Mental State Attribution to Robots: A Systematic Review of Conceptions, Methods, and Findings

SAM THELLMAN, Linköping University
MAARTJE DE GRAAF, Utrecht University
TOM ZIEMKE, Linköping University

The topic of mental state attribution to robots has been approached by researchers from a variety of disciplines, including psychology, neuroscience, computer science, and philosophy. As a consequence, the empirical studies that have been conducted so far exhibit considerable diversity in terms of how the phenomenon is described and how it is approached from a theoretical and methodological standpoint. This literature review addresses the need for a shared scientific understanding of mental state attribution to robots by systematically and comprehensively collating conceptions, methods, and findings from 155 empirical studies across multiple disciplines. The findings of the review include that: (1) the terminology used to describe mental state attribution to robots is diverse but largely homogenous in usage; (2) the tendency to attribute mental states to robots is determined by factors such as the age and motivation of the human as well as the behavior, appearance, and identity of the robot; (3) there is a *computer* < *robot* < *human* pattern in the tendency to attribute mental states that appears to be moderated by the presence of socially interactive behavior; (4) there are conflicting findings in the empirical literature that stem from different sources of evidence, including self-report and non-verbal behavioral or neurological data. The review contributes toward more cumulative research on the topic and opens up for a transdisciplinary discussion about the nature of the phenomenon and what types of research methods are appropriate for investigation.

CCS Concepts: • **Human-centered computing** → **HCI design and evaluation methods**; **Empirical studies in HCI**; **HCI theory, concepts and models**; *User models*; • **Applied computing** → *Psychology*; • **General and reference** → **Surveys and overviews;**

Additional Key Words and Phrases: Human-robot interaction, folk psychology, mentalizing, theory of mind, intentional stance, mind perception, anthropomorphism

ACM Reference format:
Sam Thellman, Maartje de Graaf, and Tom Ziemke. 2022. Mental State Attribution to Robots: A Systematic Review of Conceptions, Methods, and Findings. *ACM Trans. Hum.-Robot Interact.* 11, 4, Article 41 (September 2022), 51 pages.
https://doi.org/10.1145/3526112

This work was supported by ELLIIT, the Excellence Center at Linköping-Lund in Information Technology.
Authors' addresses: S. Thellman and T. Ziemke, Linköping University, Department of Computer and Information Science, Linköpings universitet, 581 83, Linköping Sweden; emails: {sam.thellman, tom.ziemke}@liu.se; M. de Graaf, Utrecht University, Department of Information and Computing Sciences, Princetonplein 5, Utrecht, 3584 CC, The Netherlands; email: m.m.a.degraaf@uu.nl.

2573-9522/2022/09-ART41 $15.00
https://doi.org/10.1145/3526112

1 INTRODUCTION

The term *mental state attribution* is used to refer to "the cognitive capacity to reflect upon one's own and other persons' mental states such as beliefs, desires, feelings and intentions" [Brüne et al. 2007]. In everyday social interactions, such attributions are ubiquitous, although we are typically not necessarily aware of the fact that they are *attributions*—or the fact that they are attributions of *mental states*. Any pedestrian encountering a car at a crosswalk, for example, is likely to ask themselves questions such as "Has the driver seen me?", "Do they understand I want to cross the road?", or "Are they planning to stop for me?" [cf. Ziemke 2020]. Answering these questions involves the attribution of intentional (directed) mental states to the driver, such as beliefs (e.g., there is a person on the crosswalk), desires (e.g., not to run over people), and intentions (e.g., to slow down and let the pedestrian cross the road).

In the case of human-human social interactions, such attributions are relatively unproblematic, because most people—generally speaking—know roughly what perceptual and cognitive capacities other people have, and consequently have a rough idea of what kinds of mental states they might or might not have. In the above example, the pedestrian and the driver are likely to understand the perspective of the other, and therefore, to predict each other's behavior. In the asymmetric case of human interactions with driverless cars or robots, on the other hand, things are less clear. Any driverless car, for example, is unlikely to have first-hand experience of being in the pedestrian's position. More typical robots, and humanoid robots, in particular, are maybe particularly challenging because on the one hand, they are obviously not human, but on the other hand, they can easily be interpreted as having, in Brink et al.'s [2019] terms, self-directed mechanical minds dwelling inside human-like bodies.

Hence, it probably does not come as a surprise that the scientific literature on how humans interpret robots in terms of mental or quasi-mental states is complex. For example, despite a widespread belief that robots do not have minds [Özdem et al. 2017], people frequently talk about and interact with robots *as if* they have minds. Many people might say that their robot lawnmower wants to avoid colliding with trees, although they would not say it has a mind, a will, or desires. In other words, it is not uncommon to conceptualize the behavior of robots as mind-governed without necessarily believing that robots really have minds, similar to how we interpret the behavior of fictional characters, companies, and nation-states [List et al. 2011; Wendt 1999]. For example, nations are commonly referred to as *wanting* to reach an agreement or as *believing* that a potential enemy is planning an attack, and some have even been described as "autistic" on account of their limited interaction with other states [Buzan 1993]. It has been suggested that the attribution of mental states and capacities helps us understand, explain, and predict behavior [Dennett 1989; Epley et al. 2007; Heider 1958; Mithen 1998]. However, it is still largely unclear exactly *how* helpful this is in the context of human-robot interaction, and we still do not know *why* it seems to work in our favor. Why do we sometimes attribute the behavior of robots to underlying mental states instead of, for example, computational or physical states? Moreover, how can we rely on attributed mental states as predictors of others' behavior if we are the ones doing the attributing? These and related questions have led researchers from a variety of disciplines, including psychology, neuroscience, computer science, and philosophy, to study mental state attribution to robots empirically. The studies that have been conducted so far exhibit considerable diversity in terms of how the phenomenon is described and how it is approached from a theoretical and methodological standpoint. While this interdisciplinary diversity is likely to contribute to making scientific progress at this early stage of research, there is an increasing need for researchers to develop a common language and a shared set of basic assumptions about the phenomenon at hand to be able to access and build cumulatively on each other's work.

A particularly pressing issue is that there is so far very little explicit discussion about what kinds of data constitute evidence of mental state attribution to robots. This is a significant problem because the lack of consensus implies a risk of incommensurability between obtained research findings and the absence of a basis for resolving apparent empirical contradictions. At this point, the literature is rife with conflicting findings—within the procured body of self-report evidence as well as between self-reports and non-verbal behavioral or neurological evidence—which call into question the validity of different sources of evidence. For instance, people commonly describe the behavior of robots in mental state terms [Duffy 2003; Hortensius and Cross 2018; Złotowski et al. 2015]. This might be interpreted as evidence that they attribute the behavior of robots to underlying mental causes. By contrast, the finding that (the same) people tend to deny that robots have minds or mental states when asked explicitly [Banks 2020; Fussell et al. 2008] might be interpreted as evidence that they *do not* attribute mental states to robots. Furthermore, answers to questions about the minds of robots have been found to vary significantly depending on *how such questions are asked* and *how respondents are allowed to answer* (e.g., forced-choice vs. not). For example, participants in a study by Fiala et al. [2014] refrained from describing a robot in mentalistic terms when they were provided with an additional response alternative that allowed them not to. There is also evidence that the ways that people talk about robots are not always indicative of how they interact with them. For example, people have been found to treat robots as if they were alive—for instance, hesitating to shut them off [Bartneck et al. 2007] or to hit them with a hammer [Bartneck and Hu 2008]—and to act in accordance with them having particular mental states, such as specific beliefs [Thellman et al. 2020], while at the same time verbally indicating that they do not think of them as being alive or as having those mental states. Moreover, findings based on activity in brain regions associated with mental state attribution that indicate (or not) mental state attribution may be difficult to consolidate with self-reports that suggest otherwise. For example, a study by Cross et al. [2019] found that study participants rated a robot appearing to be electrocuted as experiencing various levels of pain but could not observe any corresponding activation in participants' pain matrix during the observation of the electrocution. All these different types of conflicting findings motivate a broad and careful consideration of the research topic, including the nature of the phenomenon and the research methods employed.

The present literature review addresses the need for a shared understanding of mental state attribution to robots among scholars in the field by making visible different ways of thinking about the phenomenon (Section 4) and studying it (Section 5) across a broad range of disciplines. Previous research findings are collated to prevent unnecessary replication and inspire research questions that build systematically on previous work (Section 6). A critical assessment of the variation in obtained findings due to the types of methods employed is conducted to promote a transdisciplinary discussion of what types of research methods are appropriate for investigation (Section 7). Finally, open research questions on the topic are identified (Section 8). Previous reviews of smaller subsets of the literature on mental state attribution to robots have focused on different aspects, such as attribution of emotion [Hortensius et al. 2018] or "socialness" [Hortensius and Cross 2018], how mind attribution to robots evolves throughout the lifespan [Marchetti et al. 2018], determinants of anthropomorphism [Epley et al. 2007], and legal implications [Jaeger and Levin 2016]. Relatively few articles have taken a broader outlook on the phenomenon [Perez-Osorio and Wykowska 2020; Schellen and Wykowska 2019]. Moreover, none of the previously conducted literature reviews employed a pre-specified and auditable methodology for the purpose of systematically identifying and appraising all available evidence (i.e., they lack the characteristics of a "systematic literature review" [Kitchenham et al. 2007]). The systematic review presented in this article comprises 155 primary studies across multiple disciplines and addresses five specific questions:

RQ1. How is mental state attribution to robots *conceived* in the scientific literature in terms of (1A) terminology used to denote mental state attribution, (1B) reasons why people attribute mental states to robots, and (1C) underlying mechanisms.
RQ2. What *methods* have been used in studies of mental state attribution to robots in terms of (2A) stimulus materials and (2B) measures.
RQ3. What are the previous *findings* on mental state attribution to robots in terms of (3A) determinants, (3B) consequences, and (3C) comparisons with other agents?
RQ4. Do findings on mental state attribution to robots *vary* as a function of the methods—(4A) participant demography, (4B) how robots are presented, (4C) robot morphology, (4D) robot behavior, (4E) type of measure—employed?
RQ5. What are the *open research questions* about mental state attribution to robots?

2 REVIEW METHODS

A systematic literature review is a means of evaluating and interpreting all available research relevant to a particular research question, topic area, or phenomenon of interest, using a rigorous and auditable methodology. The present review is based on guidelines by Kitchenham et al. [2007] that have been used extensively in software engineering, human-computer interaction, and related fields.

Prior to conducting the review, the first author (S.T.) developed a review protocol. The purpose of a review protocol is to specify the methods that will be used to undertake a specific systematic review in order to reduce the possibility of researcher bias [Kitchenham et al. 2007]. The review protocol consisted of the following components: a rationale for the review, research questions, a literature search strategy including search terms and resources to be searched, study selection criteria and procedures, quality assessment procedures, a data extraction strategy specifying how the information required from each primary study will be obtained, a data synthesis strategy clarifying how the data will be analyzed, and a dissemination strategy specifying relevant publication venues. The review protocol was revised based on piloting the search strategy and evaluation by the other authors (M.d.G. and T.Z.). It was then accepted by all authors as the document guiding all subsequent review activities. In the following sections (Section 2.1–2.5), we describe the five components of the review processes.

2.1 Data Sources and Search Strategy

The aim of a systematic review is to find as many primary studies relating to the review questions as possible using an unbiased strategy [Kitchenham et al. 2007]. The search strategy employed in this review was developed in consultation with a librarian with expertise in the academic literature search. Two major academic search systems, Web of Science (Core Collection) and Scopus were used to source primary studies for the review. These search systems have been found suitable for the purpose of conducting systematic literature reviews because of the quality of the search functionalities they offer and the large size of the databases that they index [Gusenbauer and Haddaway 2020]. During the time that the review was conducted, both of these search systems indexed all journals and conference proceedings that were pre-identified by the authors as relevant (e.g., *ACM IEEE International Conference on Human-Robot Interaction*, *Frontiers in Robotics and AI*, *International Journal of Social Robotics*, *Lecture Notes in Artificial Intelligence*, *Science Robotics*, *ACM Transactions on Human-Robot Interaction*).

The data sources were searched using search terms that were matched against the titles, abstracts, and keywords of the indexed database records. An initial search query string was developed using various combinations of search terms derived from the review questions. Our aim at this stage was to search broadly and inclusively for all publications that describe the phenomenon

of interest. This means we had to identify and include all relevant terms associated with mental state attribution to robots in our search string along with their various permutations. To this end, we iteratively refined the search query string based on information from records found in multiple pilot searches. This process resulted in two functionally equivalent search query strings that were differently formatted to meet the requirements of the two search systems. The Web of Science search query string was: TS = (*robot*) AND (TS = (attribution OR ascription OR theory of mind OR mind perception OR intentional stance OR mentalizing OR anthropomorphism) OR (TS = (mind OR minds OR mental) AND TS = (attri* OR ascri* OR interpret* OR perce* OR infer* OR predict* OR expla* OR anthropomorph*))). The search was conducted in February of 2020 and resulted in two lists comprising 1,392 (Web of Science) and 1,457 (Scopus) records respectively, for a total of 2,849 records. The two lists were merged by removing duplicate items using reference management software. The merged list consisted of 2,112 publications.

2.2 Study Selection

The 2,112 publications that were identified in the search processes were assessed for their actual relevance based on a three-step study selection process guided by pre-defined selection criteria. The selection criteria were to (1) include only publications that present empirical data on the topic of mental state attribution to robots; (2) exclude publications in languages other than English; (3) exclude publications not subjected to peer review; and (4) in cases where a study is published in more than one journal and conference proceeding, exclude the least complete version. The study selection processes then proceeded sequentially according to steps 1–3 below (see also Figure 1):

(1) Each publication (N = 2,112) was independently judged by all authors (S.T., M.d.G., and T.Z.) as irrelevant, relevant, or highly relevant based on its title. Publication titles that were perceived as difficult to judge were conservatively marked as "relevant". All publications that were judged as relevant (or highly relevant) by a majority of raters were kept for the next step and the rest were discarded.
(2) Each of the remaining publications (N = 456) was independently judged by all authors (S.T., M.d.G., and T.Z.) as irrelevant, relevant, or highly relevant based on its abstract. All publications that were judged as relevant (or highly relevant) by a majority of raters were kept for the next step and the rest were discarded.
(3) Each of the remaining publications (N = 209) was judged as irrelevant or relevant by the first author (S.T.) based on a full read-through. Additional publications (N = 33) that were deemed relevant by all three authors were included. The additional publications were obtained through reference harvesting (N = 12) and a complementary database search conducted in May 2021 (N = 21). This resulted in a final selection of 155 relevant publications.

2.3 Study Quality Assessment

Assessing the quality of the primary studies included in a systematic review can be important and useful for several purposes, including establishing additional study selection criteria, investigating whether quality differences provide an explanation for differences in results, weighing the importance of individual studies in syntheses of reviewed findings, and for guiding the interpretation of findings and recommendations for further research [Kitchenham et al. 2007]. There is no agreed-upon definition of "study quality". However, according to Kitchenham et al. [2007], most quality checklists used in the context of systematic reviews include questions aimed at assessing the extent to which articles have addressed bias ("a tendency to produce results that depart systematically from the "true" results"), internal validity ("the extent to which the design and conduct of the study are likely to prevent systematic error"), and external validity ("the extent to which

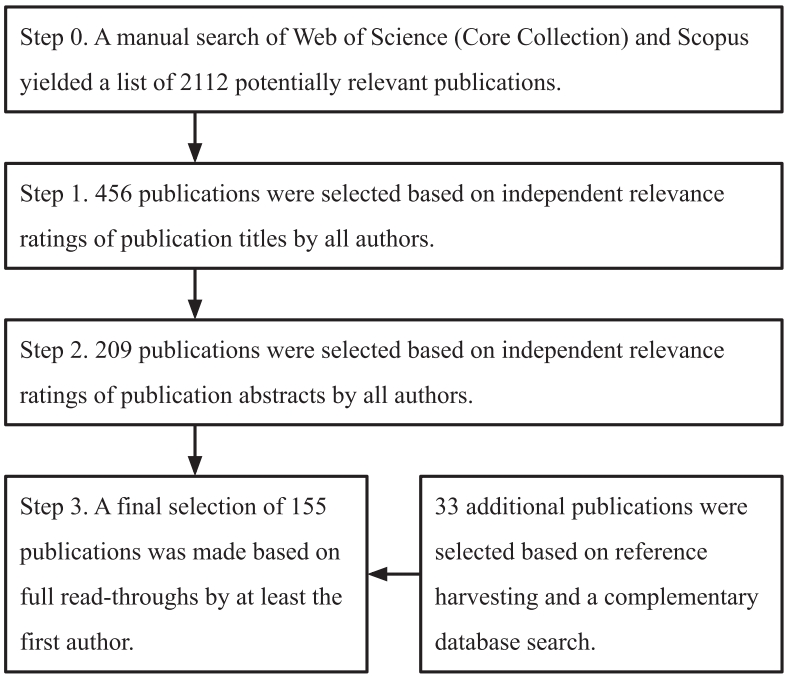

Fig. 1. Schematic overview of the study selection process.

the effects observed in the study are applicable outside of the study"). Kitchenham et al. [2007] recommend to select only quality evaluation questions that are appropriate for the specific review questions at hand. Examples of quality evaluation questions are "Was the sample size justified?", "Are the measures used in the study the most appropriate relevant ones for answering the research questions?", and "What do the main findings mean?".

