perception, human uniqueness) into the within-subject ANOVA, which returned a significant question type by timepoint interaction, $F(1,132)=10.95$, $p = .001$, $\hat{\eta}_G^2 = .004$.

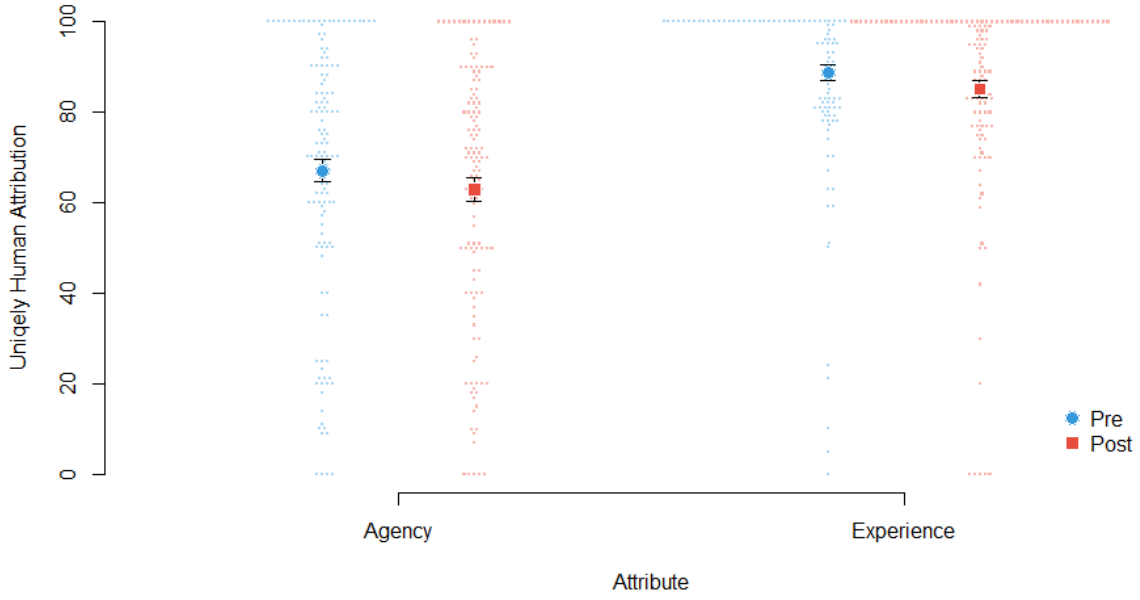

Figure 5.3: Mind perception ratings (single item) pre-and post-exposure indicating the degree to which each attribute is uniquely human. Error bars indicate SEM.

Discussion

Anthropomorphism involves the attribution of human-like qualities to nonhumans (Epley et al., 2007; Waytz et al., 2010a). Large language models (LLMs) with their increased capabilities and popularity offer a compelling opportunity to investigate the consequences of anthropomorphism on human perception. The mind perception framework is a popular psychological tool for measuring mental attribution that has previously been applied to examine self-perception and other-perception (Waytz et al., 2010b; Jacobs et al., 2023b). This framework simplifies a wide variety of qualities of mind into their principal components: agency (the ability to do), and experience (the ability to feel) (Gray et al., 2007). Using this framework, we sought to probe self-perception and the features that individuals consider uniquely human before and after exposure to ChatGPT. Drawing from previous work that has used this mind perception framework (Jacobs et al., 2023a), we predicted that after exposure to ChatGPT, people will attribute greater experience to themselves and humans, in general, compared to agentic attributions.

Our investigation returned two major findings. First, self-perception was malleable as a function of exposure to ChatGPT prompts. Participants increased the amount of agency and experience they attributed to themselves after exposure to ChatGPT. We had predicted this effect to be especially salient for experience; that a comparable result occurred for agency was unexpected. These results could be a reflection of a shift in conceptual categories, which can be understood in terms of prototype theory (Rosch, 1973). Participants may consider themselves a better conceptual member (or prototype) of an entity—both in what they can do and feel—post-exposure to a LLM like ChatGPT. This is in effect also similar to social comparison theory (Festinger, 1954; Gerber et al., 2018). However, unlike social comparison theory which focuses

on how people compare themselves to other people, in the present case, the comparison concerns LLMs. This sort of downward comparison could explain the elevated self-perception of one's capabilities. Notably, these were consistent across both methods of measurement: single-item agency and experience assessments and the Mind-perception survey.

Our second major finding was that, unlike self-perception, exposure to the LLM prompts *decreased* perceptions that agency and experience are uniquely human features. This divergence in results demonstrates that separating self-perception and other-perception (e.g., attitudes toward humanity) can reveal distinct effects. Moreover, it suggests that as programs like ChatGPT continue to develop and demonstrate more sophisticated abilities, individuals in turn may increasingly come to view agency and experience as less defining characteristics of humanness. Essentially, this implication is that there may be a change in social perspectives concerning the criteria of what defines uniquely human features. This fits neatly with the idea that if one's ability to think rationally, perform mathematical calculations, and create works of art forms a part of what makes us uniquely human, then as AI encroaches on each of these capabilities, then it follows that people will begin to question if traditional features of mind are unique. However, there was no interaction between facets of mind perception and exposure. Contrary to our hypothesis, people did not disproportionately think of agency as less unique than experience post-exposure. This is surprising considering that the developments in LLMs seem more related to increased agentic capacities compared to experiential capacities.

Importantly, although there was no interaction between exposure and facets of mind perception, people still viewed agency—the ability to do—as less unique to humans in comparison to attributions of experience. For example, many more people indicated experience as unique to humans at ceiling levels compared to agency (see Figure 5.3). This fits with past

work demonstrating high levels of attributing experience to other people and the general reluctance to ascribe experience to nonhumans (Gray et al., 2007). That said, people still considered agency as unique to humans and the difference with regard to experience was only relative, not in kind.

A number of limitations of the present study are worth considering. All participants took part remotely, and the degree to which the chosen prompts surprised participants was not measured. In light of our present results, measuring expectancy violation could shed light on the mechanisms driving the pre- and post-exposure effects. Furthermore, it is possible that the experience of the ChatGPT prompts may differ from interacting with ChatGPT in real time. The importance of real-time interactivity with LLMs remains an outstanding future research question, especially concerning mental attribution and anthropomorphism. Recent work has found that psychological distance, or the subjective experience of something being close or far from the self (Liberman et al., 2007), can be a mediating factor through which anthropomorphism positively affects evaluations of AI systems (Li & Sung, 2021) In the present study, the magnitude of effects was also relatively small. This may in part be because we adopted a conservative method for investigating differences, with the exposure manipulation to ChatGPT being relatively brief. It is possible that the short duration may have increased consistency bias effects among participants as they likely remembered their prior ratings and would desire to remain consistent. Understanding the impacts of long-term exposure as LLMs become more commonplace offers an additional fruitful direction for future research. Future directions may also include examining individual differences as they relate to anthropomorphism and self- and other-perception, given the wide inter-individual variation in mind perception noted elsewhere (e.g., Gray et al., 2011; Tharp et al., 2017).

Collectively the present findings add to the growing literature on the effects of human-computer interaction with LLMs. As the frequency and nature of LLM interactions increase, and with these interactions people become increasingly aware of its capabilities, it is crucial to consider how these interactions might shape attitudes toward ourselves in addition to wider societal attitudes (Turkle, 2017). The overall findings support our working hypothesis that anthropomorphism can in turn affect self-perception and attitudes toward other people's minds more broadly.

Chapter 6

After revealing that the mind perception framework can be used to successfully gain insights into how people view themselves and other people—and that these perceptions are influenced by exposure to LLMs—Chapter 6 focuses on how numerical scales, compared to forced-choice designs, can influence preferences regarding human-and AI-generated artworks. Chapter 6 contains one empirical study (two experiments) and a version of this study has been submitted for publication.

Study 6: Comparative designs reveal preferences for human-generated rather than AI-generated art

Introduction

The last few years have seen a tremendous rise in the use of artificial intelligence (AI) to create works of art. Popular platforms such as DALL-E or Midjourney enable millions of users to quickly generate stunning images with simplistic prompts. While this form of AI-generated art is relatively new, many researchers have already been examining attitudes and perceptions of AI-generated art. For example, there have been numerous studies empirically investigating how people value and judge a work of art based on its provenance being AI or human (Chiarella et al., 2022; Fortuna & Modliński, 2021; Ragot et al., 2020).

It has been hypothesized that people value human-generated art for the agency required to make it, including the artist's intentionality (Snapper et al., 2015), or effort (Kruger et al., 2004). On a related note, the field of research concerning the degree to which people perceive minds in others (i.e., mind perception) has revealed that while people do attribute agency to AI, albeit to a reduced degree compared to humans, there is less willingness to ascribe experiential features

such as feelings to AI (Gray et al., 2007; Jacobs et al., 2022). This has led some (Wu et al., 2021) to suggest that it could be the uncanniness (e.g., Gray & Wegner, 2012) associated with a perceived lack of emotion in AI art that leads to it being appraised less positively than human-generated art.

Most research has supported this preference for human-generated art over AI art (e.g., Fortuna & Modliński, 2021; Ragot et al., 2020). However, not all studies have observed this preference (e.g., Gangadharbatla, 2022; Hong & Curran, 2019). While a number of possible reasons might give rise to these discrepant results (e.g., the painting styles used, true provenance, sample sizes) one likely suspect concerns the way that preferences are measured. Specifically, whether individuals are asked to provide absolute numerical judgements of AI and human art separately, versus being asked to directly choose between the two types. It is our working hypothesis that when people are asked to judge AI and human art in isolation, rather than being asked to make a relative comparison, differences in preferences may become masked. Not only does this hypothesis dovetail with the mixed human vs AI art research findings, but it is consistent with what is known in the field of psychometrics. When people are instructed to give an absolute judgment of how much they value or like an item on a numerical scale of, say, 1–5, the same numbers may represent very different perceptions for different individuals, and conversely, different numbers may represent the same perception (Kreitchmann et al., 2019; Stadthagen-González et al., 2018; Wetzel et al., 2016; Wildt & Mazis, 1978). In contrast, when people are asked to make relative comparisons, for example choosing which of two items they prefer, these potential difficulties with numerical scaling are eliminated and any prevailing preferences in perception can be exposed. The present paper aims to directly test this possibility,

using a non-comparative numerical-scale design in Experiment 1 and a forced-choice comparative design in Experiment 2.

Experiment 1

In Experiment 1, participants were asked to evaluate four categories of paintings (abstract expressionism, abstract geometrical expressionism, romanticism, and naturalism). Artist labels were manipulated such that each painting category was attributed to one of two kinds of artists (human or AI) at one of two levels of expertise (Human: amateur or professional; AI: weak or strong). There was an additional control condition that did not assign any labels to the art leading to 5 conditions. We administered the dependent measures using traditional numerical scales that range from 1 to 5.

Methods

Participants

An a priori power analysis using WebPower (Zhang & Mai, 2018) for a mixed ANOVA with 90% power, a medium-sized effect (*f*=.25), and 5 groups suggested a total sample size of 252 participants. 250 participants took part in the study with 50 participants per group. Participants' mean age was 34.38 (SD = 9.61). There were 75 females and 175 males. No participants were excluded from the analyses. In terms of highest achieved level of education, 6.0% had a 'high school diploma,' 6.4% had an 'Associate degree in college (2-year),' 14.0% had 'Some college but no degree,' 59.6% had a 'Bachelor's degree in college (4-year),' 11.6% had a 'Master's degree,' and 2.0% had a 'Professional or Doctorate degree (JD, MD, PhD)'. Participants were asked to answer 'How knowledgeable are you about paintings?' to which 11.2% of participants reported 'Not knowledgeable at all', 29.2% 'Slightly knowledgeable', 31.2% 'Moderately knowledgeable', 15.2% 'Very Knowledgeable', and 13.2% 'Extremely

knowledgeable'. Participants were also asked 'How knowledgeable are you about A.I. (artificial intelligence?' to which 1.2% responded ''Not knowledgeable at all', 24.8% 'Slightly knowledgeable', 35.2% 'Moderately knowledgeable', 27.2% 'Very Knowledgeable', and 11.6% 'Extremely knowledgeable'. All participants were recruited using CloudResearch and took part from IP addresses listed in the United States.

Materials and Procedure

The study was formatted as a Qualtrics survey. Each survey question was composed of a painting, a one-line description of the artist's identity (Human [amateur or professional] or AI [weak or strong]), and two rating scales (value and liking). The description for the amateur artist was "This painting was created by an amateur artist," for the professional artist it was "This painting was created by a professional artist," for weak AI it was "This painting was created by an AI system from Calgary College of Fine Arts," and for the strong AI label it was "This painting was created by an AI system from Google and MIT."

In total, there were 20 paintings: five abstract, five patterned abstract, five romantic, and five realistic. Altogether, participants were divided into five assigned conditions. To maintain a plausible consistency of style within each artist label group while maintaining a counterbalanced factorial design, each style of art was paired with each artist identity and prestige in one of five conditions (no labels were given in the final condition).