However, the purpose of the present review was to address the lack of shared understanding of mental state attribution to robots by charting a landscape of pre-existing ideas, methods, and findings that exhibit considerable diversity. To discount individual studies on account of their particular approach of study would, in our judgment, be premature and risk precluding the possibility of a systematic and unbiased review. For this reason, we decided to refrain from quality assessment in our review. The implications of this decision are discussed in Section 10.

2.4 Data Extraction

The data extraction procedure was designed to obtain all the data necessary to answer the review questions and was conducted by the first author (S.T.). The categories of data that were extracted from each primary study are listed in Table 1. Some of the data were categorized (e.g., according to a type of method or finding). We state the rationale behind each of these categorizations in the remainder of this section.

The number of *Participants: Proportion of women/female* was calculated based on the number of female and/or women participants relative to the total number of participants (also including categories such as *male*, *man*, and *other*) reported in each publication. *Study setting* categories (Field, Lab, Online) were pre-defined (i.e., established before the review process started). Studies that took place in a controlled environment (e.g., a designated experiment room) were considered as "lab studies" even if they were set up in the field (e.g., in a school or museum). *Stimuli: Robot*

Table 1. Categories of Data Extracted from Each Primary Study

Data label [Data type]
Publication: Author [Textual]
Publication: Year [Numeric]
Publication: Source [Textual]
Participants: Number [Numeric]
Participants: Population [Textual]
Participants: Age [Numeric]
Participants: Proportion women/female [Numeric]
Study setting [Categorical: Field, Lab, Online]
Stimuli: Robot presentation [Categorical: e.g., Text, Image, Video, Present]
Stimuli: Robot morphology [Categorical: Anthropomorphic, Zoomorphic, Functional, Not Applicable]
Stimuli: Robot behavior [Categorical: Social—interaction with the participant, Social—interaction with non-participant other, Non-social, No behavior]
Measure: Operationalization [Categorical: e.g., Judged possession of mind/mental states]
Measure: Tool [Categorical: e.g., Likert scale, Binary choice, Free text]
Measure: Data type [Categorical: Verbal, Behavior (non-verbal), Neurological]
Finding [Textual]
Finding: Type [Categorical: Determinant, Consequence, State Ascribed, Agent contrast]
Terms used to describe phenomenon [Textual]
Statements including definitions of terms used to describe phenomenon [Textual]
Statements about the reason why people attribute mental states to robots [Textual]
Statements about the nature of the phenomenon [Textual]

morphology categories (anthropomorphic, zoomorphic, functional, not applicable) were based on Fong et al. [2003]. Categorization was based primarily on the authors' own descriptions of their stimulus. Text-based stimuli were included in the categorization. Although morphological cues are not strictly present in text-based stimuli, representations of the cues may be involved in processing text via semantic associations [Fiala et al. 2014]. *Stimuli: Robot behavior* categories (social—participant interaction, social—interaction with non-participant other, non-social, no behavior) were pre-defined. "Social behavior" was defined as behavior that (typically) occurs in the context of social interaction (e.g., with a study participant or a person represented in stimulus materials). Examples of "Non-social behavior" include interacting with inanimate objects and responding to events in the environment. Text-based stimuli were included in the categorization because they may evoke representations of behavioral cues through semantic associations (cf. above). *Measure: Data type* categories (verbal, behavioral, neurological) were pre-defined. Notably, the descriptor "verbal data" was applied to self-report measures such as Likert scales, semantic differential scales, and other quantitative methods used to collect continuous or ordinal data with assigned semantic content [Lavrakas 2008]. "Behavioral data" was applied to non-verbal behavioral data only. *Finding: Type* categories (determinant, consequence, comparative finding) were pre-defined. The distinction between determinants and consequences of mental state attribution to robots, and the categorization of determinants into human and robot factors was inspired by Waytz et al. [2010a]. *Stimuli: Robot presentation* categories (i.e., categories denoting how robots were presented to participants in studies; e.g., using text, image, video, or physically present robots), *Measure: Operationalization* categories (e.g., judged possession of mind, knowledge estimation), and *Measure: Tool* categories (e.g., Likert scale, binary choice, free text) emerged during the review process in a data-driven fashion based on descriptions of the methods that were employed in the primary studies.

2.5 Data Synthesis

The extracted data was collated and summarized in text and tabular form in a manner consistent with the review questions. Tables were structured to highlight similarities and differences between primary studies. A descriptive (narrative) synthesis of data [cf. Kitchenham et al. 2007] was conducted by the first author (S.T.). The procedure of the descriptive synthesis involved integrating studies comprising quantitative as well as natural language results and conclusions into a cohesive narrative. The focus was on identifying homogeneity and heterogeneity across studies in terms of conceptions (RQ1), methods (RQ2), and findings (RQ3), critically assessing how findings vary depending on employed methods (RQ4) based on results from investigating RQ2–3, and on identifying open research questions on the topic (RQ5) based on answers to RQ1–4. Quantitative synthesis (i.e., meta-analysis) was deemed unfeasible due to the diversity of research questions and methods across the primary studies.

3 OVERVIEW OF INCLUDED PUBLICATIONS

This section summarizes some information about the publications included in the review to provide an overview. For a complete list of publications, including details on participant number, population, age, sex/gender, and the study setting(s) reported in each publication, see Tables 2–4 in the Appendix.

The publication year of the included publications ranges from 1996 to 2021. Within this period, there was a growing trend in the number of publications on the topic per year (see Figure 2)—particularly in the years 2019–2021. 63% of the included publications were published in a journal, 36% in a conference proceeding, and 1% in a book section. The most common publication outlets for journal publications were *Frontiers in Psychology* (7% of all publications), *International Journal of Social Robotics* (7%), and *PLOS ONE* (5%). The most common publication venues for conference articles were *ACM/IEEE International Conference on Human-Robot Interaction, HRI* (12% of all publications), *IEEE International Conference on Robot and Human Interactive Communication, RO-MAN* (12%), and *International Conference on Social Robotics, ICSR* (5%).

Seventy-four percent of publications reported a study conducted in a laboratory setting, 25% reported a study conducted online, and 4% reported a study conducted in a field setting. 3% of publications reported studies conducted in more than one type of setting. The average number of participants per study was 68 when disregarding online studies (171 otherwise; the average N for online studies was 473).

Taken together, the participant population that the reported studies were based on can be summarized as WEIRD, i.e., Well-Educated, Industrialized, Rich, and Democratic [Henrich et al. 2010]). 25% of studies report recruiting participants from a university population. The studies were conducted in over 12 countries, including Australia, Canada, Denmark, France, Germany, Italy, Japan, New Zeeland, the Netherlands, Sweden, the United Kingdom, and the United States. 82% of the reported studies were based on participants from the adult population (i.e., 18 years or older); 18% were based on children (younger than 18 years) and a single study was based on an animal population. The mean age of study participants was 24.5 years old. 52% of participants were reported as female and/or as a woman.[1]

4 TERMINOLOGY AND CONCEPTIONS

In order to understand how mental state attribution to robots is conceived in the scientific literature (RQ1), we reviewed the terminology used to denote attribution of mind to robots (Section 4.1),

[1]This number was calculated based on the number of women and/or female participants relative to the total number of participants (also including categories such as *male, man*, and *other*) reported in each publication.

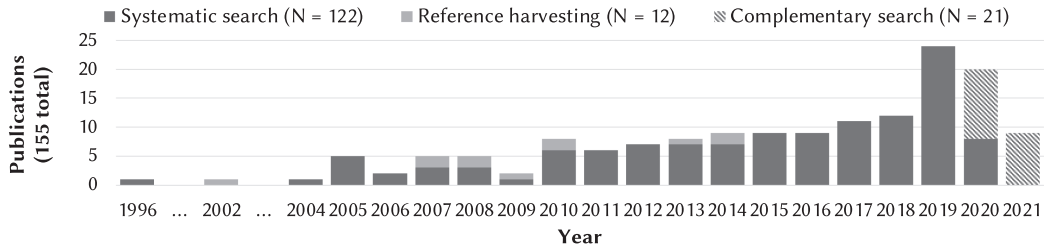

Fig. 2. Included publications sorted by publication year and clustered by source.

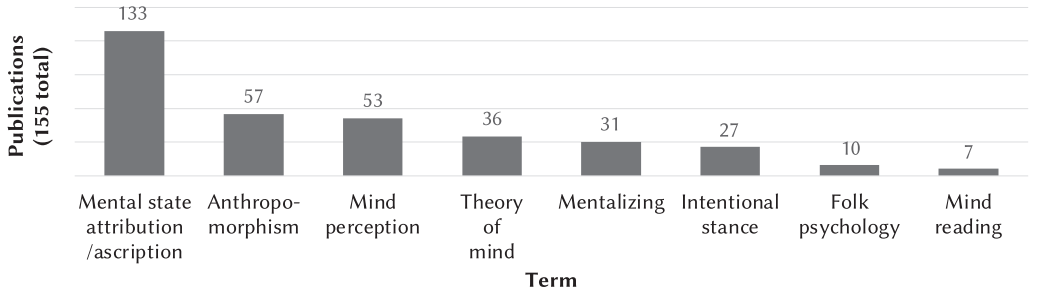

Fig. 3. Terms used in the literature to refer to attributions of mind, mental capacities, and mental states.

stated reasons for why people attribute mind to robots (Section 4.2), and statements about the underlying mechanisms (Section 4.3). All research is guided by preconceptions and hypotheses about the phenomenon at hand that is not (yet) supported by evidence. We specifically targeted the reasons for and mechanisms underlying mental state attribution to robots because, to the best of our knowledge prior to conducting the review, relatively little empirical research has so far been conducted on these issues compared with other aspects of the phenomenon, such as the determinants and consequences of mental state attribution to robots.

4.1 RQ1A: How is Mental State Attribution to Robots Conceived in the Scientific Literature in Terms of the Terminology used to Denote the Phenomenon?

We identified eight terms used in the reviewed literature to refer to attributions of mind, mental capacities, and mental states to robots (in descending order by the number of publications in which they occur): *mental state attribution/ascription* (87% of publications), *anthropomorphism* (37%), *mind perception* (34%), *theory of mind* (23%), *mentalizing* (20%), *intentional stance* (17%), *folk psychology* (6%), and *mind reading* (5%) (Figure 3)[2]. These terms overlap in meaning but have separate connotations that stem from their uses in different theories related to the underlying phenomenon. We discuss some of these connotations below.

The term *attribution* is generally understood as the processes by which "the social perceiver uses the information to arrive at causal explanations for events" [Fiske and Taylor 1991, p. 23] and is primarily associated with attribution theory within psychology [e.g., Heider 1958; Kelley 1967]. The term "*mental state/mind* attribution" is common both in the psychological and philosophical literature and refers to the attribution of events (including behavioral events) to underlying

[2]We note that with the exception of the two least common terms, *folk psychology* and *mind reading*, all of these terms were included in the search query string that was part of the search strategy employed to identify the primary studies on which the present review is based.

mental causes. The term "mental state *ascription*" (and its permutations) is commonly used as a synonym for "mental state attribution". The term "mental state *inference*" is sometimes used to emphasize that attribution involves inference of directly unobservable ("hidden", "internal") mental states based on observable evidence (e.g., behavioral cues) [for a detailed discussion, see Csibra and Gergely 2007].

Anthropomorphism is commonly used to refer to the attribution of both mental and non-mental states, such as human-like appearance [cf. Epley et al. 2007]. In recognizing that anthropomorphism is a broader concept, some authors disqualify mind-unrelated aspects of anthropomorphism by referring to mental state attribution using terms such as "psychological anthropomorphism" [Eyssel and Reich 2013; Kamide et al. 2013; Salem et al. 2011; Trovato and Eyssel 2017]. It might also be worth noting that the term anthropomorphism implies that ascribed mental states are distinctively human. This assumption is dubious for at least two reasons. Firstly, many mental states appear to be non-specific to humans. For example, Sommer et al. [2019] stated, "We initially described our measure as "anthropomorphism." However, upon feedback from reviewers, we noted that anthropomorphism refers to a tendency to ascribe human traits to nonhuman entities. Pain is not a uniquely human trait. Thus, the inclusion of pain results in anthropomorphism being an inaccurate description of this measure". Secondly, the question of whether robots and other (computational) artifacts could have a mind should be treated as an empirical possibility and not dismissed out of hand as it is the topic of an ongoing philosophical debate that has been highly active since the 1950s [e.g., Fodor 1975; Scheutz 2002; Searle 1980; Smith 2019; Turing 1950].

Mind perception is, at least in the context of the reviewed publications, closely associated with the work of Gray et al. [2007] who used self-report data to identify different "dimensions" of mind perception (distinguishing between agency-related and experiential mental states). The term suggests that mental state attribution is (at least partially) a perceptual process. We note that it seems possible to rephrase all encountered statements in which the term occurs by replacing the term "perception" with "attribution/ascription" without no apparent contradiction or loss in meaning (i.e., "she *perceived* the robot as having a mental state x" can be paraphrased as "she *ascribed* mental state x to the robot"), which suggest that the term "mind perception" does not differ significantly from "mind attribution" in theoretical import, at least not in the context of the reviewed literature.

Theory of mind has traditionally been the preferred term used by developmental and experimental psychologists to describe the capacity to attribute mental states to others [e.g., Baron-Cohen et al. 1985; Premack and Woodruff 1978]. It is often said that a person (e.g., a child) who lacks the ability to understand the mental states or viewpoint of others lacks a theory of mind. The term has also been used to describe the ability of robots to infer the mental states of humans [e.g., Scassellati 2002].

Mentalizing is frequently used in neuroscientific research in association to the "mentalizing system/network" which comprises brain regions associated with mental state attribution [Van Overwalle and Baetens 2009]. Bateman and Fonagy [2012] defined mentalizing as "the fundamental human capacity to "read" one's own and others' mental states such as thoughts and feelings".

Intentional stance is the central construct in a broader philosophical theory about the nature of the mind called intentional systems theory [Dennett 1989]. This theory views the mental state attribution as one of the multiple modes of attribution or "stances" that people can adopt to interpret and interact with objects and others in their surroundings.

Folk psychology, also sometimes called belief-desire, naive, intuitive, or commonsense psychology, refers to people's non-scientific understanding of the minds of themselves and others [Griffin and Baron-Cohen 2002]. This includes views about intentional, content-bearing, representational states (beliefs, desires, intentions, hunches, etc.) as well as phenomenal states (e.g., undirected

anxieties, feelings, and pain), traits, dispositions, and empirical generalizations such as that people who walk with a white cane might not know what is in front of them [Griffin and Baron-Cohen 2002].

Mind reading, or the ability to "read someone's mind", is a term used in the philosophical and scientific literature to describe the process by which people attribute mind, mental state and capacities [Nichols and Stich 2003].

Despite the differences in the meaning of the above terms, they were all used across the reviewed publications to refer to attributions of mind, mental capacities, and mental states to robots. In several cases, more than one term was used in the same publication. Table 5 in the Appendix provides an overview of the terms and their occurrence in the reviewed publications and Table 6 provides examples of statements that illustrate the homogenous usage of the terms. In some publications, two or more of the terms were explicitly defined as distinct in meaning. For example, Abubshait and Wiese [2017] distinguished between *mentalizing* as the use of "information from gestures, facial expression or gaze direction to make inferences about what others think, feel or intend to do" and *mind perception* as "the belief that social cues originate from an entity with a mind, capable of having internal states like emotions or intentions". Bossi et al. [2020] wrote, "When we adopt the *intentional stance* toward others, we refer to their mental states—such as beliefs, desires, or intentions—to explain and predict their behavior. We distinguish the concept of intentional stance from the process of *mentalizing*. Mentalizing refers to predicting a very specific and current instance of behavior with reference to a specific mental state. On the contrary, the intentional stance is more like a general attitude toward an agent—an assumption that the agent is an intentional entity rather than a simple mechanistic artifact". Also in these cases, the terms were all used to refer to attributions of mind, mental capacities, or mental states to robots. Furthermore, explicit distinctions were not found to be stable across multiple studies included in the review, as also demonstrated by the examples in Table 6.