After consenting to participate, participants filled in two questionnaires. The first was a standard demographic questionnaire querying the participant's age, gender, and education. Next, the participants were presented with an image of a painting and a short description of the artist's identity. Participants were asked to rate each of the paintings for value ("How much would you pay for this painting?") from 1-"None at all" to 5-"A great deal," and liking ("How much do you

like this painting?") from 1-"Dislike a great deal" to 5-"Like a great deal." After rating each painting, participants were informed that they had completed the study and were then debriefed and compensated.

Data Analysis and Availability

Data analyses were conducted in R (v4.2.1; R Core Team 2021) using the packages tidyverse (v1.3.2; Wickham et al., 2019), afex (v1.1; Singman et al., 2015), and ggplot2 (v3.4.2; Wickham et al., 2016). This study was approved by the Behavioural Research Ethics Board of the University of British Columbia (H10-00527). Data, materials, and analysis are available on OSF at: https://osf.io/wtqxz/.

Results and Discussion

We conducted a 5 (artist labels) x 4 (painting style) analysis of variance on liking attributions with labels as a between-subjects factor and painting style as a within-subjects factor. There was a significant main effect of painting style, $F(1.96,479.78) = 115.23$, $MSE = 0.57$, $p < .001$, $\hat{\eta}_G^2 = .156$, but no significant main effect of artist label $F(4,245) = 0.62$, $MSE = 1.72$, $p = .651$, $\hat{\eta}_G^2 = .006$, nor was their interaction significant, $F(7.83,479.78) = 0.46$, $MSE = 0.57$, $p = .881$, $\hat{\eta}_G^2 = .003$. Similarly, for the value attributions, the main effect of painting style was significant, $F(2.06,504.27) = 102.41$, $MSE = 0.35$, $p < .001$, $\hat{\eta}_G^2 = .054$, but no significant main effect of artist label, $F(4,245) = 0.41$, $MSE = 4.57$, $p = .802$, $\hat{\eta}_G^2 = .006$, nor was their interaction significant, $F(8.23,504.27) = 1.48$, $MSE = 0.35$, $p = .161$, $\hat{\eta}_G^2 = .003$. Altogether, participants valued and liked realistic paintings over other categories irrespective of the painter's identity (AI or human). See Figures 6.1 and 6.2 for a visualization of the results.

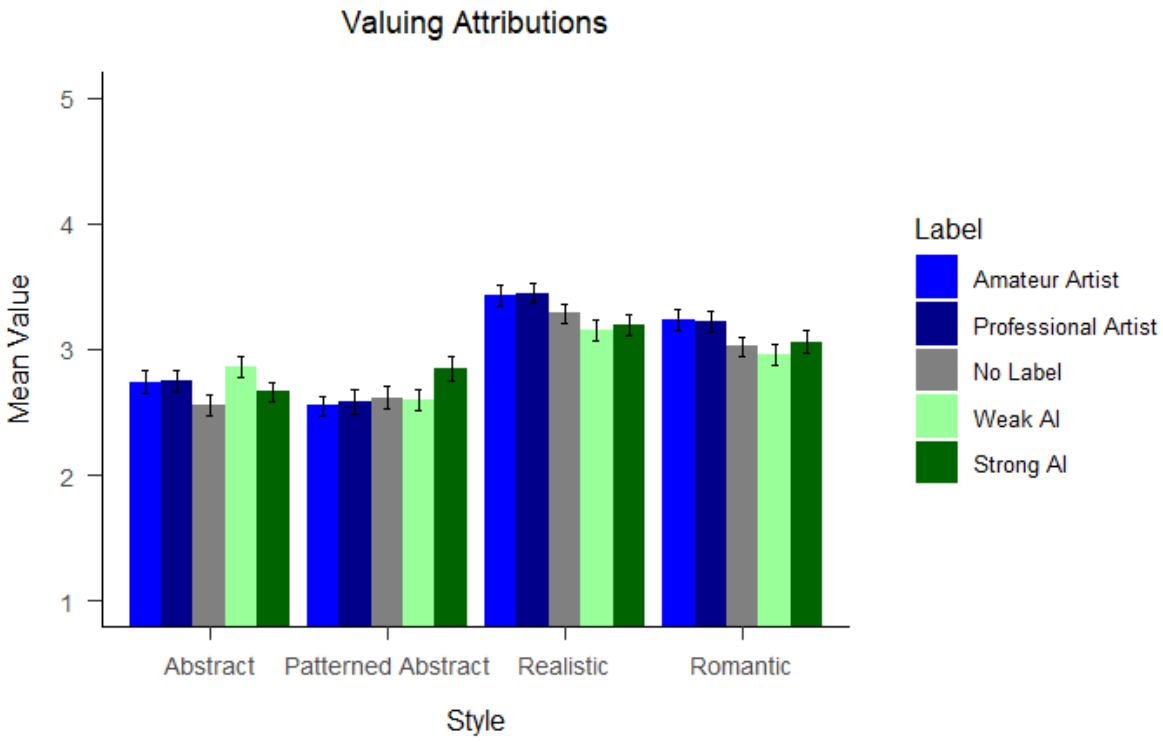

Figure 6.1: Mean valuing attributions by painting style and artist label. Error bars indicate standard error.

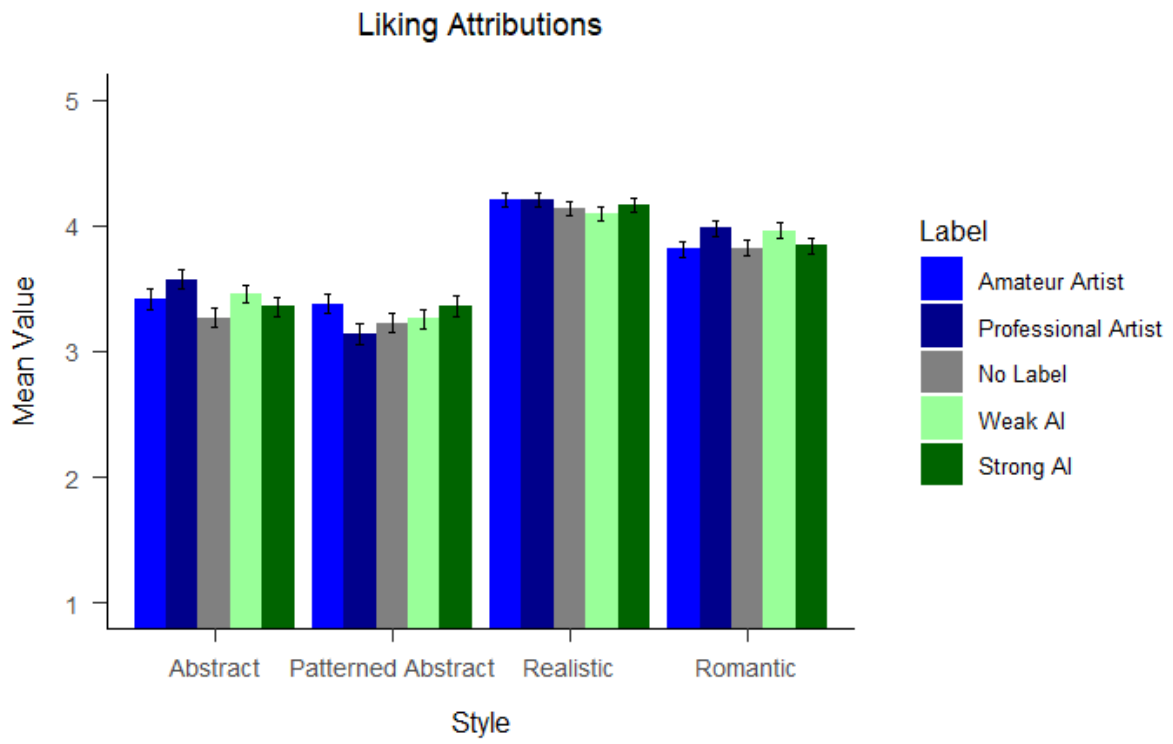

Figure 6.2: Mean liking attributions by painting style and artist label. Error bars indicate standard error.

Experiment 2

The results of Experiment 1 failed to find significant differences between appraisals for art displaying human provenance compared to AI provenance. These null findings replicate the results of a few previous studies (e.g., Gangadharbatla, 2022; Hong & Curran, 2019), but conflict with a myriad of other studies that find significant preferences for human-generated art (e.g., Ragot et al., 2020). These findings support the idea that asking people to provide numerical ratings of each piece on its own may be less sensitive to extracting a reliable preference for human-generated art relative to AI-generated art. A more direct, and sensitive, method may be to 'force' participants to choose which of two pieces of art they prefer: human-generated or AI-generated, while keeping all other factors the same.

In Experiment 2, instead of using a numerical rating task, we required participants to indicate their preference in a two-alternative forced choice (2AFC) task. And to focus purely on the effect of comparative designs, we only manipulated the artist labels of the paintings (human vs. AI) without providing information about the prestige of the artist. We again use a variety of painting styles given previous findings that certain types of art, specifically abstract, are more likely to be associated with AI generation and to remain consistent with the stimuli used in Experiment 1 (Gangadharbatla, 2022).

Methods

Participants

Another WebPower (Zhang & Mai, 2018) power analysis, this time for a 2x2 contingency table test for 90% and a medium-sized effect (w =.32) suggested 102 participants. A total of 102 individuals took part in the study. Participants' mean age was 33.72 (SD = 9.37). There were 31 females and 71 males. No participants were excluded. In terms of highest achieved level of education, 8.82% had a 'high school diploma,' 10.78% had an 'Associate degree in college (2-year),' 10.78% had 'Some college but no degree,' 52.94% had a 'Bachelor's degree in college (4-year),' and 15.69% had a 'Master's degree.' Participants were again asked to answer 'How knowledgeable are you about paintings?' to which 10.8% of participants reported 'Not knowledgeable at all', 22.5% 'Slightly knowledgeable', 37.3% 'Moderately knowledgeable', 23.5% 'Very Knowledgeable', and 5.9% 'Extremely knowledgeable'. Participants were also asked 'How knowledgeable are you about A.I. (artificial intelligence?' to which 2.9% responded ''Not knowledgeable at all', 18.6% 'Slightly knowledgeable', 43.1% 'Moderately knowledgeable', 22.5% 'Very Knowledgeable', and 12.7% 'Extremely knowledgeable'.

Participants were again recruited through CloudResearch using IP addresses based in the United States.

Materials and Procedure

Experiment 2 was also formatted as a Qualtrics survey and followed the same MTurk procedure including the consent process. Each survey question was composed of two paintings, a one-line description of the artist identity (human or AI) for each painting, and two forced-choice ratings (valuing and liking). In total, there were 16 paintings: 4 abstract, 4 patterned abstract, 4 romantic, and 4 realistic. Each identity label was assigned to the same side of the screen throughout the survey, but the arrangement was counterbalanced across participants (i.e., the AI label was always presented on the left to half of the participants and vice versa).

Data Analysis and Availability

Data analyses were conducted in R (v4.2.1; R Core Team 2021) using the packages tidyverse (v1.3.2; Wickham et al., 2019) and ggplot2 (v3.4.2; Wickham et al., 2016). This study was approved by the Behavioural Research Ethics Board of the University of British Columbia (H10-00527). Data, materials, and analysis are available on OSF at: https://osf.io/wtqxz/.

Results

A chi-squared test of independence revealed that switching labels of human and AI artists between groups made a significant difference both for liking preferences, $\chi^2(1, n = 102) = 21.66$, $p < .001$, and valuation preferences, $\chi^2(1, n = 102) = 20.33$, $p < .001$. Participants liked and valued paintings with a human artist label more frequently than an identical painting with an AI artist label—thus indicating an overall preference for human artists.

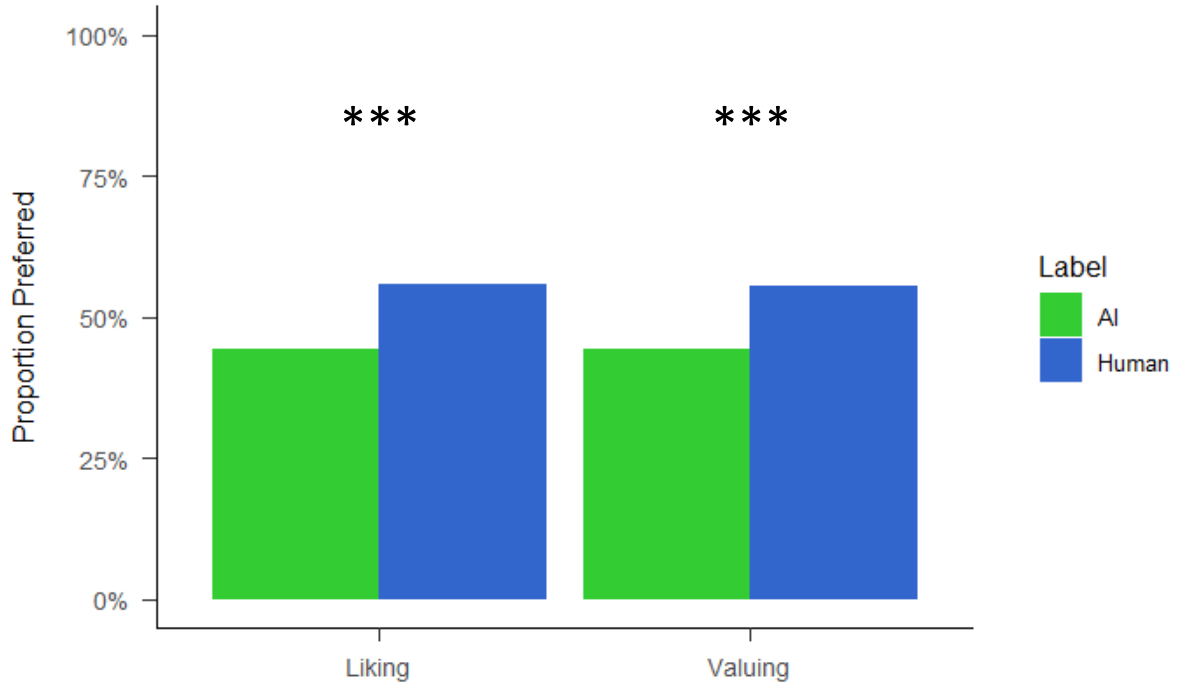

Figure 6.2: Painting selections by artist label and measure. Asterisks indicate p <.001.