Based on this finding, we conclude that the terminology employed to describe mental state attribution to robots is diverse but largely homogenous in usage (i.e., terms like *mind perception*, *[psychological] anthropomorphism*, *intentional stance*, and *mental state attribution* are to a significant extent used synonymously across and within individual studies). This suggests that researchers from various disciplines use different terms to refer to the same underlying phenomenon—a finding that we hope can facilitate cross-disciplinary dissemination of future research on the topic. It might also motivate the adoption of a standard notation to increase the accessibility of research. We suggest using "mental state attribution/ascription" based on the prevalence and relative theory-neutralness of the term. Standard notation can be deviated from when alternative descriptions are theoretically motivated (e.g., when using the term "intentional stance" in the context of an investigation where the intentional stance is contrasted with the design stance or physical stance).

Finally, we note that mental state terms are to some extent treated inconsistently across studies as either metaphorical or literal by enclosing them (or not) in quotation marks. We recommend against the metaphorical treatment on the grounds that mental state ascriptions do not necessarily involve any ontological commitments (i.e., they do not entail beliefs about whether ascribed states are real or fictive; for a more detailed discussion, see Thellman and Ziemke [2019]).

4.2 RQ1B: How is Mental State Attribution to Robots Conceived in the Scientific Literature in Terms of the Reason Why People Attribute Mental States to Robots?

We found no clear consensus in the literature about the reason(s) why people attribute mental states to robots. Several authors stressed the importance of mental state attribution to the ability to interact with robots. A recurrent assumption is that attributing mental states helps people predict and explain the behavior of robots [de Graaf and Malle 2019; Epley et al. 2007; Levin et al. 2013;

Marchesi et al. 2019; Thellman 2021]. For example, Imamura et al. [2015] stated, "Attributing mental states to others, based on their complex behaviors, enables an agent to understand another agent's current behavior and predict its future behavior". It has also been suggested that it may reduce stress and uncertainty and increase one's sense of control in the context of interactions with robots [Epley et al. 2007; Eyssel et al. 2011]. We identified two studies that provide empirical support for the claims that people ascribe mental states to robots to increase the predictability of their behavior and that robots are perceived as more predictable when people ascribe mental states to them: Waytz et al. [2010b] and Eyssel et al. [2011]. We also identified seven studies of how mental state attribution affects people's (actual) predictions of behavior: Banks [2020], Levin et al. [2012, 2013], Rueben et al. [2021], Sciutti et al. [2013], Thellman and Ziemke [2020], and Zhang et al. [2019].

4.3 RQ1C: How is Mental State Attribution to Robots Conceived in the Scientific Literature in Terms of the Underlying Mechanisms?

There is not yet a consensus about the underlying mechanisms that govern mental state attribution to robots, including to what extent they differ or overlap with those toward humans. There have been multiple suggestions that the relevant cognitive processes occur at distinct levels as denoted using terms such as *low-road vs. high-road processing* [Fiala et al. 2014], *implicit vs. explicit processing* [Banks 2020; Takahashi et al. 2013], *first-line vs. second-line reasoning* [Levin et al. 2012], and *type 1 vs. type 2 processing* [Hannibal 2014; Złotowski et al. 2018]. Such *dual processes theories* [Evans 2003] may help explain some of the apparent contradictions that exist between findings from different sources of evidence (e.g., self-report and behavioral data; see Section 7).

It has also been suggested that mental state attribution to robots is a two-step process, whereby the individual must engage in a higher-level thought process in which he or she decides to adopt an interpretative mentalistic or intentional stance toward the robot *before* lower-level attribution processes are activated [Martini et al. 2016; Wiese et al. 2017; Wykowska et al. 2014]. For example, Wiese et al. [2017] claimed, "Reasoning about the internal states of others is referred to as mentalizing, and presupposes that our social partners are believed to have a mind". This claim appears in contradiction with the observation that people frequently attribute mental states to robots (or nation states and even animated geometric figures [Heider and Simmel 1944]) while simultaneously stating that they do not believe such entities have a mind [e.g., Banks 2020; Fussell et al. 2008].

5 RESEARCH METHODS

In this section, we describe the research methods that have been used in previous studies of mental state attribution to robots (RQ2). This includes stimulus materials (Section 5.1) and measures (Section 5.2). Stimuli were classified according to how robots are presented to study participants, the physical appearance of the presented robots, and the type of behavior that they exhibit (if any). Measures were classified according to the type of data they were used to collect (i.e., verbal, behavioral, or neurological), how they operationalize the measured phenomenon (e.g., the judgment of mind possession, brain activity), and what type of measurement tool was used to collect the data (e.g., Likert scale, fMRI). See Section 2.4 for the rationale behind the categorization of research methods. For a complete list of stimulus materials and measures, see Tables 7–11 in the Appendix. All proportions (%) mentioned in Section 5 are relative to the total number of reviewed publications (N = 155) unless otherwise stated.

5.1 RQ2A: What Types of Stimulus Materials have been used in Studies of Mental State Attribution to Robots?

A majority of publications (67%) reported using some kind of representation (e.g., image or text) of a robot as part of the stimulus materials presented to participants. 43% of publications presented

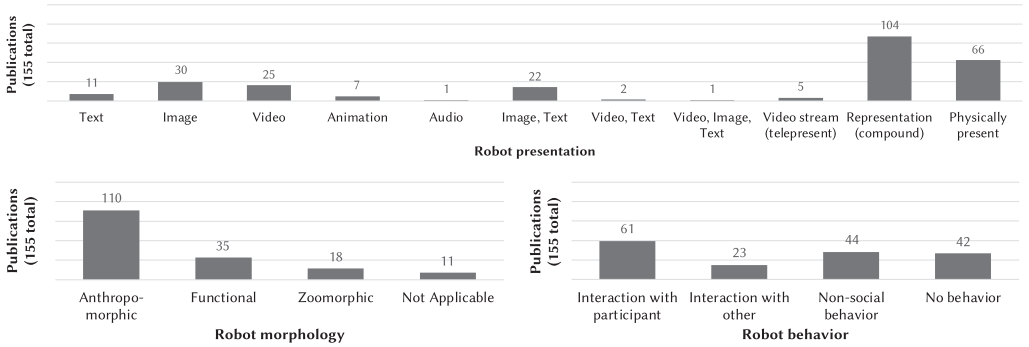

Fig. 4. Types of stimulus materials used in studies of mental state attribution to robots. *Top:* How robots were presented to study participants. "Representation" is a compound category, which includes all other categories except "Physically present". *Bottom-left:* The physical appearance of robots presented to study participants. *Bottom-right:* The type of behavior exhibited by robots presented to study participants.

study participants with a physically present robot. The most common type of representation was an image (19%), followed by video (16%), a combination of image and text (14%), text-only (7%), animation (5%), video stream (i.e., telepresence; 3%), a combination of video and text (1%), a combination of video, image, and text (1%), and audio-only (1%). 8% of publications used more than one of these types of methods of presenting robots to study participants. See Figure 4, *Top*, for a visual representation of these results.

The most common morphology of the robots presented to study participants was the *anthropomorphic* form (71% of publications), followed by *functional* (23%) and *zoomorphic* robots (12%). In 7% of the studies, participants were presented with robots without a manifest physical appearance (e.g., verbal descriptions of robots or robot sounds). 12% of publications presented study participants with more than one type of robot morphology. See Figure 4, *Bottom-Left*, for a visual representation of these results.

A majority of publications (73%) presented study participants with a robot that exhibited behavior: 39% presented a robot exhibiting social behavior in the context of a direct interaction with the study participant, 15% presented a robot exhibiting social behavior in the context of interacting with a non-participant other (e.g., the experimenter or person depicted in a video), and 28% presented a robot exhibiting non-social behavior (e.g., interacting with physical objects). 27% of publications presented a robot exhibiting no behavior. 8% of publications presented study participants with more than one of these four types of stimuli. See Figure 4, *Bottom-Right*, for a visual representation of these results.

5.2 RQ2B: What Measures have been Employed in Studies of Mental State Attribution to Robots?

A majority of publications (84%) relied on verbal measures of mental state attribution. 14% relied on (non-verbal) behavioral data and 9% relied on neurological data. 8% relied on more than one type of measure (see Figure 5).

Among the studies that relied on verbal data, the predominant type of measure—used in 59% of all reported studies—was the use of Likert or semantic differential scales (i.e., questionnaire methodology) to collect study participants' judgments of robots' possession of the mind, mental capacities and/or mental states. Other types of operationalization of the phenomenon of mental state attribution include mentalistic descriptions or explanations of robots or robot behavior (e.g.,

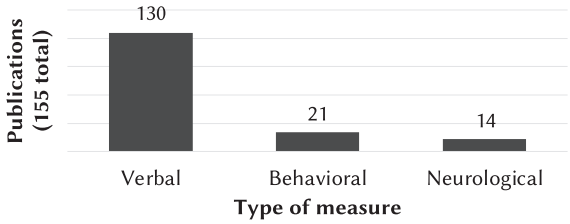

Fig. 5. Types of measures employed in studies of mental state attribution to robots.

provided in the form of free-text responses) and mentalistic predictions (i.e., predictions of the behavior of a robot that rely on assumptions about its mental states such as beliefs or desires). The predominant tool used in studies based on a child population was spoken or written questions about the mental states of robots in combination with a binary choice response format (i.e., typically yes-no questions). See Table 10 in the Appendix for a complete list of verbal measures.

The selection of behavioral measures used is diverse and cannot easily be summarized. Some studies focused on people's tendency to treat robots in manners consistent with ascriptions of mental states. For instance, studies by Bartneck and colleagues measured participants' willingness to abuse [Bartneck and Hu 2008] or turn off a robot [Bartneck et al. 2007]. Lemaignan et al. [2015] and Straub [2016] measured people's tendency to use social behaviors, such as gestures and polite speech, in interactions with a robot. A few studies measured people's tendency to help a robot achieve a goal [Martin et al. 2020; Yamaji et al. 2010] or behave altruistically toward it [Heijnen et al. 2019]. Other studies monitored study participants' performance in tasks, which required mental state ascription. Mutlu et al. [2009] measured people's belief ascriptions based on people's performance in a guessing game that depended on taking into account the non-verbal leakage of a robot opponent. Zhao et al. [2016] exploited people's tendency to take a robot's visual perspective as a measure of belief ascription. Two studies relied on people's anticipatory gazes toward end-of-motion positions of robots performing goal-directed [Sciutti et al. 2013] and/or belief-directed [Thellman and Ziemke 2020] actions as a non-verbal measure of people's mentalistic predictions of robot behavior. See Table 11 in the Appendix for a complete list of behavioral measures.

The most common neuroimaging technique used in studies relying on neurological data was functional magnetic resonance imaging (fMRI; 12 of the 14 studies that employed a neurological measure) followed by electroencephalography (EEG; 2 of 14 studies). Neurological measures targeted brain regions generally associated with mental state attribution in the neuroscientific literature, such as the mPFC and TPJ (for an overview, see Van Overwalle and Baetens [2009]). See Table 11 in the Appendix for a complete list of neurological measures.

6 RESEARCH FINDINGS

In this section, we review research findings on mental state attribution to robots (RQ3). The findings were categorized into three categories (as visualized in Figure 6): determinants (Section 6.1) and consequences (Section 6.2) of mental state attribution to robots, and comparative findings that contrast mental state attribution to robots against mental state attribution to other agents, such as humans and computers (Section 6.3). Determinants were further categorized into human factors and robot factors, and consequences were categorized as either psychological or behavioral. Comparative findings were categorized according to the type of agent compared against. See Section 2.4 for the rationale behind the categorization of research findings. See Tables 12–15 in the Appendix for a complete list of reported findings and publication references. 71% of the reviewed publications reported findings on determinants, 25% reported consequences, and 32% reported

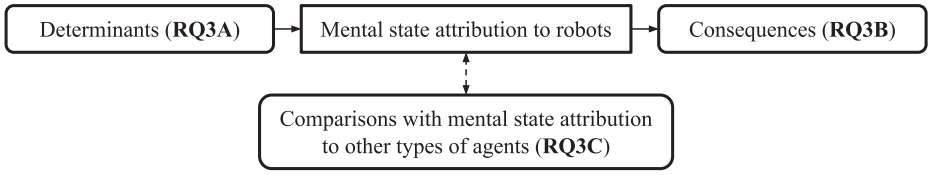

Fig. 6. Classification of research findings on mental state attribution to robots.

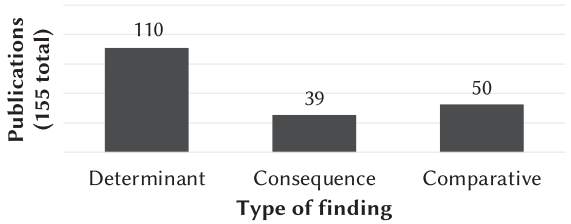

Fig. 7. Types of reported research findings on mental state attribution to robots.

comparative findings (see Figure 7). All proportions (%) mentioned in Section 6 are relative to the total number of reviewed publications (N = 155) unless otherwise stated.

6.1 RQ3A: What are the Previous Findings on the Determinants of Mental State Attribution to Robots?

A majority of the reviewed publications (71%) studied determinants of mental state attribution to robots. 40% of publications studied *human factors* that determine the tendency to attribute mental states to robots and 35% studied *robot factors*. 5% studied both human and robot factors. See Tables 12–13 in the Appendix for a complete list of studied determinants.

6.1.1 Human-Factor Determinants. Among the human factors, the most extensively studied are *age* (30 studies), followed by *motivation* (15 studies), *cultural and socioeconomic background* (eight studies), *interaction history* (six studies), *gender* (three studies), *mental disorder* (three studies), and *species* (one study). See Table 12 in the Appendix for a complete presentation of the studied human-factor determinants of mental state attribution to robots.

A total of 25 studies indicate that children tend to attribute mental states to robots. Three studies reported a generally stronger tendency compared to adults. Seven studies indicate that younger children have a particularly strong tendency (one study reported contradictory findings). Three studies indicate a stronger tendency in older compared with younger adults (one study reported contradictory findings and two studies found no effect in either direction). One study observed a tendency in infants. Based on the reported findings, we conclude that people of all ages appear to attribute mental states to robots albeit possibly to a different extent. In particular, there is relatively strong basis for concluding a stronger tendency in younger children. However, it should be noted that most of the studies reporting these findings employed verbal measures of mental state attribution (the implications of which are discussed in Section 8.1).

A total of 15 studies reported findings on various motivational determinants. These include a stronger tendency to attribute mental states to robots when motivated to predict robot behavior, when anticipating future interaction with a robot when having high expectations about a robot's capabilities, when a robot is perceived as being mistreated or subjected to harm, and when feeling lonely. Except for the reportedly stronger tendency to attribute mental states when a robot is

perceived as being mistreated or subjected to harm (four studies), corroborating evidence is generally lacking for these motivational determinants. However, it appears that the motivational component is multifaceted and includes both an effectance dimension (i.e., motivation to understand and predict robot behavior) and a sociality dimension (i.e., desire for social contact and affiliation) [cf. Epley et al. 2007]. No study has so far reported evidence on the relative importance or strength of different types of motivational determinants.

Eight studies reported findings on cultural and socioeconomic determinants. Japanese individuals were found to exhibit a stronger tendency relative to Westerners, including Germans, Australians, and Italians across four studies. However, two studies that compared Japanese and Western individuals found no effect in either direction. One study also reported a stronger tendency in Chinese compared with US individuals. We believe these results should be interpreted cautiously given the contradictory findings and the fact that these studies employed self-report measures and people of different culture have slightly different styles of ratings [Bernardi 2006]. One study also reported a similar tendency in individuals with different socioeconomic backgrounds.

Similar attributional tendencies have been reported among men and women (three studies), adults with and without schizophrenia (one study), and humans and non-human primates (one study). One study reported a stronger tendency in children without autism as compared with autism and one study reported no differences. These findings need to be corroborated by further studies before any general conclusions can be drawn.

6.1.2 Robot-Factor Determinants. The most extensively studied robot factor is *behavior* (31 studies) followed by *appearance* (17 studies), *identity* (eight studies), *capability* (four studies), and *presence* (two studies). See Table 13 in the Appendix for a complete presentation of the reported robot-factor determinants of mental state attribution to robots.

A total of 31 studies indicate that robot behavior determines the tendency to attribute mental states to robots. There is corroborating evidence that people are more inclined to attribute mental states to robots when they exhibit various types of socially interactive behavior (13 studies), such as eye gaze (five studies), gestures (one study), cheating (one study), emotional expression (two studies), and when behavior is unpredictable (two studies), complex (one study), intelligent (one study), or highly variable (one study). One study found no difference when a robot exhibited social as compared to non-social behavior. It has also been found that differences in people's tendency to attribute mental states to robots versus people decreases when social behavior is present as a behavioral stimulus (see Section 6.3). These findings suggest that robot behavior might be also considered as a motivational determinant, in line with Epley et al. [2007] (i.e., that people's attributions are motivated by the need to make sense of and predict behavior). Moreover, six studies reported a stronger tendency related to the movement of a robot, and three studies reported different types of ascriptions depending on the type of behavior enacted by a robot.

A total of 11 studies reported a stronger tendency to attribute mental states to robots with (increasingly) human-like appearance, including when a face is visible or human-like facial wounds are present as compared to when they are not. Two studies found no such effect. One study indicated a weaker tendency when features that suggest violent conflict were present. Three studies reported that people ascribed different types of mental states depending on different types of physical features. These findings support the general conclusion that increasingly human-like robot appearance gives rise to a stronger tendency to attribute mental states to robots.

Eight studies indicate that the identity of a robot affects people's tendency to attribute mental states. Four studies reported that people attribute different types of states depending on a robot's function or purpose. Two studies observed differences when robots were gendered or described as belonging to a specific culture. Two studies observed a stronger tendency when robots were

described using study participant in-group names, country of origin, or had an in-group gendered voice.

Three studies indicate robot capabilities as a determining factor, including a stronger tendency to attribute mental states when robots have similar traits of imagination as the person and difficulties in attributing mental states to robots with different-from-human capabilities. One study a similar tendency when a robot was described as having mental capabilities as when it was not.

Two studies reported a stronger tendency when a robot was physically present rather than telepresent. These findings are in line with the literature review by Li [2015] that identified physical presence rather than physical embodiment as the stronger predictor of positive outcomes in interactions with artificial agents.

6.2 RQ3B: What are the Previous Findings on the Consequences of Mental State Attribution to Robots?

A total of 25 percent of the reviewed publications studied the consequences of mental state attribution to robots. 10% studied psychological consequences and 15% studied behavioral consequences. No publication studied both psychological and behavioral consequences. See Table 14 in the Appendix for a complete list of reported consequences.

6.2.1 Psychological Consequences. Five studies reported findings on the effect of mental state attribution on people's perceptions of robots as eerie or uncanny [Mori et al. 2012]. Two studies reported an increase in uncanniness, two studies reported a decrease, and one study found no association. We abstain from drawing any general conclusions about the association between mental state attribution and the uncanny valley phenomenon based on these contradictory findings. Two studies indicate that ascribing mental states to robots can drain cognitive resources. Two studies reported an increase in trust in a robot. Two studies reported an increased moral concern for a robot. Isolated studies found: increased perceived predictability of a robot, increased ambivalence in attitude toward robots, perceived threat of damage to humans and human identity, and a reduced sense of agency on part of the human.

6.2.2 Behavioral Consequences. Seven studies observed that study participants were able to predict the behavior of robots based on ascribing mental states such as beliefs and desires. Two studies reported difficulties in predicting robot behavior that stemmed from difficulties ascribing appropriate mental states (e.g., beliefs) or capabilities (e.g., perception) to robots. Four studies observed that participants could explain robot behavior based on ascribing mental states. These findings are consistent with the claim that one of the functions of mental state attribution is to predict and explain behavior [e.g., Dennett 1989; Epley et al. 2007, see also Section 4.2]. Three studies reported decreased abuse toward robots and one study found no effect of robot abuse. Isolated studies reported: An increased tendency to help a robot, decreased likelihood of using a robot, unwillingness to switch off a robot, and the absence of an effect on the tendency to mimic the facial expressions of a robot.

6.3 RQ3D: What are the Previous Findings on the Tendency and Types of Mental States Attributed to Robots as Compared with other Agents?

A total of 31 percent of the publications reported comparative findings on mental state attribution to robots that involved other types of agents. Among these publications, all reported comparisons against mental state attribution to humans. 4% of publications also reported comparisons against computers. Taken together, the findings support the conclusion that people have a relatively stronger tendency to attribute mental states to humans compared to robots and a relatively weaker tendency to attribute mental states to computers compared to robots (i.e., supporting a

computer < *robot* < *human* pattern). See Table 15 in the Appendix for a complete list of reported comparative findings.

6.3.1 Human vs. Robot. A total of 33 studies indicate a generally stronger tendency to attribute mental states to a human person compared to a robot. Seven studies reported a similar tendency. Six studies reported a stronger tendency to attribute specifically experience-related mental states (e.g., pain, feelings, emotions) to a human, and a similar tendency to attribute agency-related states (e.g., beliefs, desires, intentions). One study reported a stronger tendency to attribute valenced (but not unvalenced) mental states to a human. Four studies reported that people attributed similar types of mental states to robots and humans. One study observed that participants were more proficient in recognizing types of emotions in humans compared to robots.

Four studies reported factors that moderated differences in the tendency to attribute mental states to humans and robots. Two studies found that the presence of gaze behavior in a robot decreased observed differences. One study found that longer response times led to a decrease and one study found a decrease over time.

6.3.2 Computer vs. Robot. Five studies reported a generally weaker tendency to attribute mental states to a computer compared to a robot. One study reported that this difference was amplified when gaze behavior was present in a robot.

7 METHODOLOGICAL FACTORS

In this section, we identify systematic variation in the research outcomes in studies on mental state attribution to robots that can be directly related to the research methods employed (RQ4), specifically the demography of study participants (Section 7.1), how robots are presented to study participants (Section 7.2), the physical appearance of presented robots (Section 7.3), the behavior exhibited by presented robots (Section 7.4), and the measures used (Section 7.5). All proportions (%) mentioned in Section 7 are relative to the total number of reviewed publications (N = 155) unless otherwise stated.

7.1 RQ4A: Do Findings on Mental State Attribution to Robots Systematically Vary as a Function of the Demography of Study Participants?

We found no clear evidence of systematic variation in obtained findings depending on the demography of study participants. A tendency to ascribe mental states to robots has been reported across all demographic categories reported in the reviewed studies (e.g., age, gender and sex, cultural and socioeconomic background, various mental disorders). Five studies found a stronger tendency in individuals in Eastern cultures as compared with individuals in Western cultures and two studies contradicted this finding (cf. Section 6.1.1). Notably, these studies relied primarily on self-report measures. It has been found that people of different cultures have slightly different styles of ratings [Bernardi 2006]. For these reasons, we believe it would be premature to conclude that Easterners exhibit a generally stronger tendency to attribute mental states to robots than Westerners. There is evidence of a stronger tendency in children (particularly young children) compared with adults (cf. Section 6.1.1) which should be taken into consideration in the selection of study participants to avoid unjustified generalizations of research findings.

7.2 RQ4B: Do Findings on Mental State Attribution to Robots Systematically Vary Depending on how Robots are Presented to Study Participants?

We found no clear evidence of systematic variation in obtained findings depending on whether robots were presented to participants using text, image, video, animation, audio, a combination of these, or if the robot was physically present or telepresent. However, two studies reported a

stronger tendency to attribute mental states to physically present robots compared with telepresent robots [Kiesler et al. 2008; Straub 2016]. These findings are in line with the literature review by Li [2015] that identified physical presence rather than physical embodiment as the stronger predictor of positive outcomes in interactions with artificial agents. We also note that a majority (95%) of the studies that did not use any behavioral stimuli employed text, image, or a combination of text and image as stimulus materials. This should be taken into consideration when designing studies of mental state attribution to robots given evidence (described in Section 7.4) that the presence of behavioral stimuli affects the outcome of studies of mental state attribution to robots.

7.3 RQ4C: Do Findings on Mental State Attribution to Robots Systematically Vary Depending on the Morphology of the Robots Presented to Study Participants?

A tendency to ascribe mental states has been reported toward robots with all kinds of physical appearance, including robots that are human-like, animal-like, insect-like, machine-like, or furniture-like (e.g., robots in the shape of a trashcan [Yamaji et al. 2010], chair [Sirkin et al. 2015], or a shoe rack [Rueben et al. 2021]). There is evidence of a stronger tendency to ascribe mental states to robots with more human-like physical appearance relative to robots with less human-like appearance (11 studies; cf. Section 6.1.2). Two studies found no difference toward robots with different degrees of human-likeness. We note that all of these studies relied on verbal measures (which may be problematic for reasons described in Section 7.5).

7.4 RQ4D: Do Findings on Mental State Attribution to Robots Systematically Vary Depending on the type (or absence) of Behavior of the Robots Presented to Study Participants?

There is evidence that the presence of behavioral stimuli affects the outcome of studies of mental state attribution to robots. In particular, 13 studies reported a stronger tendency to attribute mental states to robots when they exhibit various types of socially interactive behavior (cf. Section 6.1.2). Furthermore, comparative studies indicate that difference between people's generally stronger tendency to attribute mental states to humans relative to robots is reduced in the presence of behavioral stimulus (cf. Section 6.3). Based on this, we conclude that it is reasonable to expect different research outcomes in studies on mental state attribution to robots depending on whether the robot(s) presented to study participants as stimulus materials exhibit the behavior. Moreover, all of the reviewed studies that used non-behavioral stimuli (28% of publications) relied on verbal measures (which may be problematic for reasons described in Section 7.5).

7.5 RQ4E: Do Findings on Mental State Attribution to Robots Systematically Vary as a Function of the Measure Employed?

There appears to be systematic variation in the findings produced by studies employing different types of measures of mental state attribution to robots. This means that the selection of measures may affect what types of research outcomes can be expected.

In our review of research methods employed in studies of mental state attribution to robots (Section 5.2), we classified measures based on the type of data they procure: self-report (verbal) data, behavioral (non-verbal) data, or neurological data. Hypothetically, six different types of conflicting findings may arise in the data obtained using these different types of measures: conflicts between data procured by *similar* types of measures ("A", "B", and "C" in Figure 8) and conflicts between data procured by *different* types of measures ("D", "E", and "F").

We identified three of these six types of conflicting findings as present in the empirical literature (solid arrows in Figure 8): conflicting findings within the procured body of self-report data (type "A"), conflicting findings between verbal and behavioral data (type "D"), and conflicting findings

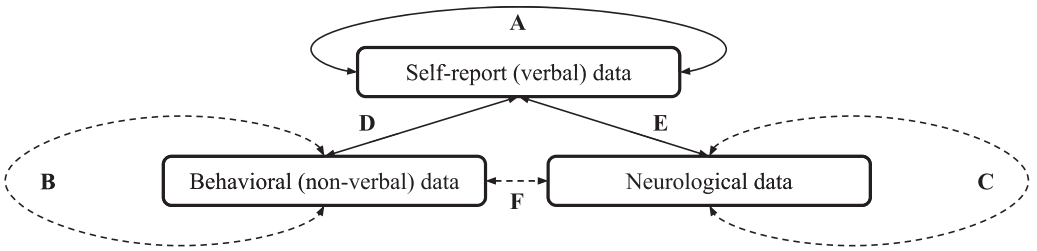

Fig. 8. Confirmed (solid arrows) and unconfirmed (dashed arrows) types of conflicting findings that stem from similar (A, B, C) or different (D, E, F) types of measures of mental state attribution to robots.

between verbal and neurological data ("E"). We did not find any clear evidence of systematic variation among the other three types of (hypothetical or unconfirmed) conflicting findings (dashed arrows): conflicts within the body of (non-verbal) behavioral data ("B"), conflicts within the body of neurological data ("C"), or conflicts between behavioral and neurological data ("F").

As a conflicting finding of type "A", there is evidence that verbal measures systematically vary depending on whether study participants are asked directly about the minds of robots (e.g., "Does this robot have a mind?") or if more "indirect" or "implicit" verbal measures are employed, such as asking participants to explain or verbally predict robot behavior. For instance, Fussell et al. [2008] found that participants in their study described robots in mental state terms in the context of an interview but also denied that robots have various mental capacities upon being asked more directly in a post-interview questionnaire. Banks [2020] compared study participants' responses to the direct question "Do you think [the robot] has a mind"? with the outcomes of four different types of implicit verbal measures and found no relationship between the two. Specifically, these findings suggest that studies that employ more direct or explicit verbal measures are more likely to observe a weaker tendency to attribute mental states to robots than studies that employ implicit verbal measures.

Another type-A conflict was reported in a study by Fiala et al. [2014]. When allowed to choose between non-mentalistic and mentalistic ways of describing the capabilities of a robot (e.g., the robot "identified the location of the box" vs. "knew the location of the box"), participants in their study preferred not to attribute mental states at all. Similarly, it was found in a series of studies [Bossi et al. 2020; Marchesi et al. 2019, 2021; Perez-Osorio et al. 2019] that a majority of study participants preferred non-mentalistic explanations over mentalistic explanations of robot behavior when given the opportunity to rate the plausibility of the two types of explanations. These findings support the generalization that studies that employ non-mentalistic verbal choice alternatives are more likely to observe a weaker tendency to attribute mental states to robots than studies that employ forced mentalistic verbal choice alternatives.

Conflicting findings of type D arise between self-report data that suggest that people attribute (or do not) mind or certain mental states or capacities to robots and non-verbal behavioral data that suggests otherwise. For instance, it has been observed in a number of studies that people are reluctant toward describing robots as having emotions, feelings, and other mental states related to phenomenal experience (cf. Section 6.3). In contradiction with these results, studies show that people are hesitant to inflict damage on or turn off robots [Bartneck and Hu 2008; Bartneck et al. 2007], which can be taken as an indication that people ascribe experiential states to robots. Other studies have shown that people's anticipatory gaze behavior is consistent with ascriptions of specific goals [Sciutti et al. 2013] and beliefs [Thellman and Ziemke 2020] and that these non-verbal behaviors are not always consistent with people's verbal self-reports [Thellman et al. 2020;

Wiese et al. 2019]. These results are consistent with the generalization that studies that employ non-verbal behavioral measures are more likely to observe a stronger tendency to attribute mental states to robots than studies that employ verbal measures.

Conflicting findings of type E arise between self-report data that suggest that people attribute (or not) mind or certain mental states or capacities to robots and neurological evidence that suggests otherwise. For example, Cross et al. [2019] found that study participants rated a robot appearing to be electrocuted as experiencing various levels of pain but could not observe any corresponding activation in participants' pain matrix during the observation of the electrocution.

We did not find any clear evidence of systematic variation among conflicting findings of type B, C, and F. This means that these types of data—to the extent that they are comparable in the sense that they bring to bear on the same research questions—appeared to be consistent with each other. However, it should be noted that some types of measures have been used primarily in the service of investigating specific aspects of mental state attribution to robots but less so for other aspects. For example, neurological measures have so far primarily been used to assess the tendency to attribute mental states to robots, mostly by focusing on comparisons with other types of agents, but were only used to investigate determinants or consequences in one study. Furthermore, since behavioral (non-verbal) measures have rarely been applied to study the relative tendency to attribute mental states to robots as compared with other agents, there are so far very few points of contact of type F. Hence, the prevalence or likelihood of conflicting findings of type B, C, or F, cannot be assessed based on the presented review results.

The conflicting findings described in this section indicate that mental state attribution to robots is a complex socio-cognitive process that, as independently suggested by several researchers, may be operating at different "levels" (cf. Section 4.3). They also call into question the validity of the different types of measures. Resolving or making sense of some of the conflicting findings may imply having to adjudicate between different types of measures. Hence, it is important going forward that the multidisciplinary research community *as a whole* engages in the question of what research methods are appropriate for studying mental state attribution to robots.

8 OPEN QUESTIONS

Based on the review findings (Sects. 4–7), we identify four open questions about mental state attribution to robots that are suitable targets for future research on the topic (RQ5). Note that we have intentionally excluded research questions about the nature of the minds ascribed to robots (e.g., "Can robots (ever) really have a mind?" and "When are mental state attributions to robots true or justified?") which pertain primarily to cognitive science [Thellman and Ziemke 2019] and philosophy of mind and cannot presently be answered using empirical methods.

8.1 What Methods are Appropriate for Studying Mental State Attribution to Robots?

There is a considerable diversity of research methods employed in studies of mental state attribution to robots (cf. Section 5). The methodological choices that individual researchers make have direct consequences on what kinds of study outcomes they can expect (cf. Section 7). Yet, there is very little explicit discussion among scholars in the field about what methods are appropriate for investigation.

Based on our review, a typical study on mental state attribution to robots is conducted in a lab setting (74% of publications), is based on WEIRD participants (cf. Section 3), presents study participants with a representation of a robot (67%), and employs a verbal measure (84%)—probably Likert or semantic differential scale (59%)—to study one of its determinants (71%). Based on concerns raised by scholars within and outside the field, we identify three potential problems with this picture. The first issue concerns the external validity of lab studies and experimental methods,

the second the construct validity of the heavily-relied-on verbal measures, and the third is a lack of research on the behavioral consequences of mental state attribution to robots. We discuss each of these issues in Sections 8.1.1–8.1.3.

8.1.1 External Validity of Laboratory Studies. It is not clear to what extent results on mental state attribution to robots obtained in experimental lab settings that rely on potentially ecologically invalid methods will carry over or generalize to the real world. We do not propose that results from the lab are necessarily ecologically invalid. However, the fact that we mostly do not know to what extent the reviewed findings generalize to situations in which interactions with robots (are anticipated to) take place stands as an important issue in itself [Jung and Hinds 2018]. The validity the various types of lab-based experimental methods employed to study mental state attribution to robots cannot be easily assessed on *a priori* grounds. Consequently, it is useful and/or necessary to conduct studies "in the wild" [Hutchins 1995] and compare results with those from the lab. So far, few field studies have been conducted (only 4% of reviewed publications reported on a field study), and they rarely compare obtained findings with similar findings from the lab. Moreover, only a few cross-cultural studies on mental state attribution to robots have been conducted to date. Cross-cultural studies are necessary to fully understand the generalizability of psychological research findings [Barrett 2020] and they are instrumental in systematically assessing the universality of the phenomenon.