General Discussion

The immense rise of public access to high-quality AI-generated art has coincided with more scientific research investigating how the nature of the art's creation influences evaluations of the artwork. A commonly cited finding has been that there exist preferences for human-generated art, perhaps for the enhanced intentionality or effort integral to human creation of art (Elgammal et al., 2017; Fortuna & Modliński, 2021; Ragot et al., 2020). However, the findings for this preference remain mixed (e.g., Gangadharbatla, 2022; Hong & Curran, 2019). Critically, the mixed findings have typically occurred in designs that do not employ comparative contexts and instead involve the use of Likert scales. Comparative designs are scarcer but may have greater sensitivity to underlying preferences when participants are directly probed in a more

comparative fashion (e.g., Chiarella, 2022). In order to shed light on these conflicting findings, over two experiments differing in experimental design, we conducted a survey experiment using human-generated art from four painting styles that were randomly paired with deceptive labels that were either presented without a comparative design (Experiment 1) or with a two-alternative forced choice design (Experiment 2).

In our first experiment, we found no evidence that people had preferences for human-generated art, nor preferences for greater degrees of artist prestige. Similar to Hong and Curran (2019) and Gangadharbatla (2022), people expressed no observable differences in their ratings of liking and valuation towards art based on its creator being human or AI. Though a null finding does not necessarily support a null hypothesis (Leppink et al., 2017), we hypothesized that the lack of emergent preferences may be related to the design involving a non-comparative context. Moreover, given limitations in probing people's attitudes towards art using Likert or continuous scale ratings of attributions (Kreitchmann et al., 2019; Watrin et al., 2019) we thought it was prudent to follow up the results by using a forced-choice dichotomous scale. In this way we could tap more directly into preferences regarding liking and valuation decisions between human- and AI-generated art by forcing participants to choose.

To this end, participants were tasked with deciding which painting to choose between in terms of their liking and valuation—a situation better approximating a prospective buyer perusing an art gallery. Indeed, this difference in method led to completely different results from Experiment 1 with a large preference for human-generated art emerging. These results suggest that people may have underlying preferences for human-created art that may not be consistently captured using more traditional Likert or continuous-scale probes. Comparative designs also provide greater ecological validity in that they better approximate art appraisal in the real world,

as a decision to purchase a particular piece is typically informed by comparisons with other art. It may be that in comparative contexts—both in research designs and in real-world situations—factors such as intentionality and effort become larger considerations in evaluating artworks.

There are some notable limitations to this work. For example, we only used artworks that were originally produced by human artists which may have influenced our results by skewing the believability of the deception. The AI artist labels were always deceptive and may have been less believable than the human artist labels which may have moderated emergent preferences for human provenanced art. We did not check if participants believed the identity labels, so we are unable to test for differences in believability between identity conditions or for an effect of believability on liking and valuation. However, we note that this limitation did not prevent the Experiment 2 finding of preferences for human-generated art. An additional limitation of Experiment 2 was the omission of manipulating artist prestige. While this was done to streamline the methodological comparison, there could be an interaction between how prestige interacts with measurement style, however unlikely. We suggest future studies use AI-generated artworks in similar comparative designs to increase external validity in this line of research. We also did not investigate the roles of individual differences or more complex situations such as co-created art as other studies have examined (e.g., Fortuna & Modliński, 2021). An interesting future direction would be to investigate if human co-creation of art with AI (Oh et al., 2018; Wu et al., 2021) influences appraisals and also benefits from more sensitive comparative designs.

In conclusion, our results revealed that previously mixed findings regarding preferences for human-generated art may be obscured by traditional numerical-rating designs and that by employing comparative designs, such as with the use of forced-choice questionnaires, underlying differences in preference can be exposed. Crucially, this finding has broader implications for the

field of human-computer interaction by suggesting that methodological designs that more explicitly reinforce direct comparisons between humans and AI can induce stronger contrast effects. Finally, comparative designs also provide greater ecological validity for distinguishing between human and AI-generated art as people typically make decisions about art in comparative contexts rather than through Likert-like appraisals in the real world.

Chapter 7:

General Discussion

The studies presented in this thesis sought to explore some of the causes and consequences of mental attribution toward AI and machines via a range of studies and different experimental paradigms. This was a compelling endeavour because of recent technological developments, including generative AI and LLMs, which have renewed essential questions related to ascribing mental states. These include the extent to which mind is afforded to non-sentient AI agents, how this may influence person perception, and which individual factors predict mental attribution toward AI. The purpose of this final chapter is to revisit the questions first proposed in the introduction in light of the results of the studies presented herein. In addition to revisiting these questions, I also provide a discussion of the implications, limitations, and future directions of the current work.

Chapter Summary

Following the Introduction in Chapter 1, in Chapter 2, I provided a taxonomic structure to categorize the ongoing psychological work with LLMs to situate the subsequent chapters. I presented three categorical approaches: the AI-Centered, the Tool-Centered, and the Human-Centered approaches. This latter approach, the Human-Centered approach, is the direction the rest of my following chapters take via the focus on probing people's perceptions of AI. In Chapter 3, I began exploring the extent of mental attribution toward AI and social robots by using the mind perception framework to probe mental attribution toward a wide range of real and fictional robotic or AI characters finding that across characters and dimensions, mind perception

is prevalent. Next, I specifically examined attitudes toward LLMs revealing that exposure to LLMs can increase the degree of agency and experience people attribute to LLMs (Study 2). Chapter 4 investigated how loneliness can influence mental attribution toward LLMs. I found that loneliness, moderated by prior exposure, predicts greater experiential attributions but not agentic attributions toward LLMs. In Chapter 5, I explored if the mind perception framework can also be used to investigate how one views their own mind and the minds of other people (i.e., person perception). Then, I examined if exposure to LLMs can influence how people view their own minds and the attributes people consider uniquely human (Study 5). I found that people, when exposed to LLMs, elevated their own capabilities of mind while expressing that the same agentic and experiential qualities are less unique to other humans. In Chapter 6, I demonstrated the importance of design choices when studying AI perception, such as choosing between numerical scales versus two alternative forced-choice designs, by illustrating that different design choices can lead to stronger preferences for human- versus AI-generated artworks.

The Extent of Mind Perception Toward AI

As David Hume (1757) famously wrote: “There is a universal tendency among mankind to conceive all beings like themselves.” However, to what extent does this apply to machines or disembodied AI systems like LLMs? Cumulatively, the present research, with some caveats and exceptions, supports the notion that there is a general tendency to anthropomorphize AI and LLMs. In Chapter 3, participants were recruited to rate the mental capabilities of a wide range of real and fictional robots using the mind perception framework. Across many different AI or robotic characters, participants attributed both agentic and experiential qualities to AI. This was

also observed when rating an LLM (Studies 3, 4, 6) and was true for both single-item measurements of agency and experience, and valid for scales differentiating between subcomponents of agentic and experiential attributions. Although these agency and experience attributions were consistent across experiments—participants were much less likely to attribute experience to the same degree as they would agency. This fits the general pattern that experiential qualities are ascribed to robots or AI systems less frequently and to a reduced degree compared to agency (e.g., Gray et al., 2007; Gray & Wegner, 2012; Yam et al., 2021). Thus, the collective evidence suggests that people do perceive a range of mental features in AI despite the lack of sentience, at least, when probed in the direct, self-report method used throughout the present studies.

Upon revisiting the takeaways from examining the extent of mind perception toward AI and LLMs, several ideas regarding what this means for our understanding of cognition are worth considering. First, the results support the idea that mental attribution is rather flexible and widespread—while it may have evolved for social facilitation (Epley et al., 2007) or predicting future behaviour (Dennett, 1988), it can extend to technological innovations and novel environments. Second, it supposes that social cognition can be broader in scope than perhaps conventionally imagined. Instead of defining social cognition more narrowly as the various psychological processes that enable individuals to socialize with other people (Frith, 2008), perhaps in time, as more research explores the various ways beliefs can be influenced by anthropomorphism, we will see a broader definition of social cognition develop; one that refers to processes related to social agents including nonhuman and/or non-sentient agents.

The general finding that people are open to anthropomorphizing robots and AI in cognitive and emotional ways has significant implications for AI safety initiatives. For example,

safe AI requires trust and greater anthropomorphism is associated with greater trust for a variety of types of AI (Epley, Waytz, et al., 2008; Inie et al., 2024). Research has also shown that people are open to blaming or holding robots morally responsible the more mind that is afforded (Hindennach et al., 2024; Malle et al., 2015) and that there is some awareness of the danger in overattributing blame to robots (Stuart & Kneer, 2021). There is also a growing area of research concerning *moral HCI* (Malle et al., 2015) with some research suggesting that people can shift between more utilitarian versus deontological considerations when automation is involved (Bonnefon et al., 2016; Schurr & Moran, 2023). These connections between moral culpability and anthropomorphism reflect the longer history of coupling mind perception and moral judgments (Gray et al., 2012a,b). However, greater anthropomorphism is not ubiquitously associated with more positive outcomes for AI safety. For example, recent AI safety initiatives have also explored methods of preventing overreliance on AI that can be dangerous such as with semi-autonomous vehicles or decisions involving severe consequences like risky financial decisions (Bonnefon et al., 2016; Klingbeil et al., 2024). A future direction in studying anthropomorphic tendencies toward AI and LLMs includes a more detailed examination of the considerations for AI safety and the moral repercussions of such tendencies. The general prevalence and flexibility of anthropomorphic tendencies found in the present thesis support this growing recognition that designing safer AI systems requires considering people's tendencies to anthropomorphize and, occasionally, over-rely on AI.

Mind Perception and Person Perception

The second key area of interest of this thesis was to investigate whether exposure to LLMs could influence person perception, or the processes involved in categorizing and making inferences about people's behaviour (Moskowitz & Gill, 2013). This issue was significantly more novel for several reasons. First, the mind perception framework has not been used to investigate person perception (Jacobs et al., 2023). And second, the influence of anthropomorphism on person perception has rarely received attention (Jacobs et al., 2024). Investigating how anthropomorphic tendencies can influence person perception has been only lightly touched upon in other works that allude to these potential effects, rather than explicitly seeking to uncover or understand them (e.g., Brette, 2022; Kiesler et al., 2008).

In Chapter 5, I began examining this research space by applying the mind perception framework to investigate self-discrepancies. Self-discrepancy theory (Higgins, 1987; Higgins et al., 1985) offers a foundational framework for examining beliefs about one's own mind. While there has been considerable research into beliefs about the self in other domains (e.g., body image), there is relatively little research regarding beliefs about one's mind, with the exception of self-discrepancy theory. The other benefit of applying the mind perception framework was that it helped address a key problem of self-discrepancy theory—distilling a wide variety of mental qualities or features into essential components that are easily measurable and have an empirical justification (e.g., Watson et al., 2010, 2016). Although Gray et al.'s (2007) mind perception framework is far from a perfect tool (see limitations below), its application proved effective for probing person perception.

Following the demonstration that the mind perception framework can be used to probe person perception, Study 6 examined whether exposure to LLMs can influence person perception. Participants were recruited to evaluate their perceived agency and experience pre-

and post-exposure to ChatGPT prompts. Additionally, participants were tasked with assessing the degree to which they perceived agency and experience as uniquely human. The results revealed diverging effects for the influence of exposure: enhancing the perception of one's own mind while reducing its uniqueness for other people. The diverging effects supported the hypothesis that social comparison theory (Festinger, 1954; Gerber et al., 2018) can be extended to include nonhuman social agents. That is, the perceived cognitive and emotional qualities in AI applications influenced how people viewed themselves and others. These findings may have also reflected that the design of the study—the pre-and post-exposure manipulation—may have led participants to shift the benchmarking of their evaluations. In other words, similar to prototype theory (Rosch, 1973), participants' self-elevations post-exposure may have been, in part, a consequence of comparing their own abilities with ChatGPT. The discrepancy between the effect of exposure on how people viewed themselves versus others may emerge for another type of self-serving bias. For example, one could imagine that people might believe that they are consistently a better writer, creative thinker, or code developer than an LLM, but they may not hold that belief for others. Similarly, an individual may believe that they will always be a better driver than autonomous cars, but would not extend this opinion to other drivers.