8.1.2 Construct Validity of Self-Reports. Several researchers have independently raised concerns about the validity of self-report measures, such as questionnaires or interviews, that rely on study participants' introspective access to attribution processes. For example, Caruana and McArthur [2019] stated: "it has become necessary to assess the extent to which simulated agents (e.g., robots) induce the adoption of an intentional stance. To date, this has largely been achieved using subjective measures, such as asking participants to rate the likelihood of an agent having a mind . . . A limitation of this approach is the inherent unreliability of conscious and subjective judgments." In a similar vein, Bossi et al. [2020] suggested: "a more detailed analysis of human attitudes toward robots with objective behavioral and neural measures alongside subjective reports is necessary." These concerns extend to both general methodological limitations of self-report measures, such as the risks of various self-report biases [Nisbett and Wilson 1977; Paulhus and Vazire 2007], and limitations that are more specific to their application in the domain of human-robot interaction [Bossi et al. 2020; Caruana and McArthur 2019; de Graaf and Malle 2019; Fiala et al. 2014; Short et al. 2010; Takahashi et al. 2016; Thellman and Ziemke 2019]. Among the more specific concerns, researchers have noted that the outcomes of self-report measures vary depending on several factors that are difficult to and/or have not been controlled for in the context of the studies in which they are used. For example, Fiala et al. [2014] stated: "When we probe people for their explicit judgments about whether robots have mental states, responses are influenced by a wide variety of factors. The apparent function of the robot, the nature of the question (forced-choice vs. not), and platitudes about robots may all contribute to producing reasoned judgment about the states of robots". Thellman and Ziemke [2019] suggested that questions about the minds of robots are "ambiguously open to interpretation as regarding the reality of the mental states of robots" and that "people tend to predict and explain robot behavior with reference to mental states without reflecting on the reality of those states". In a similar vein, Short et al. [2010] reflected: "It is not easy to measure the attribution of mental state. How do we identify whether participants think of the robot as an agent reasoning about a game, rather than a machine stepping through a task? Asking "How much does the robot think?" is not sufficient. Instead, we have to rely on more subtle cues in the participants' behavior and written responses." Moreover, there is a wealth of literature showing that verbal and non-verbal measures are not necessarily consistent

[Greenwald and Banaji 1995; Nisbett and Wilson 1977] (consider, for instance, verbal responses to the question "Are you racist?" versus behavioral measures of people's racist tendencies). This applies also in the human-robot interaction context [Spatola and Wudarczyk 2021; Thellman et al. 2020]. For instance, Spatola and Wudarczyk [2021] found that the non-verbal measure employed in their study was a better predictor of a future behavior toward the robot than explicit measures.

In recognition of some of these limitations, some researchers have turned to more "subtle" or "implicit" self-report measures that avoid *direct* questions about the minds of robots. For example, Takahashi et al. [2016] proposed to use semantic differential scales in place of direct questions, de Graaf and Malle [2019] proposed a method whereby study participants provide free-text explanations of robot behavior, Levin et al. [2013] proposed to use verbal predictions of the behavior of robots, and Banks [2020] employed the various verbal theory of mind tests that relied on more indirect questions about robots' mental states. The extent to which these implicit methods—which all still rely on probing study participants' deliberate thought processes—succeed in mitigating the issues related to more direct self-report measures are unclear. Partly for this reason, other researchers have proposed to use non-verbal behavioral measures, such as anticipatory gaze [Sciutti et al. 2013; Thellman and Ziemke 2020] or attentional cueing [Wiese et al. 2019], or neurological measures. For instance, Waytz et al. [2010b] stated that neuroimaging techniques were used in their study because "the self-report measure of the dependent variables did not distinguish between whether participants were actually attributing humanlike minds to nonhuman agents or simply using the mind as a metaphoric description of their behavior". However, a limitation that has been noted in the literature regarding neurological measures is that "although neuroimaging techniques overcome some limitations of self-reported perceptions of robots' minds, they still are not sufficiently fine-grained as they document only an overall activation of regions involved in mind inference rather than inferences of specific mental states." [de Graaf and Malle 2019].

8.1.3 Behavioral Consequences of Mental State Attribution to Robots. In addition to the concerns above, self-report measures might, at least in isolation, provide little or no insight into how mental state attributions affect how people interact with robots. Arguably, investigating the behavioral effects of ascribing mental states to robots requires, as a minimal criterion, empirical observations of people's behavior and methods to ensure that the observed behavior stems from people's mental state ascriptions, i.e., from people's understanding or mental models of robots as agents with particular (ascribed) mental states and capacities. Studying the relationship between people's understanding of a robot and how they interact with it is a prerequisite for changing or improving people's understanding of the robot so that desirable interaction outcomes can be achieved. It is therefore important to develop a methodology suited for the study of the behavioral effects of ascribing particular mental states to robots in specific circumstances. Only 15% of the reviewed publications reported on behavioral consequences of mental state attribution (see Table 14 in the Appendix). Several of these studies employed various behavioral (non-verbal) measures of mental state attribution in place of the more common self-report measures. However, it is still unclear which (if any) of these methods actually accomplish to link people's attributions with the effects that they have on people's behavior in a way that is scientifically sound or valid.

It has been proposed that these effects can be studied by measuring people's predictions of robot behavior [Levin et al. 2013, 2012; Thellman 2021; Thellman and Ziemke 2019, 2020]. For example, Levin et al. [2013] stated, "Previous research exploring explicit beliefs about the intentionality of different agents has relied upon participants to rate intentions, free will, and consciousness directly [Epley, Akalis, Waytz, & Cacioppo, 2008; Gray et al., 2007; Morewedge et al., 2007], or has reported participant comments about entities such as robots [Kanda, Hirano, Eaton, & Ishiguro, 2003], without assessing whether these complex ideas are associated with expectations

about specific observable behaviors that any given entity might exhibit. We, therefore, investigated adults' understanding of different representational systems by asking participants to make predictions about the behavior of these systems." A basic assumption underlying this approach is that people's predictions of robot behavior are indicative of how people interact with robots, because deciding what actions to take in the context of interactions depends to a significant extent on the anticipation of how others will (re)act to events that they are subjected to (including one's own actions) [FeldmanHall and Shenhav 2019].

8.2 What is the Scope and Limits of People's Ability to Predict and Explain the Behavior of Robots based on Mental State Attribution?

Upon investigating perspectives in the literature on the reason why people attribute mental states to robots (RQ1B), we found that several researchers claimed that mental state attribution is important to the ability to interact with robots. A recurrent assumption was that it helps people predict and explain robot behavior. While it is generally agreed upon that people's mental state attributions are highly predictive of human behavior [e.g., Dennett 1989; Fodor 1987; Heider 1958], it is still largely unclear exactly *how* useful it is to ascribe mental states to robots. The observation that robots have a relatively simple physical and functional constitution compared to humans might lead to the assumption that it should be possible to predict their behavior based on our knowledge about their underlying physical and functional (e.g., computational) states, and that mental state ascription, therefore, should not be necessary. However, there are examples of cases where the internal states of robots and computers can be just as inscrutable as those of humans—for experts and laypeople alike. For instance, computer chess programmers are consistently beaten in chess by their own programs despite having written every line of code themselves; the space of possible internal states and legal state-transitions of the chess program is simply too vast and complex to track [Dennett 1989]. We also see that it can be virtually impossible to predict how robotic systems that interact within stochastic environments, such as autonomous vehicles, will behave in many real-life situations [Surden and Williams 2016]. Hence, it is certainly not the case that the "technological experts" necessarily fully understand robots simply because they created them. Furthermore, non-experts are even worse off considering that they will likely often have a relatively poor understanding of the internal constitution and design of the robots they might encounter in their daily life. Nevertheless, people might still be able to understand robots based on attributing their behavior to underlying mental states and predicting that robots will generally take the kinds of actions that are "rational" given those attributed states. The question is *how far* people's folk-psychological understanding will take them in interactions with robots (whose capabilities and limitations are considerably different from those of humans).

8.3 How can People's Mental State Attributions to Robots be Improved?

People's understanding of robots affect how they interact with robots and therefore what types of interaction outcomes can be achieved. Hence, it is important to explore ways to improve people's understanding of robots. This includes their understanding of robots as having particular (ascribed) mental states, such as specific beliefs and desires. To some, this might sound as a philosophical puzzle: how can one's knowledge about a mind that is seemingly projected by oneself be *improved*? However, there is a strong prima facie case for the assumption that particular mental state attributions serve us better than others in interactions with robots, and that we can improve our understanding of robot minds in various ways—including by learning about their perceptual and cognitive capabilities and limitations. For example, mistakenly thinking that a self-driving car has *seen me* and therefore *knows where I am* can arguably be detrimental to both predicting the car's behavior and taking appropriate action (and can have disastrous interaction outcomes as a

consequence). It has been suggested that the attribution of a mental state (e.g., a specific belief) to a robot can be evaluated according to whether it is *behaviorally congruent*, that is, whether it leads to accurate prediction of the robot.s behavior [Thellman 2021].

The topic of identifying ways to improve human-robot interaction through design that facilitates mental state inference that is conducive to predicting and interacting with robots has garnered some interest [e.g., Chadalavada et al. 2015; Habibovic et al. 2018; Kaptein et al. 2017; Williams et al. 2018] but so far remains largely unexplored. There is evidence that a person who repeatedly interacts with and observes the behavior of a robot is thereby able to learn to make more appropriate mental state ascriptions and consequently improve their ability to predict its behavior [Thellman and Ziemke 2020]. This suggests that people may in some circumstances be able to gradually "tune in" to the unique perceptual and cognitive capabilities over time. In other circumstances, people may need external guidance or support to attribute the appropriate mental states. This might include information communicated to people prior to the interaction (e.g., by providing a "user manual") or during the interaction itself (e.g., communicated or signaled by the robot itself).

8.4 What are the Underlying Mechanisms that Govern Mental State Attribution to Robots?

Psychological science has been accused of being preoccupied with "effects" [van Rooij and Baggio 2020]. All the findings on mental state attribution reviewed in this article, including determinants, consequences, and comparisons, can be described as behavioral effects of underlying human social-cognitive capabilities. Effects are explananda (things to explain), not explanations. In discovering an effect (e.g., that people exhibit a stronger tendency to ascribe mental states to robots than computers), we do not thereby explain it. Ultimately, it might be possible to explain observed effects with reference to underlying mechanisms (e.g., cognitive, neurological), including the extent to which the mechanisms that underly mental state attribution to robots overlap or differ from the case of humans or computers. It has been suggested that the underlying mechanisms might operate at distinct processing levels (see Section 4.3). This suggestion is compelling given empirical contradictions in the literature (cf. Section 7). However, there is so far very little knowledge of the underlying mechanisms that govern mental state attribution to robots, and it remains to be seen how such knowledge can be put to use for the practical purpose of improving human-robot interactions.

9 SUMMARY OF PRINCIPAL FINDINGS

This section highlights the most important findings related to each review question (RQ1–RQ5).

9.1 RQ1: How is Mental State Attribution to Robots *conceived* in the Scientific Literature?

In our review of the literature, we found the terminology used to describe mental state attribution to robots to be diverse but largely homogenous in usage (i.e., terms like *mind perception*, *[psychological] anthropomorphism*, *intentional stance*, and *mental state attribution* are to a significant extent used synonymously across individual studies). This suggests that researchers from various disciplines use different terms to refer to the same underlying phenomenon—a finding that we hope can facilitate cross-disciplinary dissemination of future research on the topic. We found no clear consensus among researchers regarding the reason(s) why people attribute mental states to robots or what the underlying mechanisms might be. However, a common conception is that mental state attribution helps people interact with robots by providing an interpretative framework for predicting and explaining robot behavior. Several researchers also independently suggested

that the underlying mechanisms might operate on distinct cognitive levels (in parallel and/or in sequence).

9.2 RQ2: What *methods* have been used in Studies of Mental State Attribution to Robots?

The methodological landscape is considerably diverse (cf. Tables 7–11 in the Appendix). However, a majority of studies are experimental, conducted in a laboratory setting, and employ verbal measures (76% of studies rely on questionnaire methodology) and representations of robots (as opposed to physically present robots) as a stimulus. The robots presented to study participants are most commonly anthropomorphic (i.e., humanoid) and exhibit some kind of behavior (social or non-social). Study participants can be described as mostly WEIRD (i.e., from western, educated, industrialized, rich, and democratic societies). Some of the potential limitations of the various research methods are discussed in Section 8.1.

9.3 RQ3: What are the Previous *findings* on Mental State Attribution to Robots?

Our review supports three general conclusions. Firstly, people's tendency to attribute mental states to robots is determined by multiple factors, including (but not necessarily limited to) age and motivation as well as robot behavior, appearance, and identity. Secondly, mental state attribution to robots has psychological as well as behavioral effects, and appears to facilitate prediction and explanation in interaction with robots. The consequences are so far not as well-documented as the determinants. Thirdly, people's tendency to attribute mental states to robots versus other agents follows a *computer* < *robot* < *human* pattern and appears to be moderated by the presence of socially interactive behavior. These conclusions should be considered as tentative as they are made on the basis of findings that depend on different sources of evidence (including self-report data) which may or may not be valid (cf. Section 8.1).

9.4 RQ4: Do Findings on Mental State Attribution to Robots *vary* as a Function of the Methods Employed?

We found that studies relying on different types of methodological approaches to the study of mental state attribution to robots (e.g., verbal vs. non-verbal measures) give rise to different and sometimes seemingly contradictory research findings (see Section 7). This means that the methodological choices that individual researchers make have direct consequences for what kinds of study outcomes they can expect. We hope that this finding, along with the breakdown of research methods and findings provided in Sections 5–6 and Tables 7–15 in the Appendix, can help researchers make more well-informed methodological decisions in the context of their own work. The finding also calls for a broader, interdisciplinary discussion of what research methods are appropriate for investigating mental state attribution to robots (see Section 8.1).

9.5 RQ5: What are the *open research questions* about Mental State Attribution to Robots?

We identified four open research questions: (1) What methods are appropriate for studying mental state attribution to robots? (2) What are the scope and limits of people's ability to predict and explain the behavior of robots based on mental state attribution? (3) How can people's mental state attributions to robots be improved? (4) What are the underlying mechanisms that govern mental state attribution to robots? Among these, we consider the first question to be the most pressing because it precludes the possibility to confidently assess the remaining questions. Assessing what methods are appropriate for studying mental state attribution to robots might involve adjudicating between different, contradictory sources of evidence. Alternatively, it might be possible to account

for conflicting findings in reference to the theory that can accommodate different, seemingly contradictory types of evidence simultaneously. Dual-process theories (see Section 4.3) might be a contender for such a solution as they may help explain why evidence that stems from higher cognitive processes (e.g., self-reports) may seem in contradiction with evidence that stem from lower or introspectively inaccessible cognitive processes (e.g., non-verbal behavior or neurological activity).

10 STRENGTHS AND LIMITATIONS

One of the main strengths of the reported literature review is that it employs a comprehensive and systematic method for the purpose of identifying and reviewing all available evidence relevant to the review questions (Section 2). It should be noted, as a general methodological limitation, that this method does not guarantee full coverage of all relevant primary studies (this is practically unattainable under most circumstances [Kitchenham et al. 2007]). However, the transparent and complete description of the review methods facilitates replication and a fair assessment of the basis for the conclusions drawn in the review. It should also be noted that the review was delimited to publications that described robotic systems using the term "robot". As a consequence, the review may have excluded research on mental state attribution to autonomous vehicles, drones, and other types of robotic systems that are not typically described using the term, as well as research that targeted computers, virtual characters, or animals—all of which might provide methods and insights that are relevant to and applicable in the study of mental state attribution to robots. Moreover, the review was delimited to empirical publications only. While we hope to have identified most findings and empirical methods, this means that some of the conceptions of the phenomenon that are present in the non-empirical literature might shed additional insights to the discussion presented in this review.

Another methodological limitation is that we decided not to conduct a quality assessment of the primary studies included in the review (cf. Section 2.3). Aside from the practical infeasibility due to the large variety of methods, we think that in this case, such assessment would risk to prematurely discounting individual studies based on their particular approach of study, and thereby limit the possibility of a systematic and unbiased review. Whereas most of the primary studies were sourced from well-established scientific journals and conference proceedings, the lack of a detailed quality assessment means that these studies are presented in the review as if they all are of equal quality, and that therefore the general conclusions drawn from them might be biased. Consequently, the burden is placed on the reader to judge the validity of included primary studies as well as the general review conclusions drawn from them. This emphasizes the importance of a transparent and detailed description of the review methods so that the reader can easily trace and understand the rationale behind the review conclusions.