Individual Differences in Mind Perception Toward AI

Probing the role of individual differences was a theme throughout the collection of the studies in this thesis. Past work has found that a multitude of individual factors predict greater anthropomorphism of robots and AI including demographic features such as age, gender, and culture in addition to personality factors (Eyssel et al., 2012; Rossi et al., 2020; Syrdal et al.,

2020). In Study 1, only age and sex measures were examined for any associations with the mind perception ratings. Analyses revealed that age, but not sex, was associated with both agency and experience attributions toward a wide set of robots. Study 2 also revealed a similar pattern, this time for attributions toward LLMs. Specifically, younger participants gave higher scores than older adults suggesting that there may be generational differences in attitudes toward perceiving human features in robots or AI systems. Or alternatively, as people age, they become more reluctant to ascribing human features to robots. This tendency for younger people to anthropomorphize more frequently has been observed elsewhere, although it typically has been reported in contexts comparing younger and older children and does not appear to be a robust effect (Manzi et al., 2021; Pak et al., 2020; Thellman et al., 2022). In Study 3, contrary to Study 1, participant age was not observed to be predictive of agentic or experiential attributions to LLMs. Younger individuals were more likely to have used or interacted with LLMs, and the relationship between exposure and mental attribution appears to be a more robust effect (e.g., Study 3, 4, 6). Age may be a moderator, wherein younger individuals are more likely to ascribe mental states to robots and LLMs, but the connection is weak or unreliable. Moreover, the association appears largely driven by differences in prior exposure to robots and AI.

In Study 3, individual differences were investigated beyond simple demographics like sex and age. Specifically, the Ten-Item Personality Inventory (TIPI) (Gosling et al., 2003) , the Individual Differences in Anthropomorphism Questionnaire (IDAQ) (Waytz, Cacioppo, et al., 2010), and the Autism Quotient 10 (AQ10) (Allison et al., 2012; Baron-Cohen et al., 2001) were administered in addition to collecting age, gender, and education. Prior exposure predicted agency and experience attributions across the two online experiments (Experiments 1 and 3). The IDAQ also yielded more consistent effects again for both agency and experience. As expected,

mental attribution toward LLMs was strongly related to anthropomorphism toward other targets like animals and other non-sentient entities like an ocean. Agreeableness and conscientiousness were also consistently related to agency attributions while extraversion was consistently related to experience attributions. Emotional stability and openness were more inconsistent in the pattern of results or any associations were null.

Study 4 also examined loneliness as a predictor of mental attribution toward LLMs. This was a particularly interesting individual difference factor because of the frequency of people using LLM interfaces akin to a therapist and the strong connections between social isolation and anthropomorphism (Epley et al., 2008a,b; Eyssel & Reich, 2013; Lee et al., 2006; Reich & Eyssel, 2013). While previous research had extended the connection between loneliness and anthropomorphizing chatbots (Sheehan et al., 2020, Folk et al., 2023), the benefit of using the mind perception framework in this thesis was that it differentiates between types of psychological anthropomorphism (i.e., agency and experience).

This differentiation proved to be key. While agency attributions were not associated with loneliness scores, experience attributions were significantly related. This relationship was also moderated by the degree of prior exposure participants had with LLMs. These findings support theories about why people anthropomorphize. Experience is more closely related to the emotional, feeling-oriented types of mental attribution in contrast to the more cognitive, goal-orientated mental attribution that agency is associated with. Loneliness is marked more by social-emotional deficits than cognitive or goal-oriented deficits. This fits with the Epley et al.'s (2007) theory that social connection is a primary motivator of anthropomorphism. It also fits with the the need for control motivator of anthropomorphism—individuals may seek to symbolically

fulfill social or emotional needs by reappraising social perceptions in their environment (Epley et al., 2007; Kay et al., 2009).

A future direction concerning individual differences and their associations with mind perception toward AI is to explore several traits that were absent from the studies presented here. The most obvious omission was a measure of general intelligence. Intelligence is widely regarded as one of the most important predictors of individual behaviour (Nisbett et al., 2012), leaving its absence particularly conspicuous. While education was examined and could, in some ways, be considered an indirect measure, education level is not an ideal proxy for intelligence. It would also have been interesting to examine education from the perspective of field of study, rather than education from the perspective of highest level achieved. Similarly, differences may emerge as a function of job occupation. Anecdotally, there appears to be a general trend where engineers and computer scientists are more skeptical of anthropomorphizing LLMs and AI more generally compared to people working in other fields. There are some notable exceptions to this observation such as the engineer fired by Google for claiming one of its chatbots is sentient (Grant, 2022) or Geoffery Hinton's bolder statements describing LLMs as capable of reasoning (Metz, 2023). Future studies could examine how technical understandings of LLMs may influence any mental attribution toward them.

Limitations

While each of the studies had idiosyncratic limitations discussed in more detail in each of their sections, there are several overarching thematic limitations to the present work. A major limitation has been the choice method of probing mental attribution throughout my prior work—

the mind perception framework. For example, it is not evident that the Gray et al. 2-factor method is the best dimensional structure as others have pointed out (Tamir & Thornton, 2018; Tzelios et al., 2022). It also could be the case that for different applications different factor structures could be more optimal (Lee et al., 2006; Malle, 2019; Tamir & Thornton, 2018; Tzelios et al., 2022). This was most evident when applying the mind perception framework toward person perception where it had previously not been applied. While the framework did provide a parsimonious means of examining self-discrepancies, it could be the case that how people cluster mental features depends on whether they are reflecting on themselves, or others, and if they are doing so regarding the present or regarding the future.

Another issue with the mind perception framework was also highlighted by the work on person perception in Chapter 4. There appear to be asymmetries in the items loading onto agency and experience in terms of their social desirability. While most features of agency prima facie are desirable, positive features, many of the subcomponents of experience have negative associations. For example, fear, pain, and embarrassment all can be considered negative emotions or mental states. One direction for future research would be to explicitly examine if these asymmetries are of serious consequence. One way to go about answering this would be to use antonyms for many of the negatively associated items underlying the experience factor. For example, this could include switching select words such as hunger to satiety or embarrassment to comfort.

Another limitation of the present work is the many unknowns regarding the future and the lasting impact of the present findings. The field of AI and the study of AI perception are moving so fast that it becomes hard to predict where LLMs or AI will stand in the future. Experts in the field vary drastically in their predictions and even experts on any given subject can be

notoriously poor at predicting how things will change in the future (Burgman et al., 2011). I do believe that we can turn toward some prior trends to make somewhat more informed speculation about the direction of anthropomorphism toward AI in the future. In general, both my research, and the field more broadly, tend to align on the notion that anthropomorphism is a relatively universal tendency—perhaps born out of the need and desire to foster social connection, facilitate communication, and make sense of the world. These evolutionary tendencies are likely so ingrained that there is no reason to suggest they will disappear. As philosophical treaties from over 2000 years ago attest—anthropomorphism is a part of the human condition (Lesher, 2023). Moreover, throughout a number of the studies in this dissertation (e.g., Studies 3, 4, 6), greater levels of prior exposure predicted greater anthropomorphism, suggesting that perhaps in the future when people are more familiar with AI applications, they will anthropomorphize more, not less.

One understandable reason for being skeptical that people will anthropomorphize LLMs more over time comes from Dennett's work on the intentional stance. One can argue that LLMs are so novel and foreign, that while relying on the intentional stance in the present for making sense of their behaviour, as people develop more experience with them, it will become easier and easier to rely on the design stance. Just as other technologies, like smartphones, have become more integrated with daily life, the range of affordances offered by LLMs may become clearer over time leading users to be able to interpret its "behaviour" as understandable in terms of what LLMs are designed to do. The challenge associated with predicting how mental attribution to technology will develop over time is one of the compelling reasons for this dissertation as the thesis provides insight into an important moment in the history of AI perception.

Conclusion

This thesis provided an empirical approach to investigating mental attribution toward AI. Through eleven experiments, I demonstrated that people frequently attribute both agency and experience to AI, with substantial individual variation in these tendencies reflecting factors such as age, prior exposure to AI, and perceived social isolation. Furthermore, I revealed that these tendencies to attribute agency and experience to AI influence how people view their own minds—and those of other people. Together, these insights deepen our understanding of human-AI interaction and the psychological underpinnings of mental attribution.

References

Adam, E. K., Hawkley, L. C., Kudielka, B. M., & Cacioppo, J. T. (2006). Day-to-day dynamics of experience–cortisol associations in a population-based sample of older adults. *Proceedings of the National Academy of Sciences*, *103*(45), 17058–17063. https://doi.org/10.1073/pnas.0605053103

Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications*, *2*, 100006. https://doi.org/10.1016/j.mlwa.2020.100006

Addlesee, A., Cherakara, N., Nelson, N., Hernández García, D., Gunson, N., Sieińska, W., Romeo, M., Dondrup, C., & Lemon, O. (2024). A multi-party conversational social robot using LLMs. *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 1273–1275. https://doi.org/10.1145/3610978.3641112

Ahadzadeh, A. S., Pahlevan Sharif, S., & Ong, F. S. (2017). Self-schema and self-discrepancy mediate the influence of Instagram usage on body image satisfaction among youth. *Computers in Human Behavior*, *68*, 8–16. https://doi.org/10.1016/j.chb.2016.11.011

Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Toward brief "red flags" for autism screening: The short autism spectrum quotient and the short quantitative checklist in 1,000 cases and 3,000 controls. *Journal of the American Academy of Child & Adolescent Psychiatry*, *51*(2), 202-212.e7. https://doi.org/10.1016/j.jaac.2011.11.003

Altabe, M., & Thompson, J. K. (1996). Body image: A cognitive self-schema construct? *Cognitive Therapy and Research*, *20*(2), 171–193. https://doi.org/10.1007/BF02228033

Appel, M., Izydorczyk, D., Weber, S., Mara, M., & Lischetzke, T. (2020). The uncanny of mind in a machine: Humanoid robots as tools, agents, and experiencers. *Computers in Human Behavior*, *102*, 274–286. https://doi.org/10.1016/j.chb.2019.07.031

Avramides, A. (2023). Other Minds. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2023). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2023/entries/other-minds/

Aydın, Ö., & Karaarslan, E. (2022). OpenAI ChatGPT generated literature review: Digital twin in healthcare. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4308687

Baars, B. J., & Gage, N. M. (2010). *Cognition, brain, and consciousness: Introduction to cognitive neuroscience* (Second edition). Elsevier, Academic Press.

Bao, M. (2019). Can home use of speech-enabled artificial intelligence mitigate foreign language anxiety – investigation of a concept. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3431734

Barker, E., & Stalley, R. F. (1995). *Aristotle politics*. Oxford University Press.

Barnett, M. D., Moore, J. M., & Harp, A. R. (2017). Who we are and how we feel: Self-discrepancy theory and specific affective states. *Personality and Individual Differences*, *111*, 232–237. https://doi.org/10.1016/j.paid.2017.02.024

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

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism,

males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, *31*(1), 5–17. https://doi.org/10.1023/A:1005653411471

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

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*(1), 71–81. https://doi.org/10.1007/s12369-008-0001-3

Bartz, J. A., Tchalova, K., & Fenerci, C. (2016). Reminders of social connection can attenuate anthropomorphism: A replication and extension of Epley, Akalis, Waytz, and Cacioppo (2008). *Psychological Science*, *27*(12), 1644–1650. https://doi.org/10.1177/0956797616668510

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2003). *lme4: Linear Mixed-Effects Models using "Eigen" and S4* (p. 1.1-35.5) [Dataset]. https://doi.org/10.32614/CRAN.package.lme4

Baumeister, R. F., & Finkel, E. J. (Eds.). (2010). *Advanced social psychology: The state of the science*. Oxford University Press.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bergstrom, R. L., & Neighbors, C. (2006). Body image disturbance and the social norms approach: An integrative review of the literature. *Journal of Social and Clinical Psychology*, *25*(9), 975–1000. https://doi.org/10.1521/jscp.2006.25.9.975

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

Boden, M. A. (1998). Creativity and artificial intelligence. *Artificial Intelligence*, *103*(1–2), 347–356. https://doi.org/10.1016/S0004-3702(98)00055-1

Boden, M. A. (2009). Computer models of creativity. *AI Magazine*, *30*(3), 23–34. https://doi.org/10.1609/aimag.v30i3.2254

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573–1576. https://doi.org/10.1126/science.aaf2654

Booth, T., Murray, A. L., McKenzie, K., Kuenssberg, R., O'Donnell, M., & Burnett, H. (2013). Brief report: An evaluation of the AQ-10 as a brief screening instrument for ASD in adults. *Journal of Autism and Developmental Disorders*, *43*(12), 2997–3000. https://doi.org/10.1007/s10803-013-1844-5

Borji, A. (2023). Stochastic parrots or intelligent systems? A perspective on true depth of understanding in LLMs. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4507038

Bozza, S., Roten, C.-A., Jover, A., Cammarota, V., Pousaz, L., & Taroni, F. (2023). A model-independent redundancy measure for human versus ChatGPT authorship discrimination

using a Bayesian probabilistic approach. *Scientific Reports*, *13*(1), 19217. https://doi.org/10.1038/s41598-023-46390-8

Brette, R. (2022). Brains as computers: Metaphor, analogy, theory or fact? *Frontiers in Ecology and Evolution*, *10*, 878729. https://doi.org/10.3389/fevo.2022.878729