The review also contributes comprehensive lookup tables (included in the Appendix) that can be used by researchers interested in specific types of methods or findings about mental state attribution to robots to find related research. This can potentially lead to better-informed methodological choices and promote appropriate (and demote redundant) replication of research findings—in short, to a more cumulative science.

11 CONCLUSION

What prompted us to conduct this comprehensive review of conceptions, methods, and findings regarding human mental state attribution to robots is the need for a shared scientific understanding of this complex phenomenon within the interdisciplinary HRI community. Among other things, we found that the terminology used is diverse, but also to some degree homogeneous or convergent in the sense that "mental state attribution/ascription" is the by far most used term. This term has the

advantage, in our opinion [cf. Thellman and Ziemke 2021], that—unlike terms like "mind reading" or "mind perception"—it makes it relatively clear that the mental states in question are ascribed by people, and not necessarily real. We hope that this article contributes to facilitating a broader dissemination among researchers in different disciplines as well as further convergence among different strands of research.

Our systematic review found that the literature abounds with conflicting research findings that stem from different sources of evidence, including self-reports, non-verbal behavioral and neurological data. This means that the methodological choices that individual researchers make—including the selection of stimulus materials (e.g., behavioral vs. non-behavioral) and measures (e.g., verbal vs. non-verbal behavioral or neurological)—have direct consequences for what research outcomes can be expected. This indicates a clear need for an inter- and transdisciplinary discussion about the nature of the phenomenon and what research methods are appropriate. In particular, we believe that there is a need for researchers to develop a common language and a shared set of basic assumptions about the phenomenon at hand, in order to be able to access and build cumulatively on each other's work. We hope that our review findings can inform or at least inspire a broader discussion about these issues.

It should be noted that there are significant practical incentives for studying how people's mental state attributions affect how they interact with robots. Researchers have begun to explore ways of improving human-robot interaction through design that facilitates mental state inference that is conducive to achieving desirable outcomes in human-robot interactions (e.g., collaboration, accident prevention). At the present, the largest obstacle is the lack of appropriate research methods for obtaining such knowledge. We believe the best way to advance the field in terms of methodology is through scientific discussions that transcend disciplinary boundaries.

Last, but not least, we conclude that the role of mental state attribution in interactions with robots is actually still surprisingly unclear. The scope and limits of people's capability to interact (socially) with robots based on attributions of their behavior to underlying mental states, such as beliefs, feelings, and intentions, are still largely unknown. As a consequence, possible ways to improve human-robot interactions by guiding people's mental state attributions—and managing their expectations [cf. Ziemke 2020]—remain largely unexplored so far. These issues are bound to become increasingly relevant as robots become more ubiquitous in the daily lives of people and, as illustrated with the pedestrian example in the introduction, need to be understandable and interactable for a very broad range of people.

APPENDIX

Table 2. Publications Included in the Review (N = 155)

Author (study #)	Year	N	Population	Age*	% fem/wom	Setting
[Abubshait and Wiese 2017]	2017	63	US students	21	75	Lab
[Abubshait and Wykowska 2020]	2020	27		24.4	63	Lab
[Abubshait et al. 2020]	2020	109	University students	18–37	63	Lab
[Akechi et al. 2018]	2018	78	Children with and without autism	7–24	26	Lab
[Alimardani and Qurashi 2019]	2019	53	Young and elderly Dutch natives			Lab
[Appel et al. 2020] (1)	2020	44	US and German adult residents		44	Online
[Appel et al. 2016]	2016	93	US residents	34.6	52	Online
[Banks 2020]	2020	469	White/caucasian majority	22.3	57	Online
[Bartneck and Hu 2008] (2)	2008	25	Dutch students	19–25	40	Lab
[Bartneck et al. 2007]	2007	49	Dutch	24.6	33	Lab
[Beran and Ramirez-Serrano 2010]	2010	184	Canadian children	8.2	53	Lab
[Bernotat and Eyssel 2018]	2018	102	German and Japanese adults		54	Online
[Bernstein and Crowley 2008]	2008	60	Children	4–7	48	Lab
[Bossi et al. 2020]	2020	52	Adults			Lab
[Brink et al. 2019]	2019	240	Children	3–18	49	Field, Lab
[Broadbent et al. 2013]	2013	30	University students and staff	22.5	47	Lab
[Buckwalter and Phelan 2013]	2013	253	US			Online
[Carter and Pelphrey 2006]	2006	9	Children	9.2	40	Lab
[Carter et al. 2011]	2011	17	Adults	27.5	47	Lab
[Chaminade et al. 2018]	2018	21	French speakers			Lab
[Chaminade et al. 2012]	2012	19	University students	21.5	0	Lab
[Ciardo et al. 2020]	2020	90		25.9	65	Lab
[Cross et al. 2019]	2019	26	UK university students	19.9	68	Lab
[Dang and Liu 2021]	2021	379	Chinese, US		41	Online
[de Graaf and Malle 2019]	2019	121		38.5	51	Online
[Desideri et al. 2021]	2020	94	Primary school children		47	Lab
[Di Dio et al. 2018]	2018	37	Italian children	7	43	Lab
[Di Dio et al. 2020]	2020	31	Italian children	5.9	42	Lab
[Eimler et al. 2011]	2011	211	US university students, German		40	Online
[Eyssel and Pfundmair 2015]	2015	68	German	23.7	49	Lab
[Eyssel and Kuchenbrandt 2012]	2012	78	German university students	23.3	51	Lab
[Eyssel and Reich 2013]	2013	34	German university students	29.1		Lab
[Eyssel et al. 2012]	2012	58	Students	23	53	Lab
[Eyssel et al. 2010]	2010	31	German	22.6		Lab
[Eyssel et al. 2011]	2011	58	German university students	25.2	52	Lab
[Eyssel et al. 2017]	2017	81	German	25.6	49	Lab
[Fiala et al. 2014]	2014	52				
[Fiore et al. 2013]	2013	74	US	19.2	50	Lab
[Fraune et al. 2020]	2020	599	US and Japanese		35	Online, Lab
[Fu et al. 2021]	2020	24	University students	25.3		Lab
[Giusti and Marti 2006]	2006	5	Elderly	74.8	80	Field
[Gobbini et al. 2011]	2011	12		26	67	Lab
[Gray et al. 2007]	2007	2399	Majority white	30.5	66	Online
[Gray and Wegner 2012]	2012	120	US	25	36	Lab
[Hannibal 2014]	2014	14	Danish schoolchildren			Field
[Haring et al. 2015]	2015	42	Australian and Japanese		43	Lab
[Haring et al. 2019]	2019	35	US	17–24		Lab
[Heijnen et al. 2019]	2019	54	Dutch university student majority	22.3	65	Lab
[Henkel et al. 2017]	2017	30	US	8–12	60	Lab
[Hoenen et al. 2016]	2016	57		23.2	74	Lab
[Hofree et al. 2014]	2014	595	US		75	Lab
[Holbrook 2018]	2018	855	US	18–74	50	Online
[Huebner 2010]	2010	2014	US university students			Lab
[Imamura et al. 2015]	2015	22		31	55	Lab
[Ishii and Watanabe 2019]	2019	392	Japanese university students	19.5	53	Lab
[Itakura et al. 2008]	2008	50	Toddlers	2.6	52	Lab
[Ito et al. 2004]	2004					Lab
[Jipson and Gelman 2007]	2007	72	Children and adults		50	Lab
[Kahn Jr et al. 2012]	2012	40	University students	20.3	52	Lab
[Kamewari et al. 2005]	2005	32	Infants	0.5	52	Lab
[Kamide et al. 2013]	2013	1200	Japanese	38.4	50	Online

*Numbers in column "Age" are means unless marked as age span.

Table 3. Publications Included in the Review (N = 155) (Continued)

Author (study #)	Year	N	Population	Age*	% fem/wom	Setting
[Keijsers and Bartneck 2018]	2018	302	Majority US residents	38	57	Lab
[Keijsers et al. 2021]	2019	217	Majority US residents		59	Online, Lab
[Kiesler et al. 2008]	2008	113	US	26	48	Lab
[Konijn and Hoorn 2020]	2020	265		31.5	53	Online, Lab
[Korman et al. 2019]	2019	82				Online
[Krach et al. 2008]	2008	20		24.5	0	Lab
[Kupferberg et al. 2013]	2013	43	Monkeys (common marmosets)	1–11		Lab
[Kupferberg et al. 2018]	2018	20	German	26.6	50	Lab
[Küster and Swiderska 2021]	2021	253		38.4	64	Online
[Lakatos et al. 2014] (1)	2014	48		24.5	46	Lab
[Law et al. 2021]	2020	84	University students			Lab
[Lee et al. 2021]	2021	428	US resident majority		37	Online
[Lee et al. 2005]	2005	108	Chinese university students		69	Lab
[Lefkeli et al. 2021]	2020	245	University population	22.4	62	Lab
[Lemaignan et al. 2015]	2015	26	Children	4.5		Lab
[Levin et al. 2013]	2013	102	US university students, hospital staff		64	Lab
[Levin et al. 2012]	2012	81	US university students		62	Lab
[Mahzoon et al. 2019]	2019	32	Japanese university students		47	Lab
[Mandell et al. 2017]	2017	160	English speakers	18–77	53	Online
[Manzi et al. 2020]	2020	144	Italian children	5–9	49	Lab
[Marchesi et al. 2019]	2019	106		33.3	68	Online
[Marchesi et al. 2021]	2020	41		24.8	61	Lab
[Martin et al. 2020]	2020	40	Children, anglo-australian majority	3.6	43	Lab
[Martini et al. 2016]	2016	318	English speakers, white majority		56	Online
[Martini et al. 2015]	2015	7	US university students	22.1	86	Lab
[Melson et al. 2009]	2009	152	Children	3–15		Lab
[Miyake et al. 2019]	2019	25	Japanese	21.4	25	Lab
[Morewedge et al. 2007] (2)	2007	63	University students	29.8	25	Lab
[Mou et al. 2020]	2020	32	University population		53	Lab
[Müller et al. 2020]	2020	127		17-24	86	Lab
[Mutlu et al. 2009]	2009	26	Japanese university students	20.4	65	Lab
[Nigam and Klahr 2000]	2000	39	Children			Lab
[Nijssen et al. 2019]	2019	135	Dutch university students		88	Lab
[Okanda et al. 2019]	2019	48	Japanese university students		54	Lab
[Okanda et al. 2021]	2021	95	Children and adults	3–62		Lab
[Okita et al. 2005]	2005	93	Children	3–5		Lab
[Özdem et al. 2017]	2017	21		18–28		Lab
[Paetzel et al. 2018]	2018	46	University students	26.2	24	Lab
[Peressini 2014]	2014	73	University students		53	Lab
[Perez-Osorio et al. 2019]	2019	44		24.8	55	Lab
[Powers et al. 2005]	2005	33	Native English speakers	21	48	Lab
[Quadflieg et al. 2016]	2016	265	University students and staff	18–54	54	Online
[Raffard et al. 2016]	2016	38	Adults with and without schizophrenia	18–55		Lab
[Rueben et al. 2021]	2021	6	University population		50	Field
[Saerbeck and Bartneck 2010]	2010	18	Adult		44	Lab
[Salem et al. 2011]	2011	62	German native speakers	30.9	52	Lab
[Saylor and Levin 2005]	2005	11	Children	4.6	64	Lab
[Sciutti et al. 2013]	2013	10		31	20	Lab
[Severson and Lemm 2016]	2016	90	Children	5–9	50	Lab
[Short et al. 2010]	2010	60	US student majority		53	Lab
[Sirkin et al. 2015]	2015	20	US university students	20.8	60	Lab
[Somanader et al. 2011]	2011	66	Children	4–5	52	Lab
[Sommer et al. 2019]	2019	126	Children, caucasian majority	7.6	50	Lab
[Spektor-Precel and Mioduser 2015]	2015	24	Children	4–6		Lab
[Stafford et al. 2014]	2014	25	Older adults	86.1	72	Field
[Straub 2016]	2016		Café visitors			Field
[Sturgeon et al. 2019]	2019	60	College educated majority	20–79	40	Online
[Subrahmanyam et al. 2002]	2002	48	Children and adults			Lab
[Swiderska and Küster 2018]	2018	217		22.3	62	Online
[Sytsma and Machery 2010]	2010	1135	Adults	18–75		Online
[Takahashi et al. 2016]	2016	500	Japanese	45	50	Online
[Takahashi et al. 2013]	2013	46	Adults	18–36	41	Lab
[Takahashi et al. 2014]	2014	16		18–25	69	Lab

*Numbers in column "Age" are means unless marked as age span.

Table 4. Publications Included in the Review (N = 155) (Continued)

Author (study #)	Year	N	Population	Age*	% fem/wom	Setting
[Tan et al. 2018]	2018	775	US and Chinese adults		45	Online
[Tanibe et al. 2017]	2017	129	Japanese speakers	21.8	50	Online
[Tatsukawa et al. 2019]	2019	84	Japanese			Lab
[Terada and Yamada 2017]	2017	110	Japanese university students	19–25	41	Online
[Terada et al. 2007]	2007	32	University students and staff	20–27		Lab
[Thellman and Ziemke 2020]	2020	155	Swedish university students	25		Lab
[Thellman et al. 2017]	2017	90	Swedish university students	24	52	Lab
[Trovato and Eyssel 2017]	2017	66	Italian and Japanese students		40	Lab
[van den Berghe et al. 2021]	2021	104	Children	5.7	48	Lab
[van der Woerdt and Haselager 2019]	2019	63	University population		33	Online
[van Duuren and Scaife 1996]	1996	80	Children and adults	5–45		Lab
[van Straten et al. 2020]	2020	144	Primary school children	8.9	50	Lab
[Van Straten et al. 2021]	2021	168	Primary school children	9	56	Lab
[Wallkötter et al. 2020]	2020	62		38.7		Online
[Wang and Krumhuber 2018]	2018	443			45	Lab, Online
[Wang and Quadflieg 2015]	2015	26	English speakers	18–35	54	Lab
[Ward et al. 2013]	2013	121		34.9	66	Online
[Waytz et al. 2010b] (2–6)	2010	215	Adults		46	Lab
[Weisman et al. 2017]	2017	1442	US adults			Online
[Wiese et al. 2019] (1)	2019	114	University students		68	Lab
[Wiese et al. 2021]	2021	142	English speakers	35.3	49	Online
[Wiese et al. 2012]	2012	70		18–32	70	Lab
[Wykowska et al. 2014]	2014	44		19–34	55	Lab
[Xie et al. 2019]	2019	400		39	48	Online
[Xu and Sar 2018]	2018	522			44	Online
[Yamaji et al. 2010]	2010	108	Children	4–11		Field
[Zhang et al. 2019]	2019	40	Children with and without autism	5–8	13	Lab
[Zhao et al. 2016]	2016	1648			57	Online
[Złotowski et al. 2014]	2014	35	Majority university population	26.7	54	Lab
[Złotowski et al. 2017]	2017	52	Japanese-speaking university students	21.5	35	Lab
[Złotowski et al. 2018]	2018	40	Japanese-speaking university population	21.5	35	Lab

*Numbers in column "Age" are means unless marked as age span.