Brüne, M., Abdel-Hamid, M., Lehmkämper, C., & Sonntag, C. (2007). Mental state attribution, neurocognitive functioning, and psychopathology: What predicts poor social competence in schizophrenia best? *Schizophrenia Research*, *92*(1–3), 151–159. https://doi.org/10.1016/j.schres.2007.01.006

Burgman, M. A., McBride, M., Ashton, R., Speirs-Bridge, A., Flander, L., Wintle, B., Fidler, F., Rumpff, L., & Twardy, C. (2011). Expert status and performance. *PLoS ONE*, *6*(7), e22998. https://doi.org/10.1371/journal.pone.0022998

Cabanac, G., Labbé, C., & Magazinov, A. (2021). *Tortured phrases: A dubious writing style emerging in science. Evidence of critical issues affecting established journals* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2107.06751

Cacioppo, J. T., & Cacioppo, S. (2018). The growing problem of loneliness. *The Lancet*, *391*(10119), 426. https://doi.org/10.1016/S0140-6736(18)30142-9

Cacioppo, J. T., & Patrick, W. (2008). Are humans unique? *Nature Neuroscience*, *11*(10), 1119–1119. https://doi.org/10.1038/nn1008-1119

Cairó, O. (2011). External measures of cognition. *Frontiers in Human Neuroscience*, *5*. https://doi.org/10.3389/fnhum.2011.00108

Carver, C. S., Lawrence, J. W., & Scheier, M. F. (1999). Self-discrepancies and affect: Incorporating the role of feared selves. *Personality and Social Psychology Bulletin*, *25*(7), 783–792. https://doi.org/10.1177/0146167299025007002

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*(6), 2156–2160. https://doi.org/10.1016/j.chb.2013.05.009

Cassam, Q. (2007). *The Possibility of Knowledge*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199208319.001.0001

Černý, M. (2022). The history of chatbots: The journey from psychological experiment to educational object. *Journal of Applied Technical and Educational Sciences*, Vol. 12 No. 3 (2022): 2022/3. https://doi.org/10.24368/JATES322

Chalmers, D. (2017). The hard problem of consciousness. In S. Schneider & M. Velmans (Eds.), *The Blackwell Companion to Consciousness* (1st ed., pp. 32–42). Wiley. https://doi.org/10.1002/9781119132363.ch3

Chalmers, D. J. (1992). *Subsymbolic computation and the chinese room*.

Chalmers, D. J. (2023). *Could a Large Language Model be conscious?* https://doi.org/10.48550/ARXIV.2303.07103

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A survey on

evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1–45. https://doi.org/10.1145/3641289

Chemero, A. (2023). LLMs differ from human cognition because they are not embodied. *Nature Human Behaviour*, *7*(11), 1828–1829. https://doi.org/10.1038/s41562-023-01723-5

Chen, X., Roberts, R., Liu, Z., & Tong, W. (2023). A generative adversarial network model alternative to animal studies for clinical pathology assessment. *Nature Communications*, *14*(1), 7141. https://doi.org/10.1038/s41467-023-42933-9

Chiarella, S. G., Torromino, G., Gagliardi, D. M., Rossi, D., Babiloni, F., & Cartocci, G. (2022). Investigating the negative bias towards artificial intelligence: Effects of prior assignment of AI-authorship on the aesthetic appreciation of abstract paintings. *Computers in Human Behavior*, *137*, 107406. https://doi.org/10.1016/j.chb.2022.107406

Choudhury, A., & Shamszare, H. (2023). Investigating the impact of user trust on the adoption and use of ChatGPT: Survey analysis. *Journal of Medical Internet Research*, *25*, e47184. https://doi.org/10.2196/47184

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). *Training verifiers to solve math word problems* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2110.14168

Cole, D. (2024). The Chinese Room Argument. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2024). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2024/entries/chinese-room/

Colombatto, C., & Fleming, S. M. (2024). Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness*, *2024*(1), niae013. https://doi.org/10.1093/nc/niae013

Corcoran, R., Mercer, G., & Frith, C. D. (1995). Schizophrenia, symptomatology and social inference: Investigating "theory of mind" in people with schizophrenia. *Schizophrenia Research*, *17*(1), 5–13. https://doi.org/10.1016/0920-9964(95)00024-G

Cornette, M. M., Strauman, T. J., Abramson, L. Y., & Busch, A. M. (2009). Self-discrepancy and suicidal ideation. *Cognition & Emotion*, *23*(3), 504–527. https://doi.org/10.1080/02699930802012005

Darwin, C. (with Huxley, J.). (1872). *The Origin of Species: 150th Anniversary Edition* (2nd ed). Penguin Publishing Group.

De Freitas, J., Agarwal, S., Schmitt, B., & Haslam, N. (2023). Psychological factors underlying attitudes toward AI tools. *Nature Human Behaviour*, *7*(11), 1845–1854. https://doi.org/10.1038/s41562-023-01734-2

De Graaf, M. M. A., & Malle, B. F. (2018). People's judgments of human and robot behaviors: A robust set of behaviors and some discrepancies. *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 97–98. https://doi.org/10.1145/3173386.3177051

De Vega, M., Glenberg, A., & Graesser, A. (2008). *Symbols and embodiment debates on meaning and cognition*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199217274.001.0001

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using Large Language Models in psychology. *Nature Reviews Psychology*. https://doi.org/10.1038/s44159-023-00241-5

Dennett, D. C. (1988). Précis of the intentional stance. *Behavioral and Brain Sciences*, *11*(03), 495. https://doi.org/10.1017/S0140525X00058611

Dhar, V. (2024). The paradigm shifts in artificial intelligence. *Communications of the ACM*, *67*(11), 50–59. https://doi.org/10.1145/3664804

Dijkstra, R., Genc, Z., Kayal, S., & Kamps, J. (n.d.). *Reading comprehension quiz generation using generative pre-trained transformers*.

DiSalvo, C. F., Gemperle, F., Forlizzi, J., & Kiesler, S. (2002). All robots are not created equal: The design and perception of humanoid robot heads. *Proceedings of the 4th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*, 321–326. https://doi.org/10.1145/778712.778756

Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLOS ONE*, *18*(3), e0279720. https://doi.org/10.1371/journal.pone.0279720

Duong, D., & Solomon, B. D. (2024). Analysis of large-language model versus human performance for genetics questions. *European Journal of Human Genetics*, *32*(4), 466–468. https://doi.org/10.1038/s41431-023-01396-8

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Edwards, S., Jenkins, R., Jacobs, O., & Kingstone, A. (2024). The medium modulates the medusa effect: Perceived mind in analogue and digital images. *Cognition*, *249*, 105827. https://doi.org/10.1016/j.cognition.2024.105827

Elali, F. R., & Rachid, L. N. (2023). AI-generated research paper fabrication and plagiarism in the scientific community. *Patterns*, *4*(3), 100706. https://doi.org/10.1016/j.patter.2023.100706

Elgammal, A., Liu, B., Elhoseiny, M., & Mazzone, M. (2017). *Can: Creative adversarial networks, generating "Art" by learning about styles and deviating from style norms* (No. arXiv:1706.07068). arXiv. http://arxiv.org/abs/1706.07068

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, *19*(1), 17. https://doi.org/10.1007/s40979-023-00140-5

Elyoseph, Z., & Levkovich, I. (2023). Beyond human expertise: The promise and limitations of ChatGPT in suicide risk assessment. *Frontiers in Psychiatry*, *14*, 1213141. https://doi.org/10.3389/fpsyt.2023.1213141

Epley, N., Akalis, S., Waytz, A., & Cacioppo, J. T. (2008). Creating social connection through inferential reproduction: Loneliness and perceived agency in gadgets, gods, and greyhounds. *Psychological Science*, *19*(2), 114–120. https://doi.org/10.1111/j.1467-9280.2008.02056.x

Epley, N., Waytz, A., Akalis, S., & Cacioppo, J. T. (2008). When we need a human: Motivational determinants of anthropomorphism. *Social Cognition*, *26*(2), 143–155. https://doi.org/10.1521/soco.2008.26.2.143

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Esteban-Lozano, I., Castro-González, Á., & Martínez, P. (2024). Using a LLM-based conversational agent in the social robot mini. In H. Degen & S. Ntoa (Eds.), *Artificial Intelligence in HCI* (Vol. 14736, pp. 15–26). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-60615-1_2

European Commission. (2024, November 19). *AI Act | Shaping Europe's digital future*. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

European Commission. Joint Research Centre. (2017). *European framework for the digital competence of educators: DigCompEdu.* Publications Office. https://data.europa.eu/doi/10.2760/159770

Eyssel, F., & Kuchenbrandt, D. (2011). Manipulating anthropomorphic inferences about NAO: The role of situational and dispositional aspects of effectance motivation. *2011 RO-MAN*, 467–472. https://doi.org/10.1109/ROMAN.2011.6005233

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*, 61–68. https://doi.org/10.1145/1957656.1957673

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. *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, 125–126. https://doi.org/10.1145/2157689.2157717

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)*, 121–122. https://doi.org/10.1109/HRI.2013.6483531

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*(2), 117–140. https://doi.org/10.1177/001872675400700202

Firmin, M. W., & Genesi, D. J. (2013). History and implementation of classroom technology. *Procedia - Social and Behavioral Sciences*, *93*, 1603–1617. https://doi.org/10.1016/j.sbspro.2013.10.089

Folk, D. P., Wu, C., & Heine, S. (2023). *Cultural variation in attitudes towards social chatbots*. https://doi.org/10.31234/osf.io/wc895

Forby, L., Anderson, N. C., Cheng, J. T., Foulsham, T., Karstadt, B., Dawson, J., Pazhoohi, F., & Kingstone, A. (2023). Reading the room: Autistic traits, gaze behaviour, and the ability to

infer social relationships. *PLOS ONE*, *18*(3), e0282310. https://doi.org/10.1371/journal.pone.0282310

Fortuna, P., & Modliński, A. (2021). A(I)rtist or counterfeiter? Artificial intelligence as (D)evaluating factor on the art market. *The Journal of Arts Management, Law, and Society*, *51*(3), 188–201. https://doi.org/10.1080/10632921.2021.1887032

Frith, C. D., & Corcoran, R. (1996). Exploring 'theory of mind' in people with schizophrenia. *Psychological Medicine*, *26*(3), 521–530. https://doi.org/10.1017/S0033291700035601

Gabajiwala, E., Mehta, P., Singh, R., & Koshy, R. (2022). Quiz Maker: Automatic quiz generation from text using NLP. In P. K. Singh, S. T. Wierzchoń, J. K. Chhabra, & S. Tanwar (Eds.), *Futuristic Trends in Networks and Computing Technologies* (Vol. 936, pp. 523–533). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-5037-7_37

Gangadharbatla, H. (2022). The role of AI attribution knowledge in the evaluation of artwork. *Empirical Studies of the Arts*, *40*(2), 125–142. https://doi.org/10.1177/0276237421994697

Garety, P. A., & Freeman, D. (1999). Cognitive approaches to delusions: A critical review of theories and evidence. *British Journal of Clinical Psychology*, *38*(2), 113–154. https://doi.org/10.1348/014466599162700

Gazzaniga, M. S. (2008). *Human: The science behind what makes us unique*. Ecco.

Gazzola, V., Rizzolatti, G., Wicker, B., & Keysers, C. (2007). The anthropomorphic brain: The mirror neuron system responds to human and robotic actions. *NeuroImage*, *35*(4), 1674–1684. https://doi.org/10.1016/j.neuroimage.2007.02.003

Gerber, J. P., Wheeler, L., & Suls, J. (2018). A social comparison theory meta-analysis 60+ years on. *Psychological Bulletin*, *144*(2), 177–197. https://doi.org/10.1037/bul0000127

Gernsbacher, M. A., & Yergeau, M. (2019). Empirical failures of the claim that autistic people lack a theory of mind. *Archives of Scientific Psychology*, *7*(1), 102–118. https://doi.org/10.1037/arc0000067

Gerst-Emerson, K., & Jayawardhana, J. (2015). Loneliness as a public health issue: The impact of loneliness on health care utilization among older adults. *American Journal of Public Health*, *105*(5), 1013–1019. https://doi.org/10.2105/AJPH.2014.302427

Gervais, W. M. (2013). Perceiving minds and gods: How mind perception enables, constrains, and is triggered by belief in gods. *Perspectives on Psychological Science*, *8*(4), 380–394. https://doi.org/10.1177/1745691613489836

Goddu, M. K., Noë, A., & Thompson, E. (2024). LLMs don't know anything: Reply to Yildirim and Paul. *Trends in Cognitive Sciences*, S1364661324001682. https://doi.org/10.1016/j.tics.2024.06.008

Gómez-Robles, A., Hopkins, W. D., Schapiro, S. J., & Sherwood, C. C. (2015). Relaxed genetic control of cortical organization in human brains compared with chimpanzees. *Proceedings of the National Academy of Sciences*, *112*(48), 14799–14804. https://doi.org/10.1073/pnas.1512646112

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Government of Canada. (2024, November 12). *Canadian Artificial Intelligence safety institute*. Innovation, Science and Economic Development Canada. https://ised-isde.canada.ca/site/ised/en/canadian-artificial-intelligence-safety-institute

Govier, T. (1994). Is it a jungle out there? Trust, distrust and the construction of social reality. *Dialogue*, *33*(2), 237–252. https://doi.org/10.1017/S0012217300010519

Grant, N. (2022, July 23). Google fires engineer who claims its A.I. is conscious. *The New York Times*. https://www.nytimes.com/2022/07/23/technology/google-engineer-artificial-intelligence.html

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619–619. https://doi.org/10.1126/science.1134475

Gray, K., Jenkins, A. C., Heberlein, A. S., & Wegner, D. M. (2011). Distortions of mind perception in psychopathology. *Proceedings of the National Academy of Sciences*, *108*(2), 477–479. https://doi.org/10.1073/pnas.1015493108

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, *125*(1), 125–130. https://doi.org/10.1016/j.cognition.2012.06.007

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*(2), 101–124. https://doi.org/10.1080/1047840X.2012.651387

Guilford, J. P. (1967). Creativity: Yesterday, today and tomorrow. *The Journal of Creative Behavior*, *1*(1), 3–14. https://doi.org/10.1002/j.2162-6057.1967.tb00002.x

Guzik, E. E., Byrge, C., & Gilde, C. (2023). The originality of machines: AI takes the Torrance Test. *Journal of Creativity*, *33*(3), 100065. https://doi.org/10.1016/j.yjoc.2023.100065

Hadi, M. U., Tashi, Q. A., Qureshi, R., Shah, A., Muneer, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., & Mirjalili, S. (2023). *A survey on large language models: Applications, challenges, limitations, and practical usage*. https://doi.org/10.36227/techrxiv.23589741.v1

Haluza, D., & Jungwirth, D. (2023). Artificial intelligence and ten societal megatrends: An exploratory study using GPT-3. *Systems*, *11*(3), 120. https://doi.org/10.3390/systems11030120

Hardin, E. E., & Lakin, J. L. (2009). The integrated self-discrepancy index: A reliable and valid measure of self-discrepancies. *Journal of Personality Assessment*, *91*(3), 245–253. https://doi.org/10.1080/00223890902794291

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review*, *10*(3), 252–264. https://doi.org/10.1207/s15327957pspr1003_4

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400–407. https://doi.org/10.3758/s13428-015-0578-z

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*(2), 243. https://doi.org/10.2307/1416950

Heinrich, L. M., & Gullone, E. (2006). The clinical significance of loneliness: A literature review. *Clinical Psychology Review*, *26*(6), 695–718. https://doi.org/10.1016/j.cpr.2006.04.002

Heller, B., Proctor, M., Mah, D., Jewell, L., & Cheung, B. (2005). *Freudbot: An investigation of chatbot technology in distance education*. 3913–3918. https://www.learntechlib.org/primary/p/20691/

Higgins, E. (1987). Self-discrepancy: A theory relating self and affect. *Psychological Review*, *94*, 319–340. https://doi.org/10.1037/0033-295X.94.3.319

Higgins, E. T., Bond, R. N., Klein, R., & Strauman, T. (1986). Self-discrepancies and emotional vulnerability: How magnitude, accessibility, and type of discrepancy influence affect. *Journal of Personality and Social Psychology*, *51*(1), 5–15. https://doi.org/10.1037/0022-3514.51.1.5

Higgins, E. T., Klein, R., & Strauman, T. (1985). Self-concept discrepancy theory: A psychological model for distinguishing among different aspects of depression and anxiety. *Social Cognition*, *3*(1), 51–76. https://doi.org/10.1521/soco.1985.3.1.51

Hindennach, S., Shi, L., MiletiĆ, F., & Bulling, A. (2024). Mindful explanations: Prevalence and impact of mind attribution in XAI research. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1–43. https://doi.org/10.1145/3641009

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2023). Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science*, *5*, 1096257. https://doi.org/10.3389/fcomp.2023.1096257

Hong, J.-W., & Curran, N. M. (2019). Artificial intelligence, artists, and art: Attitudes toward artwork produced by humans vs. Artificial intelligence. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *15*(2s), 1–16. https://doi.org/10.1145/3326337

Hu, K. (2023, February 2). ChatGPT sets record for fastest-growing user base—Analyst note. *Reuters*. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Huebner, B. (2010). Commonsense concepts of phenomenal consciousness: Does anyone care about functional zombies? *Phenomenology and the Cognitive Sciences*, *9*(1), 133–155. https://doi.org/10.1007/s11097-009-9126-6

Hughes, M. E., Waite, L. J., Hawkley, L. C., & Cacioppo, J. T. (2004). A short scale for measuring loneliness in large surveys: Results from two population-based studies. *Research on Aging*, *26*(6), 655–672. https://doi.org/10.1177/0164027504268574

Hume, D., & Bell, J. M. (1990). *Dialogues concerning natural religion* (J. M. Bell, Ed.). Penguin Books.

Hyun Baek, T., & Kim, M. (2023). Is ChatGPT scary good? How user motivations affect creepiness and trust in generative artificial intelligence. *Telematics and Informatics*, *83*, 102030. https://doi.org/10.1016/j.tele.2023.102030

Inie, N., Druga, S., Zukerman, P., & Bender, E. M. (2024). From "AI" to probabilistic automation: How does anthropomorphization of technical systems descriptions influence trust? *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2322–2347. https://doi.org/10.1145/3630106.3659040

Introducing ChatGPT. (2022). https://openai.com/index/chatgpt/

Introducing Claude 3.5 Sonnet \ Anthropic. (2024). https://www.anthropic.com/news/claude-3-5-sonnet

Jacobs, O. L., Gazzaz, K., & Kingstone, A. (2022). Mind the robot! Variation in attributions of mind to a wide set of real and fictional robots. *International Journal of Social Robotics*, *14*(2), 529–537. https://doi.org/10.1007/s12369-021-00807-4

Jacobs, O. L., Pazhoohi, F., & Kingstone, A. (2023). Self-discrepancies in mind perception for actual, ideal, and ought selves and partners. *PLOS ONE*, *18*(12), e0295515. https://doi.org/10.1371/journal.pone.0295515

Jacobs, O. L., Pazhoohi, F., & Kingstone, A. (2024). Large language models have divergent effects on self-perceptions of mind and the attributes considered uniquely human. *Consciousness and Cognition*, *124*, 103733. https://doi.org/10.1016/j.concog.2024.103733

Jacobs, O., Pazhoohi, F., & Kingstone, A. (2023). *Brief exposure increases mind perception to ChatGPT and is moderated by the individual propensity to anthropomorphize*. https://doi.org/10.31234/osf.io/pn29d

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

Jiang, H., Zhang, X., Cao, X., Breazeal, C., Roy, D., & Kabbara, J. (2023). *PersonaLLM: Investigating the ability of large language models to express personality traits* (Version 5). arXiv. https://doi.org/10.48550/ARXIV.2305.02547

Jorgensen, T. (2022). *Is the human brain a biological computer? | Princeton University Press*. https://press.princeton.edu/ideas/is-the-human-brain-a-biological-computer

Kamide, H., Eyssel, F., & Arai, T. (2013). Psychological anthropomorphism of robots. In G. Herrmann, M. J. Pearson, A. Lenz, P. Bremner, A. Spiers, & U. Leonards (Eds.), *Social Robotics* (Vol. 8239, pp. 199–208). Springer International Publishing. https://doi.org/10.1007/978-3-319-02675-6_20

Karinshak, E., Liu, S. X., Park, J. S., & Hancock, J. T. (2023). Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 1–29. https://doi.org/10.1145/3579592

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kaufman, J. C., & Kaufman, A. B. (2004). Applying a creativity framework to animal cognition. *New Ideas in Psychology*, *22*(2), 143–155. https://doi.org/10.1016/j.newideapsych.2004.09.006

Kay, A. C., Whitson, J. A., Gaucher, D., & Galinsky, A. D. (2009). Compensatory control: Achieving order through the mind, our institutions, and the heavens. *Current Directions in Psychological Science*, *18*(5), 264–268. https://doi.org/10.1111/j.1467-8721.2009.01649.x

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. https://doi.org/10.1521/soco.2008.26.2.169

Kim, J. (2018). *Philosophy of mind* (Third edition). Routledge.

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

Koch, K. (2021). Purposiveness in nature: Hegel and Spinoza on anthropomorphism and backward causation. *Intellectual History Review*, *31*(3), 463–478. https://doi.org/10.1080/17496977.2021.1956043

Koch, S. (Ed.). (1959). *Psychology: A study of a science.* McGraw-Hill.

Köhler, W. (1925). An aspect of gestalt psychology. *The Pedagogical Seminary and Journal of Genetic Psychology*, *32*(4), 691–723. https://doi.org/10.1080/08856559.1925.9944846

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, *121*(45), e2405460121. https://doi.org/10.1073/pnas.2405460121

Kreitchmann, R. S., Abad, F. J., Ponsoda, V., Nieto, M. D., & Morillo, D. (2019). Controlling for response biases in self-report scales: Forced-choice vs. psychometric modeling of likert items. *Frontiers in Psychology*, *10*, 2309. https://doi.org/10.3389/fpsyg.2019.02309

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 25). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Kruger, J., Wirtz, D., Van Boven, L., & Altermatt, T. W. (2004). The effort heuristic. *Journal of Experimental Social Psychology*, *40*(1), 91–98. https://doi.org/10.1016/S0022-1031(03)00065-9

Laban, G., George, J.-N., Morrison, V., & Cross, E. S. (2020). Tell me more! Assessing interactions with social robots from speech. *Paladyn, Journal of Behavioral Robotics*, *12*(1), 136–159. https://doi.org/10.1515/pjbr-2021-0011

Laban, G., Kappas, A., Morrison, V., & Cross, E. S. (2024). Building long-term human–robot relationships: Examining disclosure, perception and well-being across time. *International Journal of Social Robotics*, *16*(5), 1–27. https://doi.org/10.1007/s12369-023-01076-z

Laban, G., Laban, T., & Gunes, H. (2024a). *LEXI: Large Language Models experimentation interface* (No. arXiv:2407.01488). arXiv. http://arxiv.org/abs/2407.01488

Laban, G., Laban, T., & Gunes, H. (2024b). *LEXI: Large Language Models Experimentation Interface* (No. arXiv:2407.01488). arXiv. http://arxiv.org/abs/2407.01488

Lakatos, G., Gácsi, M., Konok, V., Brúder, I., Bereczky, B., Korondi, P., & Miklósi, Á. (2014). Emotion attribution to a non-humanoid robot in different social situations. *PLoS ONE*, *9*(12), e114207. https://doi.org/10.1371/journal.pone.0114207

Langer, M., Hunsicker, T., Feldkamp, T., König, C. J., & Grgić-Hlača, N. (2022). "Look! It's a computer program! It's an algorithm! It's AI!": Does terminology affect human perceptions and evaluations of algorithmic decision-making systems? *CHI Conference on Human Factors in Computing Systems*, 1–28. https://doi.org/10.1145/3491102.3517527

Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A. Y. S., & Coiera, E. (2018). Conversational agents in healthcare: A systematic review. *Journal of the American Medical Informatics Association*, *25*(9), 1248–1258. https://doi.org/10.1093/jamia/ocy072

Lawrence, M. (2009). *ez: Easy analysis and visualization of factorial experiments* (p. 4.4-0) [Dataset]. https://doi.org/10.32614/CRAN.package.ez

Lee, C.-Y. S., & Goldstein, S. E. (2016). Loneliness, stress, and social support in young adulthood: Does the source of support matter? *Journal of Youth and Adolescence*, *45*(3), 568–580. https://doi.org/10.1007/s10964-015-0395-9

Lee, K. M., Jung, Y., Kim, J., & Kim, S. R. (2006). Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. *International Journal of Human-Computer Studies*, *64*(10), 962–973. https://doi.org/10.1016/j.ijhcs.2006.05.002

Lee, S., Lee, N., & Sah, Y. J. (2020). Perceiving a mind in a chatbot: Effect of mind perception and social cues on co-presence, closeness, and intention to use. *International Journal of*

Human–Computer Interaction, *36*(10), 930–940. https://doi.org/10.1080/10447318.2019.1699748

Leppink, J., O'Sullivan, P., & Winston, K. (2017). Evidence against vs. In favour of a null hypothesis. *Perspectives on Medical Education*, *6*(2), 115–118. https://doi.org/10.1007/S40037-017-0332-6

Lesher, J. (2023). Xenophanes. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Summer 2023). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2023/entries/xenophanes/

Lewicki, R. J., McAllister, D. J., & Bies, R. J. (1998). Trust and distrust: New relationships and realities. *The Academy of Management Review*, *23*(3), 438. https://doi.org/10.2307/259288

Lissitz, R. W., & Willhoft, J. L. (1985). A methodological study of the torrance tests of creativity. *Journal of Educational Measurement*, *22*(1), 1–11. https://doi.org/10.1111/j.1745-3984.1985.tb01044.x

Llama 3.2. (n.d.). Meta Llama. Retrieved October 31, 2024, from https://www.llama.com/

Loconte, R., Orrù, G., Tribastone, M., Pietrini, P., & Sartori, G. (2023). *Challenging ChatGPT "Intelligence" with human tools: A neuropsychological investigation on prefrontal functioning of a large language model*. SSRN. https://doi.org/10.2139/ssrn.4377371

Looser, C. E., & Wheatley, T. (2010). The tipping point of animacy: How, when, and where we perceive life in a face. *Psychological Science*, *21*(12), 1854–1862. https://doi.org/10.1177/0956797610388044

MacNeil, S., Tran, A., Mogil, D., Bernstein, S., Ross, E., & Huang, Z. (2022). Generating diverse code explanations using the GPT-3 large language model. *Proceedings of the 2022 ACM Conference on International Computing Education Research - Volume 2*, 37–39. https://doi.org/10.1145/3501709.3544280

Malle, B. F. (2019). How many dimensions of mind perception really are there? *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, 2268–2274.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 117–124. https://doi.org/10.1145/2696454.2696458

Mann, F., Wang, J., Pearce, E., Ma, R., Schlief, M., Lloyd-Evans, B., Ikhtabi, S., & Johnson, S. (2022). Loneliness and the onset of new mental health problems in the general population. *Social Psychiatry and Psychiatric Epidemiology*, *57*(11), 2161–2178. https://doi.org/10.1007/s00127-022-02261-7

Manzi, F., Massaro, D., Di Lernia, D., Maggioni, M. A., Riva, G., & Marchetti, A. (2021). Robots are not all the same: Young adults' expectations, attitudes, and mental attribution to two humanoid social robots. *Cyberpsychology, Behavior, and Social Networking*, *24*(5), 307–314. https://doi.org/10.1089/cyber.2020.0162

Mauldin, M. L. (1994). ChatterBots, TinyMuds, and the Turing Test: Entering the Loebner Prize Competition. *AAAI*, *94*, 16–21.