Table 5. Terms used in Reviewed Studies to Denote Mental State Attribution to Robots

Term	Study
Anthropomorphism	[Alimardani and Qurashi 2019; Banks 2020; Bernotat and Eyssel 2018; Bossi et al. 2020; Broadbent et al. 2013; Chaminade et al. 2012; Desideri et al. 2021; Di Dio et al. 2018; Eyssel et al. 2012, 2010; Eyssel and Kuchenbrandt 2012; Eyssel et al. 2011; Eyssel and Reich 2013; Eyssel et al. 2017; Eyssel and Pfundmair 2015; Fraune et al. 2020; Hannibal 2014; Heijnen et al. 2019; Hofree et al. 2014; Holbrook 2018; Huebner 2010; Kamide et al. 2013; Keijsers and Bartneck 2018; Keijsers et al. 2021; Kiesler et al. 2008; Krach et al. 2008; Küster and Swiderska 2021; Lakatos et al. 2014; Law et al. 2021; Lemaignan et al. 2015; Levin et al. 2013; Manzi et al. 2020; Marchesi et al. 2019, 2021; Melson et al. 2009; Müller et al. 2020; Nijssen et al. 2019; Okanda et al. 2021; Salem et al. 2011; Severson and Lemm 2016; Short et al. 2010; Sommer et al. 2019; Stafford et al. 2014; Takahashi et al. 2014; Tan et al. 2018; Terada et al. 2007; Trovato and Eyssel 2017; van den Berghe et al. 2021; van der Woerdt and Haselager 2019; van Straten et al. 2020; Van Straten et al. 2021; Wallkötter et al. 2020; Wang and Krumhuber 2018; Waytz et al. 2010b; Zhao et al. 2016; Złotowski et al. 2014, 2017, 2018]
Folk psychology	[Carter et al. 2011; de Graaf and Malle 2019; Fiala et al. 2014; Huebner 2010; Kiesler et al. 2008; Peressini 2014; Sytsma and Machery 2010; Thellman et al. 2017; Thellman and Ziemke 2020; Weisman et al. 2017]
Intentional stance	[Abubshait and Wykowska 2020; Bossi et al. 2020; Carter and Pelphrey 2006; Chaminade et al. 2018, 2012; Ciardo et al. 2020; Desideri et al. 2021; Huebner 2010; Krach et al. 2008; Lee et al. 2021; Mandell et al. 2017; Marchesi et al. 2019, 2021; Martini et al. 2016, 2015; Özdem et al. 2017; Perez-Osorio et al. 2019; Rueben et al. 2021; Terada et al. 2007; Terada and Yamada 2017; Thellman et al. 2017; Thellman and Ziemke 2020; Wallkötter et al. 2020; Wiese et al. 2019, 2012; Wykowska et al. 2014; Yamaji et al. 2010]
Mental state attribution/ ascription	[Abubshait et al. 2020; Abubshait and Wiese 2017; Abubshait and Wykowska 2020; Akechi et al. 2018; Alimardani and Qurashi 2019; Appel et al. 2020; Banks 2020; Beran and Ramirez-Serrano 2010; Bernotat and Eyssel 2018; Bernstein and Crowley 2008; Bossi et al. 2020; Brink et al. 2019; Buckwalter and Phelan 2013; Carter et al. 2011; Carter and Pelphrey 2006; Cross et al. 2019; Dang and Liu 2021; de Graaf and Malle 2019; Di Dio et al. 2018, 2020; Eimler et al. 2011; Eyssel et al. 2012, 2010; Eyssel and Kuchenbrandt 2012; Eyssel et al. 2011; Eyssel and Reich 2013; Eyssel et al. 2017; Eyssel and Pfundmair 2015; Fiala et al. 2014; Fiore et al. 2013; Fu et al. 2021; Giusti and Marti 2006; Gobbini et al. 2011; Gray and Wegner 2012; Hannibal 2014; Haring et al. 2019; Heijnen et al. 2019; Hoenen et al. 2016; Holbrook 2018; Huebner 2010; Imamura et al. 2015; Ishii and Watanabe 2019; Itakura et al. 2008; Ito et al. 2004; Jipson and Gelman 2007; Kahn Jr et al. 2012; Kamewari et al. 2005; Kamide et al. 2013; Keijsers and Bartneck 2018; Keijsers et al. 2021; Konijn and Hoorn 2020; Korman et al. 2019; Krach et al. 2008; Kupferberg et al. 2013, 2018; Küster and Swiderska 2021; Lakatos et al. 2014; Law et al. 2021; Lee et al. 2021, 2005; Lefkeli et al. 2021; Lemaignan et al. 2015; Levin et al. 2013; Mahzoon et al. 2019; Mandell et al. 2017; Marchesi et al. 2019, 2021; Martin et al. 2020; Martini et al. 2016, 2015; Melson et al. 2009; Miyake et al. 2019; Mou et al. 2020; Müller et al. 2020; Mutlu et al. 2009; Nigam and Klahr 2000; Nijssen et al. 2019; Okanda et al. 2019, 2021; Okita et al. 2005; Özdem et al. 2017; Paetzel et al. 2018; Peressini 2014; Perez-Osorio et al. 2019; Quadflieg et al. 2016; Rueben et al. 2021; Saerbeck and Bartneck 2010; Salem et al. 2011; Saylor and Levin 2005; Sciutti et al. 2013; Severson and Lemm 2016; Short et al. 2010; Sirkin et al. 2015; Somanader et al. 2011; Sommer et al. 2019; Spektor-Precel and Mioduser 2015; Stafford et al. 2014; Straub 2016; Subrahmanyam et al. 2002; Swiderska and Küster 2018; Sytsma and Machery 2010; Takahashi et al. 2016, 2013, 2014; Tan et al. 2018; Tanibe et al. 2017; Tatsukawa et al. 2019; Terada et al. 2007; Terada and Yamada 2017; Thellman et al. 2017; Thellman and Ziemke 2020; Trovato and Eyssel 2017; van den Berghe et al. 2021; van der Woerdt and Haselager 2019; van Duuren and Scaife 1996; van Straten et al. 2020; Van Straten et al. 2021; Wallkötter et al. 2020; Wang and Krumhuber 2018; Wang and Quadflieg 2015; Ward et al. 2013; Waytz et al. 2010b; Weisman et al. 2017; Wiese et al. 2019, 2021, 2012; Wykowska et al. 2014; Xie et al. 2019; Xu and Sar 2018; Yamaji et al. 2010; Zhang et al. 2019; Zhao et al. 2016; Złotowski et al. 2014, 2017, 2018]
Mentalizing	[Abubshait et al. 2020; Abubshait and Wiese 2017; Abubshait and Wykowska 2020; Banks 2020; Bossi et al. 2020; Carter et al. 2011; Carter and Pelphrey 2006; Chaminade et al. 2012; Ciardo et al. 2020; Cross et al. 2019; Desideri et al. 2021; Di Dio et al. 2018; Fiore et al. 2013; Gobbini et al. 2011; Kamewari et al. 2005; Krach et al. 2008; Mandell et al. 2017; Marchesi et al. 2021; Martini et al. 2016, 2015; Okanda et al. 2021; Özdem et al. 2017; Sturgeon et al. 2019; Takahashi et al. 2013, 2014; Wang and Quadflieg 2015; Waytz et al. 2010b; Wiese et al. 2019, 2012; Wykowska et al. 2014; Zhao et al. 2016]
Mind perception	[Abubshait et al. 2020; Abubshait and Wiese 2017; Akechi et al. 2018; Alimardani and Qurashi 2019; Appel et al. 2020; Bernotat and Eyssel 2018; Brink et al. 2019; Broadbent et al. 2013; Dang and Liu 2021; Desideri et al. 2021; Di Dio et al. 2018; Eyssel and Kuchenbrandt 2012; Eyssel and Reich 2013; Eyssel et al. 2017; Eyssel and Pfundmair 2015; Fiore et al. 2013; Fu et al. 2021; Gray et al. 2007; Gray and Wegner 2012; Haring et al. 2015; Holbrook 2018; Ishii and Watanabe 2019; Kamide et al. 2013; Keijsers and Bartneck 2018; Küster and Swiderska 2021; Lee et al. 2021; Lefkeli et al. 2021; Mahzoon et al. 2019; Mandell et al. 2017; Manzi et al. 2020; Martini et al. 2016; Miyake et al. 2019; Müller et al. 2020; Okanda et al. 2021; Quadflieg et al. 2016; Raffard et al. 2016; Stafford et al. 2014; Swiderska and Küster 2018; Sytsma and Machery 2010; Takahashi et al. 2016; Tanibe et al. 2017; Tatsukawa et al. 2019; Trovato and Eyssel 2017; van der Woerdt and Haselager 2019; Wallkötter et al. 2020; Wang and Krumhuber 2018; Ward et al. 2013; Waytz et al. 2010b; Weisman et al. 2017; Wiese et al. 2019, 2021; Xu and Sar 2018]
Mind reading	[Carter and Pelphrey 2006; Imamura et al. 2015; Ito et al. 2004; Kupferberg et al. 2018; Lee et al. 2021; Takahashi et al. 2014; Terada and Yamada 2017]
Theory of mind	[Akechi et al. 2018; Alimardani and Qurashi 2019; Banks 2020; Broadbent et al. 2013; Carter and Pelphrey 2006; de Graaf and Malle 2019; Di Dio et al. 2018, 2020; Fiore et al. 2013; Gobbini et al. 2011; Holbrook 2018; Ishii and Watanabe 2019; Itakura et al. 2008; Kamide et al. 2013; Keijsers and Bartneck 2018; Korman et al. 2019; Krach et al. 2008; Law et al. 2021; Lee et al. 2021; Lefkeli et al. 2021; Levin et al. 2012; Mandell et al. 2017; Marchesi et al. 2019; Martini et al. 2015; Mou et al. 2020; Mutlu et al. 2009; Spektor-Precel and Mioduser 2015; Stafford et al. 2014; Sturgeon et al. 2019; Terada et al. 2007; Terada and Yamada 2017; Waytz et al. 2010b; Weisman et al. 2017; Wiese et al. 2012; Zhang et al. 2019]

Table 6. Examples of Statements in Which Various Terms (boldfaced) are used to Denote Mental State Attribution to Robots

Statement	Study
"From mundane conversations across dinner tables to strategic negotiations in poker games and first dates, people automatically and purposefully **mentalize** other agents. That is, they engage in meta-representational sense-making associated with inferring others' mental states, and those processes are core to human social interaction [1]. These processes are collectively known as **theory of mind** (ToM), a system by which one **ascribes mental states** to self and other and then uses that comparative ascription to make predictions about others' behaviors".	[Banks 2020]
"In this way, the NMSPI prompts participants to make **mental state attributions**, that is, to engage in **theory of mind mentalizing**"	[Fiore et al. 2013]
"One of the biases that intriguing cognitive scientist and psychologist most would be **Theory of Mind**, the cognitive capacity to **ascribe mental states** to others, animals, and even non-living entities"	[Ishii and Watanabe 2019]
"Emery [21] suggests that people combine information from gaze cues with "higher-order cognitive strategies (including experience and empathy) to determine that an individual is attending to a particular stimulus because they intend to do something with the object, or believe something about the object"—an ability called "**mental state attribution**" or "**theory of mind**"."	[Mutlu et al. 2009]
"Research in social cognitive neuroscience has demonstrated that when we interact with others we often make inferences about the others' internal states (i.e., intentions, beliefs) in order to explain, understand, and predict their behavior—a process commonly referred to as **mentalizing** ... The belief that an agent has a mind intuitively or consciously triggers the adoption of the **intentional stance** (Dennett, 2003), which involves treating the agent as a rational being with beliefs, desires, and action goals ..."	[Özdem et al. 2017]
"... two mechanisms have been proposed to underlie intention understanding. One of these is a **folk psychology** or '**theory theory of mind**' that has been advanced as an alternate method of intention understanding that uses mental state information."	[Carter et al. 2011]
The human capability to **mentalize**, also defined as "**Theory of Mind**" (ToM), has been studied for 40 years as a socio-cognitive function that enables individuals to think about others' mental states, such as thoughts, intentions, motivations, desires, and emotions underlying behavior (Tomasello, 1999; see also, Frith & Frith, 1999). Through the **attribution of states of mind**, humans can predict and eventually manipulate others' thoughts and actions (for a review, see Waytz, Gray, Epley, & Wegner, 2010)	[Di Dio et al. 2018]
The MPFC, the TPJ, and the anterior temporal cortex are the major components of the **mentalizing** or **ToM** system	[Gobbini et al. 2011]
"The **theory of mind perception** is related to **anthropomorphism**, in that people **attribute capacities of mind** to non- human characters."	[Stafford et al. 2014]
"**Theory of Mind (ToM)** or **mentalizing** refers to the ability to make inferences about the thoughts, beliefs, or intentions of another individual"	[Sturgeon et al. 2019]
"Taken together, the present activity modulation observed in this set of brain regions likely reflects the participants' **mentalizing** processes employed to **read** opponents' mental states"	[Takahashi et al. 2014]
"The optimal way to cope with this type of intelligent agent, which has behavioral variability in both competitive and cooperative situations, is to **attribute abstract mental states** to it as the causes of its behavior, as in **mind-reading** (Whiten, 1996), a **theory of mind** (Premack and Woodruff, 1978), or an **intentional stance** (Dennett, 1987)"	[Terada and Yamada 2017]

Table 7. Stimuli Employed in Reviewed Studies of Mental State Attribution to Robots

Stimulus	Study
Robot presentation	
Text	[Appel et al. 2020, 2016; Dang and Liu 2021; de Graaf and Malle 2019; Fiala et al. 2014; Peressini 2014; Takahashi et al. 2016; Tanibe et al. 2017; Ward et al. 2013; Waytz et al. 2010b]
Image	[Abubshait et al. 2020; Abubshait and Wiese 2017; Bernotat and Eyssel 2018; Bernstein and Crowley 2008; Di Dio et al. 2018, 2020; Eimler et al. 2011; Eyssel and Kuchenbrandt 2012; Eyssel and Reich 2013; Mandell et al. 2017; Manzi et al. 2020; Martin et al. 2020; Martini et al. 2016, 2015; Müller et al. 2020; Nigam and Klahr 2000; Nijssen et al. 2019; Okanda et al. 2019; Quadflieg et al. 2016; Raffard et al. 2016; Saylor and Levin 2005; Subrahmanyam et al. 2002; Swiderska and Küster 2018; Trovato and Eyssel 2017; Wiese et al. 2019, 2021, 2012; Wykowska et al. 2014; Zhao et al. 2016]
Video	[Brink et al. 2019; Cross et al. 2019; Eyssel et al. 2012; Fraune et al. 2020; Gobbini et al. 2011; Gray and Wegner 2012; Hoenen et al. 2016; Hofree et al. 2014; Holbrook 2018; Itakura et al. 2008; Jipson and Gelman 2007; Kamewari et al. 2005; Konijn and Hoorn 2020; Korman et al. 2019; Kupferberg et al. 2013; Lefkeli et al. 2021; Morewedge et al. 2007; Mou et al. 2020; Sommer et al. 2019; Sturgeon et al. 2019; Tan et al. 2018; Terada and Yamada 2017; Thellman and Ziemke 2020; van der Woerdt and Haselager 2019; Xu and Sar 2018; Zhao et al. 2016]
Animation	[Carter et al. 2011; Carter and Pelphrey 2006; Keijsers and Bartneck 2018; Keijsers et al. 2021; Kiesler et al. 2008; Özdem et al. 2017; Paetzel et al. 2018]
Audio	[Banks 2020]
Image + Text	[Akechi et al. 2018; Banks 2020; Bossi et al. 2020; Buckwalter and Phelan 2013; Gray et al. 2007; Huebner 2010; Ishii and Watanabe 2019; Kamide et al. 2013; Küster and Swiderska 2021; Lee et al. 2021; Levin et al. 2013, 2012; Mahzoon et al. 2019; Marchesi et al. 2019, 2021; Miyake et al. 2019; Perez-Osorio et al. 2019; Sytsma and Machery 2010; Thellman et al. 2017; Wang and Krumhuber 2018; Wang and Quadflieg 2015; Weisman et al. 2017; Xie et al. 2019]
Video + Text	[Eyssel et al. 2011; Lee et al. 2005]
Video + Image + Text	[Levin et al. 2013]
Telepresent	[Chaminade et al. 2018, 2012; Kiesler et al. 2008; Krach et al. 2008; Takahashi et al. 2014]
Physically present	[Abubshait and Wykowska 2020; Alimardani and Qurashi 2019; Bartneck and Hu 2008; Bartneck et al. 2007; Beran and Ramirez-Serrano 2010; Broadbent et al. 2013; Ciardo et al. 2020; Cross et al. 2019; Desideri et al. 2021; Eyssel et al. 2010, 2017; Eyssel and Pfundmair 2015; Fiore et al. 2013; Fraune et al. 2020; Fu et al. 2021; Giusti and Marti 2006; Hannibal 2014; Haring et al. 2019, 2015; Heijnen et al. 2019; Henkel et al. 2017; Hofree et al. 2014; Imamura et al. 2015; Ito et al. 2004; Kahn Jr et al. 2012; Keijsers et al. 2021; Kiesler et al. 2008; Kupferberg et al. 2018; Lakatos et al. 2014; Law et al. 2021; Lemaignan et al. 2015; Martin et al. 2020; Melson et al. 2009; Mou et al. 2020; Mutlu et al. 2009; Okanda et al. 2021; Okita et al. 2005; Powers et al. 2005; Rueben et al. 2021; Saerbeck and Bartneck 2010; Salem et al. 2011; Sciutti et al. 2013; Severson and Lemm 2016; Short et al. 2010; Sirkin et al. 2015; Somanader et al. 2011; Spektor-Precel and Mioduser 2015; Stafford et al. 2014; Straub 2016; Takahashi et al. 2013, 2014; Tatsukawa et al. 2019; Terada et al. 2007; van den Berghe et al. 2021; van Duuren and Scaife 1996; van Straten et al. 2020; Van Straten et al. 2021; Wallkötter et al. 2020; Yamaji et al. 2010; Zhang et al. 2019; Złotowski et al. 2014, 2017, 2018]

Note: Studies that employ multiple types of stimuli may be placed in more than one category.