McClelland, H., Evans, J. J., Nowland, R., Ferguson, E., & O'Connor, R. C. (2020). Loneliness as a predictor of suicidal ideation and behaviour: A systematic review and meta-analysis

of prospective studies. *Journal of Affective Disorders*, *274*, 880–896. https://doi.org/10.1016/j.jad.2020.05.004

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

McCrae, R., & Costa, P. (1996). *The five factor model of personality: Theoretical Perspective*. Guilford Press.

McNeill, W. E. S. (2012). On seeing that someone is angry. *European Journal of Philosophy*, *20*(4), 575–597. https://doi.org/10.1111/j.1468-0378.2010.00421.x

McQuate, S. (2023, July 27.). Q&A: UW researcher discusses just how much energy ChatGPT uses. *UW News*. Retrieved November 24, 2024, from https://www.washington.edu/news/2023/07/27/how-much-energy-does-chatgpt-use/

Metz, C. (2023, May 1). 'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead. *The New York Times*. https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html

Miller, G. A. (2003). The cognitive revolution: A historical perspective. *Trends in Cognitive Sciences*, *7*(3), 141–144. https://doi.org/10.1016/S1364-6613(03)00029-9

Miner, A. S., Milstein, A., Schueller, S., Hegde, R., Mangurian, C., & Linos, E. (2016). Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA Internal Medicine*, *176*(5), 619. https://doi.org/10.1001/jamainternmed.2016.0400

Mintz, L. B., & Betz, N. E. (1986). Sex differences in the nature, realism, and correlates of body image. *Sex Roles*, *15*(3–4), 185–195. https://doi.org/10.1007/BF00287483

Miotto, M., Rossberg, N., & Kleinberg, B. (2022). *Who is GPT-3? An exploration of personality, values and demographics* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2209.14338

Mishra, R., Welch, K. C., & Popa, D. O. (2024). *Human-mediated large language models for robotic intervention in children with autism spectrum disorders* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2402.00260

Molnár, Z., Clowry, G. J., Šestan, N., Alzu'bi, A., Bakken, T., Hevner, R. F., Hüppi, P. S., Kostović, I., Rakic, P., Anton, E. S., Edwards, D., Garcez, P., Hoerder-Suabedissen, A., & Kriegstein, A. (2019). New insights into the development of the human cerebral cortex. *Journal of Anatomy*, *235*(3), 432–451. https://doi.org/10.1111/joa.13055

Moore, J. W. (2016). What is the sense of agency and why does it matter? *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01272

Moskowitz, G. B., & Gill, M. J. (2013). *Person perception*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195376746.013.0058

Müller, B. C. N., Gao, X., Nijssen, S. R. R., & Damen, T. G. E. (2021). I, robot: How human appearance and mind attribution relate to the perceived danger of robots. *International Journal of Social Robotics*, *13*(4), 691–701. https://doi.org/10.1007/s12369-020-00663-8

Nagel, T. (1980). What is it like to be a bat? In N. Block (Ed.), *The Language and Thought Series*. Harvard University Press. https://doi.org/10.4159/harvard.9780674594623.c15

Newell, A. (1980). Physical symbol systems*. *Cognitive Science*, *4*(2), 135–183. https://doi.org/10.1207/s15516709cog0402_2

Nilsson, N. J. (2007). The physical symbol system hypothesis: Status and prospects. In M. Lungarella, F. Iida, J. Bongard, & R. Pfeifer (Eds.), *50 Years of Artificial Intelligence* (Vol. 4850, pp. 9–17). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-77296-5_2

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F., & Turkheimer, E. (2012). Intelligence: New findings and theoretical developments. *American Psychologist*, *67*(2), 130–159. https://doi.org/10.1037/a0026699

Nuijten, M. B., Hartgerink, C. H. J., Van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*(4), 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

Oh, C., Song, J., Choi, J., Kim, S., Lee, S., & Suh, B. (2018). I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3173574.3174223

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, *24*(5), 939–973. https://doi.org/10.1017/S0140525X01000115

Orrù, G., Piarulli, A., Conversano, C., & Gemignani, A. (2023). Human-like problem-solving abilities in large language models using ChatGPT. *Frontiers in Artificial Intelligence*, *6*, 1199350. https://doi.org/10.3389/frai.2023.1199350

Ozgul, S., Heubeck, B., Ward, J., & Wilkinson, R. (2003). Self-discrepancies: Measurement and relation to various negative affective states. *Australian Journal of Psychology*, *55*(1), 56–62. https://doi.org/10.1080/00049530412331312884

Pak, R., Crumley-Branyon, J. J., De Visser, E. J., & Rovira, E. (2020). Factors that affect younger and older adults' causal attributions of robot behaviour. *Ergonomics*, *63*(4), 421–439. https://doi.org/10.1080/00140139.2020.1734242

Pan, K., & Zeng, Y. (2023). *Do LLMs possess a personality? Making the MBTI test an amazing evaluation for large language models* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2307.16180

Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology*, *5*(2), 125–137. https://doi.org/10.1111/j.2044-835X.1987.tb01048.x

Perrig, S. A. C., Scharowski, N., & Brühlmann, F. (2023). Trust issues with trust scales: Examining the psychometric quality of trust measures in the context of AI. *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–7. https://doi.org/10.1145/3544549.3585808

Phillips, A. G., & Silvia, P. J. (2010). Individual differences in self-discrepancies and emotional experience: Do distinct discrepancies predict distinct emotions? *Personality and Individual Differences*, *49*(2), 148–151. https://doi.org/10.1016/j.paid.2010.03.010

Phillips, E., Zhao, X., Ullman, D., & Malle, B. F. (2018). What is human-like?: Decomposing robots' human-like appearance using the anthropomorphic roBOT (ABOT) database. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 105–113. https://doi.org/10.1145/3171221.3171268

Picard, R. W. (2000). *Affective computing* (1. paperback ed). MIT Press.

Pinar Saygin, A., Cicekli, I., & Akman, V. (2000). Turing Test: 50 years later. *Minds and Machines*, *10*(4), 463–518. https://doi.org/10.1023/A:1011288000451

Porsdam Mann, S., Vazirani, A. A., Aboy, M., Earp, B. D., Minssen, T., Cohen, I. G., & Savulescu, J. (2024). Guidelines for ethical use and acknowledgement of large language models in academic writing. *Nature Machine Intelligence*. https://doi.org/10.1038/s42256-024-00922-7

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

Preus, J. S. (1995). Anthropomorphism and Spinoza's innovations. *Religion*, *25*(1), 1–8. https://doi.org/10.1006/reli.1995.0001

Ragot, M., Martin, N., & Cojean, S. (2020). Ai-generated vs. Human artworks. A perception bias towards artificial intelligence? *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–10. https://doi.org/10.1145/3334480.3382892

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. ‘Sandy,’ … Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477–486. https://doi.org/10.1038/s41586-019-1138-y

Rao, H., Leung, C., & Miao, C. (2023). *Can ChatGPT assess human personalities? A general evaluation framework* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2303.01248

Reich, N., & Eyssel, F. (2013). Attitudes towards service robots in domestic environments: The role of personality characteristics, individual interests, and demographic variables. *Paladyn, Journal of Behavioral Robotics*, *4*(2). https://doi.org/10.2478/pjbr-2013-0014

Ritter, S., Barrett, D. G. T., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 2940–2949). PMLR. https://proceedings.mlr.press/v70/ritter17a.html

Roberts, B. W., & Yoon, H. J. (2022). Personality psychology. In *Annual Review of Psychology* (Vol. 73, Issue Volume 73, 2022, pp. 489–516). Annual Reviews. https://doi.org/10.1146/annurev-psych-020821-114927

Rossi, S., Conti, D., Garramone, F., Santangelo, G., Staffa, M., Varrasi, S., & Di Nuovo, A. (2020). The role of personality factors and empathy in the acceptance and performance of a social robot for psychometric evaluations. *Robotics*, *9*(2), 39. https://doi.org/10.3390/robotics9020039

Rothman, J. (2023, November 13). Why the godfather of A.I. fears what he’s built. *The New Yorker*. https://www.newyorker.com/magazine/2023/11/20/geoffrey-hinton-profile-ai

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal*, *24*(1), 92–96. https://doi.org/10.1080/10400419.2012.650092

Rutinowski, J., Franke, S., Endendyk, J., Dormuth, I., & Pauly, M. (2023). *The self-perception and political biases of ChatGPT* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2304.07333

Ruzich, E., Allison, C., Smith, P., Watson, P., Auyeung, B., Ring, H., & Baron-Cohen, S. (2015). Measuring autistic traits in the general population: A systematic review of the Autism-Spectrum Quotient (AQ) in a nonclinical population sample of 6,900 typical adult males and females. *Molecular Autism*, *6*(1), 2. https://doi.org/10.1186/2040-2392-6-2

Ryan, R. M., & Deci, E. L. (2006). Self-Regulation and the Problem of Human Autonomy: Does Psychology Need Choice, Self-Determination, and Will? *Journal of Personality*, *74*(6), 1557–1586. https://doi.org/10.1111/j.1467-6494.2006.00420.x

Sacino, A., Cocchella, F., De Vita, G., Bracco, F., Rea, F., Sciutti, A., & Andrighetto, L. (2022). Human- or object-like? Cognitive anthropomorphism of humanoid robots. *PLOS ONE*, *17*(7), e0270787. https://doi.org/10.1371/journal.pone.0270787

Salah, M., Al Halbusi, H., & Abdelfattah, F. (2023). May the force of text data analysis be with you: Unleashing the power of generative AI for social psychology research. *Computers in Human Behavior: Artificial Humans*, *1*(2), 100006. https://doi.org/10.1016/j.chbah.2023.100006

Sallam, M., Salim, N. A., Al-Tammemi, A. B., Barakat, M., Fayyad, D., Hallit, S., Harapan, H., Hallit, R., & Mahafzah, A. (2023). ChatGPT output regarding compulsory vaccination

and COVID-19 vaccine conspiracy: A descriptive study at the outset of a paradigm shift in online search for information. *Cureus*. https://doi.org/10.7759/cureus.35029

Schank, R. C., & Colby, K. M. (1973). Computer models of thought and language. In *Computer models of thought and language*. W. H. Freeman.

Schurr, A., & Moran, S. (2023). The presence of automation enhances deontological considerations in moral judgments. *Computers in Human Behavior*, *140*, 107590. https://doi.org/10.1016/j.chb.2022.107590

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., Abdulhai, M., Faust, A., & Matarić, M. (2023). *Personality traits in large language models* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2307.00184

Shahsavar, Y., & Choudhury, A. (2023). User intentions to use ChatGPT for self-diagnosis and health-related purposes: Cross-sectional survey study. *JMIR Human Factors*, *10*, e47564. https://doi.org/10.2196/47564

Shank, D. B., Graves, C., Gott, A., Gamez, P., & Rodriguez, S. (2019). Feeling our way to machine minds: People's emotions when perceiving mind in artificial intelligence. *Computers in Human Behavior*, *98*, 256–266. https://doi.org/10.1016/j.chb.2019.04.001

Sheehan, B., Jin, H. S., & Gottlieb, U. (2020). Customer service chatbots: Anthropomorphism and adoption. *Journal of Business Research*, *115*, 14–24. https://doi.org/10.1016/j.jbusres.2020.04.030

Shen, X., Chen, Z., Backes, M., & Zhang, Y. (2023). *In ChatGPT we trust? Measuring and characterizing the reliability of ChatGPT* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2304.08979

Shiffrin, R., & Mitchell, M. (2023). Probing the psychology of AI models. *Proceedings of the National Academy of Sciences*, *120*(10), e2300963120. https://doi.org/10.1073/pnas.2300963120

Shin, H. I., & Kim, J. (2020). My computer is more thoughtful than you: Loneliness, anthropomorphism and dehumanization. *Current Psychology*, *39*(2), 445–453. https://doi.org/10.1007/s12144-018-9975-7

Shravya Bhat, Nguyen, H., Moore, S., Stamper, J., Sakr, M., & Nyberg, E. (2022). Towards automated generation and evaluation of questions in educational domains. *Proceedings of the 15th International Conference on Educational Data Mining*, 701–704. https://doi.org/10.5281/ZENODO.6853084

Shum, H., He, X., & Li, D. (2018). From Eliza to XiaoIce: Challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, *19*(1), 10–26. https://doi.org/10.1631/FITEE.1700826

Snapper, L., Oranç, C., Hawley-Dolan, A., Nissel, J., & Winner, E. (2015). Your kid could not have done that: Even untutored observers can discern intentionality and structure in abstract expressionist art. *Cognition*, *137*, 154–165. https://doi.org/10.1016/j.cognition.2014.12.009

Song, X. (2023). Energy metabolism and brain functions. *Harvard Brain Science Initiative*. https://brain.harvard.edu/hbi_news/energy-metabolism-and-brain-functions/

Spatola, N., & Wudarczyk, O. A. (2021). Ascribing emotions to robots: Explicit and implicit attribution of emotions and perceived robot anthropomorphism. *Computers in Human Behavior*, *124*, 106934. https://doi.org/10.1016/j.chb.2021.106934

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., … Wu, Z. (2022). *Beyond the imitation game: Quantifying and extrapolating the capabilities of language models*. https://doi.org/10.48550/ARXIV.2206.04615

Stadthagen-González, H., López, L., Parafita Couto, M. C., & Párraga, C. A. (2018). Using two-alternative forced choice tasks and Thurstone's law of comparative judgments for code-switching research. *Linguistic Approaches to Bilingualism*, *8*(1), 67–97. https://doi.org/10.1075/lab.16030.sta

Stafford, R. Q., MacDonald, B. A., Jayawardena, C., Wegner, D. M., & Broadbent, E. (2014). Does the robot have a mind? Mind perception and attitudes towards robots predict use of an eldercare robot. *International Journal of Social Robotics*, *6*(1), 17–32. https://doi.org/10.1007/s12369-013-0186-y

Stevenson, C., Smal, I., Baas, M., Grasman, R., & van der Maas, H. (2022). *Putting GPT-3's creativity to the (Alternative Uses Test)* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2206.08932

Stuart, M. T., & Kneer, M. (2021). Guilty artificial minds: Folk attributions of mens rea and culpability to artificially intelligent agents. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–27. https://doi.org/10.1145/3479507

Syrdal, D. S., Dautenhahn, K., Koay, K. L., & Walters, M. L. (2020). The Negative Attitudes Towards Robots scale and reactions to robot behaviour in a live human-robot interaction study. *International Journal of Social Robotics*, *13*, 691–701. https://doi.org/10.1007/s12369-020-00663-8

Tack, A., & Piech, C. (2022). *The AI teacher test: Measuring the pedagogical ability of blender and GPT-3 in educational dialogues* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2205.07540

Tamir, D. I., & Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive Sciences*, *22*(3), 201–212. https://doi.org/10.1016/j.tics.2017.12.005

Tang, L., Sun, Z., Idnay, B., Nestor, J. G., Soroush, A., Elias, P. A., Xu, Z., Ding, Y., Durrett, G., Rousseau, J. F., Weng, C., & Peng, Y. (2023). Evaluating large language models on medical evidence summarization. *Npj Digital Medicine*, *6*(1), 158. https://doi.org/10.1038/s41746-023-00896-7

Tangney, J. P., Niedenthal, P. M., Covert, M. V., & Barlow, D. H. (1998). Are shame and guilt related to distinct self-discrepancies? A test of Higgins's (1987) hypotheses. *Journal of Personality and Social Psychology*, *75*(1), 256–268. https://doi.org/10.1037/0022-3514.75.1.256

Tharp, M., Holtzman, N. S., & Eadeh, F. R. (2017). Mind perception and individual differences: A replication and extension. *Basic and Applied Social Psychology*, *39*(1), 68–73. https://doi.org/10.1080/01973533.2016.1256287

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*, *11*(4), 1–51. https://doi.org/10.1145/3526112

Thellman, S., Silvervarg, A., & Ziemke, T. (2017). Folk-psychological interpretation of human vs. Humanoid robot behavior: Exploring the intentional stance toward robots. *Frontiers in Psychology*, *8*, 1962. https://doi.org/10.3389/fpsyg.2017.01962

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, *29*(8), 1930–1940. https://doi.org/10.1038/s41591-023-02448-8

Tian, E. (2024). *AI Detector—The Original AI Checker for ChatGPT & More*. GPTZero. https://gptzero.me/

Torchiano, M. (2013). *effsize: Efficient effect size computation* (p. 0.8.1) [Dataset]. https://doi.org/10.32614/CRAN.package.effsize

Trott, S., Jones, C., Chang, T., Michaelov, J., & Bergen, B. (2023). Do Large Language Models know what humans know? *Cognitive Science*, *47*(7), e13309. https://doi.org/10.1111/cogs.13309

Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. Basic Books.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315. https://doi.org/10.1037/0033-295X.90.4.293

Tzelios, K., Williams, L. A., Omerod, J., & Bliss-Moreau, E. (2022). Evidence of the unidimensional structure of mind perception. *Scientific Reports*, *12*(1), 18978. https://doi.org/10.1038/s41598-022-23047-6

UK Government. (2022). *Introducing the AI safety institute*. GOV.UK. https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute

Van Dis, E. A. M., Bollen, J., Zuidema, W., Van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. *Nature*, *614*(7947), 224–226. https://doi.org/10.1038/d41586-023-00288-7

Van Noorden, R., & Perkel, J. M. (2023). AI and science: What 1,600 researchers think. *Nature*, *621*(7980), 672–675. https://doi.org/10.1038/d41586-023-02980-0

Varela, F. J., Thompson, E., & Rosch, E. (1993). *The embodied mind: Cognitive science and human experience*. MIT press.

Walter, K. V., Conroy-Beam, D., Buss, D. M., Asao, K., Sorokowska, A., Sorokowski, P., Aavik, T., Akello, G., Alhabahba, M. M., Alm, C., Amjad, N., Anjum, A., Atama, C. S., Atamtürk Duyar, D., Ayebare, R., Batres, C., Bendixen, M., Bensafia, A., Bizumic, B., … Zupančič, M. (2020). Sex differences in mate preferences across 45 countries: A large-scale replication. *Psychological Science*, *31*(4), 408–423. https://doi.org/10.1177/0956797620904154

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. (2024). A survey on large language model based

autonomous agents. *Frontiers of Computer Science*, *18*(6), 186345. https://doi.org/10.1007/s11704-024-40231-1

Wang, X., & Krumhuber, E. G. (2018). Mind perception of robots varies with their economic versus social function. *Frontiers in Psychology*, *9*, 1230. https://doi.org/10.3389/fpsyg.2018.01230

Watrin, L., Geiger, M., Spengler, M., & Wilhelm, O. (2019). Forced-choice versus likert responses on an occupational big five questionnaire. *Journal of Individual Differences*, *40*(3), 134–148. https://doi.org/10.1027/1614-0001/a000285

Watson, N., Bryan, B. C., & Thrash, T. M. (2010). Self-discrepancy: Comparisons of the psychometric properties of three instruments. *Psychological Assessment*, *22*(4), 878–892. https://doi.org/10.1037/a0020644

Watson, N., Bryan, B. C., & Thrash, T. M. (2016). Self-discrepancy: Long-term test–retest reliability and test–criterion predictive validity. *Psychological Assessment*, *28*(1), 59–69. https://doi.org/10.1037/pas0000162

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human?: The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, *5*(3), 219–232. https://doi.org/10.1177/1745691610369336

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences*, *14*(8), 383–388. https://doi.org/10.1016/j.tics.2010.05.006

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117. https://doi.org/10.1016/j.jesp.2014.01.005

Webster, P. (2023). Six ways large language models are changing healthcare. *Nature Medicine*, *29*(12), 2969–2971. https://doi.org/10.1038/s41591-023-02700-1

Wegner, D. M. (2003). *The Illusion of Conscious Will*. MIT Press.

Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences*, *114*(43), 11374–11379. https://doi.org/10.1073/pnas.1704347114

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168

Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J. R. (2016). The stability of extreme response style and acquiescence over 8 years. *Assessment*, *23*(3), 279–291. https://doi.org/10.1177/1073191115583714

Whitson, J. A., & Galinsky, A. D. (2008). Lacking control increases illusory pattern perception. *Science*, *322*(5898), 115–117. https://doi.org/10.1126/science.1159845

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2014). *dplyr: A grammar of data manipulation* (p. 1.1.4) [Dataset]. https://doi.org/10.32614/CRAN.package.dplyr

Wiese, E., Metta, G., & Wykowska, A. (2017). Robots as intentional agents: Using neuroscientific methods to make robots appear more social. *Frontiers in Psychology*, *8*, 1663. https://doi.org/10.3389/fpsyg.2017.01663

Wildt, A. R., & Mazis, M. B. (1978). Determinants of scale response: Label versus position. *Journal of Marketing Research*, *15*(2), 261–267. https://doi.org/10.1177/002224377801500209

Will, P., Merritt, E., Jenkins, R., & Kingstone, A. (2021). The Medusa effect reveals levels of mind perception in pictures. *Proceedings of the National Academy of Sciences*, *118*(32), e2106640118. https://doi.org/10.1073/pnas.2106640118

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(4), 625–636. https://doi.org/10.3758/BF03196322

Wimmer, H. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128. https://doi.org/10.1016/0010-0277(83)90004-5

Wu, Z., Ji, D., Yu, K., Zeng, X., Wu, D., & Shidujaman, M. (2021). AI creativity and the human-AI co-creation model. In M. Kurosu (Ed.), *Human-Computer Interaction. Theory, Methods and Tools* (Vol. 12762, pp. 171–190). Springer International Publishing. https://doi.org/10.1007/978-3-030-78462-1_13

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)*, 472–475. https://doi.org/10.1109/ROMAN.2018.8525586

Ye, Y., You, H., & Du, J. (2023). Improved trust in human-robot collaboration with ChatGPT. *IEEE Access*, *11*, 55748–55754. https://doi.org/10.1109/ACCESS.2023.3282111

Yildirim, G., Elban, M., & Yildirim, S. (2018). Analysis of use of virtual reality technologies in history education: A case study. *Asian Journal of Education and Training*, *4*(2), 62–69. https://doi.org/10.20448/journal.522.2018.42.62.69

Young, J. G. (1985). What is creativity? *The Journal of Creative Behavior*, *19*(2), 77–87. https://doi.org/10.1002/j.2162-6057.1985.tb00640.x

Zagic, D., Rapee, R. M., & Wuthrich, V. M. (2024). A novel experimental approach to identifying the cognitive mechanisms underlying loneliness. *Cognitive Therapy and Research*, *48*(5), 1014–1026. https://doi.org/10.1007/s10608-024-10494-w

Zhang, X., Wu, C., Zhang, Y., Xie, W., & Wang, Y. (2023). Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, *14*(1), 4542. https://doi.org/10.1038/s41467-023-40260-7

Zhang, Z., & Mai, Y. (2018). *WebPower: Basic and advanced statistical power analysis* (p. 0.9.4) [Dataset]. https://doi.org/10.32614/CRAN.package.WebPower

Złotowski, J., Strasser, E., & Bartneck, C. (2014). Dimensions of anthropomorphism: From humanness to humanlikeness. *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, 66–73. https://doi.org/10.1145/2559636.2559679

Appendix

Table S1: Study 2 Experiment 1: Individual difference measures and their correlations with mind perception (agency and experience) ratings.

	Pre Agency	Pre Experience	Post Agency	Post Experience
Age	0.05	-0.21 **	0.07	-0.24 **
Education	0.05	0.25 **	0.03	0.18 *
Prior Exposure	0.17 *	0.38 ***	0.17 *	0.38 ***
AQ-10 Total	-0.01	0.11	-0.01	0.12
IDAQ Total	0.28 ***	0.43 ***	0.25 **	0.48 ***
Openness	0.07	-0.2 *	0.08	-0.18 *
Extraversion	0.09	0.22 **	0.13	0.22 **
Agreeableness	0.27 ***	-0.1	0.21 **	-0.08
Conscientiousness	0.2 *	-0.21 **	0.23 **	-0.12
Emotional Stability	0.13	-0.08	0.14	-0.09

Table S2: Study 2 Experiment 3: Individual difference measures and their correlations with mind perception (agency and experience) ratings.

	Pre Agency	Pre Experience	Post Agency	Post Experience
Age	0.06	-0.04	0.00	0.01
Education	-0.07	-0.03	-0.12 *	-0.02
Prior Exposure	0.18 **	0.11	0.09	0.08
AQ-10 Total	-0.16 **	0.05	-0.14 *	0.04
IDAQ Total	0.18 **	0.37 ***	0.16 **	0.36 ***
Openness	-0.05	-0.12 *	-0.06	-0.11
Extraversion	0.20 **	0.16 *	0.15 *	0.14 *
Agreeableness	0.14 *	-0.10	0.10	0.01
Conscientiousness	0.18 *	0.09	0.15*	0.10
Emotional Stability	0.14 *	0.12	0.09	0.12