Table 8. Stimuli Employed in Reviewed Studies of Mental State Attribution to Robots (Continued)

Stimulus	Study
Robot morphology	
Anthropomorphic	[Abubshait et al. 2020; Abubshait and Wiese 2017; Abubshait and Wykowska 2020; Akechi et al. 2018; Alimardani and Qurashi 2019; Banks 2020; Bernotat and Eyssel 2018; Bernstein and Crowley 2008; Bossi et al. 2020; Brink et al. 2019; Broadbent et al. 2013; Buckwalter and Phelan 2013; Carter et al. 2011; Carter and Pelphrey 2006; Chaminade et al. 2018; Desideri et al. 2021; Di Dio et al. 2018, 2020; Eyssel et al. 2012; Eyssel and Kuchenbrandt 2012; Eyssel et al. 2011; Eyssel and Reich 2013; Eyssel et al. 2017; Eyssel and Pfundmair 2015; Fu et al. 2021; Gobbini et al. 2011; Gray et al. 2007; Gray and Wegner 2012; Hannibal 2014; Haring et al. 2019, 2015; Heijnen et al. 2019; Henkel et al. 2017; Hofree et al. 2014; Holbrook 2018; Ishii and Watanabe 2019; Itakura et al. 2008; Ito et al. 2004; Kahn Jr et al. 2012; Kamewari et al. 2005; Kamide et al. 2013; Keijsers and Bartneck 2018; Keijsers et al. 2021; Kiesler et al. 2008; Konijn and Hoorn 2020; Korman et al. 2019; Krach et al. 2008; Kupferberg et al. 2018; Küster and Swiderska 2021; Lee et al. 2021, 2005; Levin et al. 2013, 2012; Mandell et al. 2017; Manzi et al. 2020; Marchesi et al. 2019, 2021; Martin et al. 2020; Martini et al. 2016, 2015; Miyake et al. 2019; Morewedge et al. 2007; Mou et al. 2020; Müller et al. 2020; Mutlu et al. 2009; Nijssen et al. 2019; Okanda et al. 2019, 2021; Özdem et al. 2017; Paetzel et al. 2018; Peressini 2014; Perez-Osorio et al. 2019; Powers et al. 2005; Quadflieg et al. 2016; Raffard et al. 2016; Salem et al. 2011; Saylor and Levin 2005; Sciutti et al. 2013; Short et al. 2010; Somanader et al. 2011; Sommer et al. 2019; Straub 2016; Sturgeon et al. 2019; Swiderska and Küster 2018; Takahashi et al. 2013, 2014; Tan et al. 2018; Tatsukawa et al. 2019; Terada and Yamada 2017; Thellman et al. 2017; Thellman and Ziemke 2020; Trovato and Eyssel 2017; van den Berghe et al. 2021; van der Woerdt and Haselager 2019; van Duuren and Scaife 1996; van Straten et al. 2020; Van Straten et al. 2021; Wallkötter et al. 2020; Wang and Krumhuber 2018; Wang and Quadflieg 2015; Weisman et al. 2017; Wiese et al. 2019, 2021, 2012; Wykowska et al. 2014; Xu and Sar 2018; Zhang et al. 2019; Zhao et al. 2016; Złotowski et al. 2014, 2017, 2018]
Functional	[Banks 2020; Bartneck and Hu 2008; Beran and Ramirez-Serrano 2010; Bernstein and Crowley 2008; Brink et al. 2019; Broadbent et al. 2013; Ciardo et al. 2020; Cross et al. 2019; Fiore et al. 2013; Fraune et al. 2020; Hoenen et al. 2016; Imamura et al. 2015; Kamide et al. 2013; Krach et al. 2008; Law et al. 2021; Lefkeli et al. 2021; Lemaignan et al. 2015; Mahzoon et al. 2019; Martini et al. 2015; Müller et al. 2020; Nijssen et al. 2019; Peressini 2014; Rueben et al. 2021; Saerbeck and Bartneck 2010; Sirkin et al. 2015; Stafford et al. 2014; Sytsma and Machery 2010; Takahashi et al. 2014; Tan et al. 2018; Terada et al. 2007; Terada and Yamada 2017; Xie et al. 2019; Xu and Sar 2018; Yamaji et al. 2010]
Zoomorphic	[Bartneck et al. 2007; Eimler et al. 2011; Eyssel et al. 2010; Giusti and Marti 2006; Jipson and Gelman 2007; Kupferberg et al. 2013; Lakatos et al. 2014; Melson et al. 2009; Okanda et al. 2019; Okita et al. 2005; Saerbeck and Bartneck 2010; Saylor and Levin 2005; Severson and Lemm 2016; Sommer et al. 2019; Tan et al. 2018; Terada and Yamada 2017; Xu and Sar 2018]
Not applicable	[Appel et al. 2020, 2016; Dang and Liu 2021; de Graaf and Malle 2019; Fiala et al. 2014; Takahashi et al. 2016; Tanibe et al. 2017; Wang and Krumhuber 2018; Ward et al. 2013; Waytz et al. 2010b]

Note: Studies that employ multiple types of stimuli may be placed in more than one category.

Table 9. Stimuli Employed in Reviewed Studies of Mental State Attribution to Robots (Continued)

Stimulus	Study
Robot behavior	
Social (participant interacton)	[Bartneck and Hu 2008; Bartneck et al. 2007; Broadbent et al. 2013; Chaminade et al. 2018, 2012; Ciardo et al. 2020; Cross et al. 2019; Desideri et al. 2021; Eyssel et al. 2010, 2017; Eyssel and Pfundmair 2015; Fiore et al. 2013; Fraune et al. 2020; Giusti and Marti 2006; Hannibal 2014; Haring et al. 2019, 2015; Heijnen et al. 2019; Henkel et al. 2017; Imamura et al. 2015; Ito et al. 2004; Kahn Jr et al. 2012; Keijsers and Bartneck 2018; Keijsers et al. 2021; Kiesler et al. 2008; Krach et al. 2008; Lakatos et al. 2014; Law et al. 2021; Lefkeli et al. 2021; Lemaignan et al. 2015; Mahzoon et al. 2019; Martin et al. 2020; Melson et al. 2009; Miyake et al. 2019; Mou et al. 2020; Mutlu et al. 2009; Okanda et al. 2021; Okita et al. 2005; Powers et al. 2005; Rueben et al. 2021; Salem et al. 2011; Severson and Lemm 2016; Short et al. 2010; Sirkin et al. 2015; Stafford et al. 2014; Straub 2016; Takahashi et al. 2013, 2014; Terada et al. 2007; Terada and Yamada 2017; van den Berghe et al. 2021; van Straten et al. 2020; Van Straten et al. 2021; Wallkötter et al. 2020; Waytz et al. 2010b; Yamaji et al. 2010; Złotowski et al. 2014, 2017, 2018]
Social (interaction with non-participant other)	[Alimardani and Qurashi 2019; Banks 2020; Bossi et al. 2020; de Graaf and Malle 2019; Fraune et al. 2020; Fu et al. 2021; Haring et al. 2015; Holbrook 2018; Jipson and Gelman 2007; Konijn and Hoorn 2020; Korman et al. 2019; Lee et al. 2021, 2005; Marchesi et al. 2019, 2021; Morewedge et al. 2007; Okita et al. 2005; Paetzel et al. 2018; Perez-Osorio et al. 2019; Quadflieg et al. 2016; Sturgeon et al. 2019; Tan et al. 2018; Thellman et al. 2017; Wang and Quadflieg 2015; Waytz et al. 2010b]
Non-social	[Abubshait and Wykowska 2020; Beran and Ramirez-Serrano 2010; Bossi et al. 2020; Brink et al. 2019; Buckwalter and Phelan 2013; Carter et al. 2011; Carter and Pelphrey 2006; Cross et al. 2019; de Graaf and Malle 2019; Eimler et al. 2011; Eyssel et al. 2012, 2011; Fraune et al. 2020; Gobbini et al. 2011; Gray and Wegner 2012; Hoenen et al. 2016; Hofree et al. 2014; Itakura et al. 2008; Kamewari et al. 2005; Kupferberg et al. 2013, 2018; Levin et al. 2013, 2012; Marchesi et al. 2019; Özdem et al. 2017; Perez-Osorio et al. 2019; Saerbeck and Bartneck 2010; Sciutti et al. 2013; Somanader et al. 2011; Spektor-Precel and Mioduser 2015; Tan et al. 2018; Tatsukawa et al. 2019; Thellman et al. 2017; Thellman and Ziemke 2020; van der Woerdt and Haselager 2019; van Duuren and Scaife 1996; Waytz et al. 2010b; Wiese et al. 2019, 2021, 2012; Wykowska et al. 2014; Zhang et al. 2019; Zhao et al. 2016]
No behavior	[Abubshait et al. 2020; Abubshait and Wiese 2017; Akechi et al. 2018; Appel et al. 2020, 2016; Bernotat and Eyssel 2018; Bernstein and Crowley 2008; Dang and Liu 2021; Di Dio et al. 2018, 2020; Eyssel and Kuchenbrandt 2012; Eyssel and Reich 2013; Fiala et al. 2014; Gray et al. 2007; Huebner 2010; Ishii and Watanabe 2019; Kamide et al. 2013; Küster and Swiderska 2021; Mandell et al. 2017; Manzi et al. 2020; Martin et al. 2020; Martini et al. 2016, 2015; Müller et al. 2020; Nigam and Klahr 2000; Nijssen et al. 2019; Okanda et al. 2019; Peressini 2014; Raffard et al. 2016; Saylor and Levin 2005; Sommer et al. 2019; Subrahmanyam et al. 2002; Swiderska and Küster 2018; Sytsma and Machery 2010; Takahashi et al. 2016; Tanibe et al. 2017; Trovato and Eyssel 2017; Wang and Krumhuber 2018; Ward et al. 2013; Weisman et al. 2017; Xie et al. 2019]

Note: Studies that employ multiple types of stimuli may be placed in more than one category.

Table 10. Measures Employed in Studies of Mental State Attribution to Robots

Data type, **Operationalization**, Measurement tool	**Study**
Verbal	
Judged possession of mind/mental capacities/mental states	
Likert/semantic differential scale	[Abubshait et al. 2020; Abubshait and Wiese 2017; Abubshait and Wykowska 2020; Akechi et al. 2018; Alimardani and Qurashi 2019; Appel et al. 2020, 2016; Banks 2020; Bernotat and Eyssel 2018; Bossi et al. 2020; Brink et al. 2019; Broadbent et al. 2013; Buckwalter and Phelan 2013; Cross et al. 2019; Dang and Liu 2021; Eimler et al. 2011; Eyssel et al. 2012, 2010; Eyssel and Kuchenbrandt 2012; Eyssel et al. 2011; Eyssel and Reich 2013; Eyssel et al. 2017; Eyssel and Pfundmair 2015; Fraune et al. 2020; Fu et al. 2021; Gray et al. 2007; Gray and Wegner 2012; Haring et al. 2019; Heijnen et al. 2019; Hofree et al. 2014; Holbrook 2018; Huebner 2010; Imamura et al. 2015; Ishii and Watanabe 2019; Kamide et al. 2013; Keijsers and Bartneck 2018; Keijsers et al. 2021; Kiesler et al. 2008; Konijn and Hoorn 2020; Korman et al. 2019; Krach et al. 2008; Küster and Swiderska 2021; Lakatos et al. 2014; Law et al. 2021; Lee et al. 2021, 2005; Lefkeli et al. 2021; Mahzoon et al. 2019; Mandell et al. 2017; Manzi et al. 2020; Marchesi et al. 2019, 2021; Martin et al. 2020; Martini et al. 2016, 2015; Miyake et al. 2019; Morewedge et al. 2007; Mou et al. 2020; Müller et al. 2020; Mutlu et al. 2009; Nijssen et al. 2019; Okanda et al. 2021; Peressini 2014; Perez-Osorio et al. 2019; Quadflieg et al. 2016; Raffard et al. 2016; Salem et al. 2011; Severson and Lemm 2016; Stafford et al. 2014; Sturgeon et al. 2019; Swiderska and Küster 2018; Sytsma and Machery 2010; Takahashi et al. 2016, 2014; Tan et al. 2018; Tanibe et al. 2017; Tatsukawa et al. 2019; Terada and Yamada 2017; Thellman et al. 2017; Trovato and Eyssel 2017; van der Woerdt and Haselager 2019; van Straten et al. 2020; Van Straten et al. 2021; Wallkötter et al. 2020; Wang and Krumhuber 2018; Ward et al. 2013; Waytz et al. 2010b; Weisman et al. 2017; Wiese et al. 2019; Xie et al. 2019; Xu and Sar 2018; Złotowski et al. 2014, 2017, 2018]
Binary choice	[Banks 2020; Bernstein and Crowley 2008; Brink et al. 2019; Di Dio et al. 2018, 2020; Kahn Jr et al. 2012; Mandell et al. 2017; Melson et al. 2009; Nigam and Klahr 2000; Okanda et al. 2019; Okita et al. 2005; Saylor and Levin 2005; Somanader et al. 2011; Spektor-Precel and Mioduser 2015; Subrahmanyam et al. 2002; Terada et al. 2007; van den Berghe et al. 2021; van Duuren and Scaife 1996]
Multiple choice	[Fiala et al. 2014; Lakatos et al. 2014; Paetzel et al. 2018]
Graphic scale	[Fiore et al. 2013; Saerbeck and Bartneck 2010; Sommer et al. 2019]
Free text/speech	[Henkel et al. 2017]
Interview	[Beran and Ramirez-Serrano 2010; Jipson and Gelman 2007; Lemaignan et al. 2015; Sirkin et al. 2015; Subrahmanyam et al. 2002]
Mentalistic description of robot	
Free speech	[Giusti and Marti 2006; Short et al. 2010; Straub 2016]
Interview	[Hannibal 2014; Rueben et al. 2021]
Mentalistic explanation of robot behavior	
Likert/semantic differential scale	[Banks 2020]
Free text	[Banks 2020; de Graaf and Malle 2019; Korman et al. 2019]
Interview	[Rueben et al. 2021]
Mentalistic prediction of robot behavior	
Likert/semantic differential scale	[Banks 2020]
Binary choice	[Banks 2020; Levin et al. 2013, 2012; Zhang et al. 2019]
Interview	[Rueben et al. 2021]

Table 11. Measures Employed in Studies of Mental State Attribution to Robots (Continued)

Data type, **Operationalization**, Measurement tool	**Study**
Behavioral (non-verbal)	
Level of robot abuse	[Bartneck and Hu 2008]
Tendency to switch off robot	[Bartneck et al. 2007]
Sense of agency in interaction with robot	[Ciardo et al. 2020]
Tendency to avoid eye contact with robot	[Desideri et al. 2021]
Presence of joint Simon effect	[Heijnen et al. 2019]
Altruistic behavior in competition against a robot in the dictator game	[Heijnen et al. 2019]
Imitation of goal-directed robot behavior	[Itakura et al. 2008]
Teaching task completion time	[Ito et al. 2004]
Level of attention toward robot	[Ito et al. 2004]
Tendency to respond to robot questions	[Ito et al. 2004]
Violation of expectations about robot behavior	[Kamewari et al. 2005]
Preferential looking time	[Kupferberg et al. 2013]
Tendency to address robot using polite speech	[Lemaignan et al. 2015]
Tendency to communicate with robot using social gestures	[Lemaignan et al. 2015]
Time taken to accept mentalistic description of behavior	[Marchesi et al. 2021]
Task performance in a guessing game where a robot exhibited non-verbal leakage	[Mutlu et al. 2009]
Anticipatory gaze toward end-of-motion-position of robot performing a goal-directed action	[Sciutti et al. 2013]
Presence of non-verbal social behavior in interaction with robot	[Straub 2016]
Randomness of decision making in a competitive game played against a robot	[Takahashi et al. 2013]
Anticipatory gaze toward end-of-motion-position of robot performing a belief-directed action	[Thellman and Ziemke 2020]
Use of mixed or exploitative strategies in a competitive game against a robot	[Terada and Yamada 2017]
Tendency to take the visual perspective of a robot	[Zhao et al. 2016]
Response in visual priming task	[Złotowski et al. 2018]
Tendency to help robot achieve a goal	[Martin et al. 2020; Yamaji et al. 2010]
*Neurological**	
Activation in the inferior parietal lobule (IPL)	[Kupferberg et al. 2018]
Activation in the pain matrix	[Cross et al. 2019]
Activation in the posterior paracingulate cortex (PCC)	[Takahashi et al. 2014]
Activation in the premotor cortex (PMC)	[Kupferberg et al. 2018]
Activation in the precuneus (PrC)	[Wang and Quadflieg 2015]
Activation in the superior temporal gyrus (STG)	[Chaminade et al. 2018]
Activation in the superior temporal sulcus (STS)	[Carter et al. 2011; Carter and Pelphrey 2006; Chaminade et al. 2018]
Activation in the temporoparietal junction (TPJ)	[Chaminade et al. 2012; Gobbini et al. 2011; Krach et al. 2008; Özdem et al. 2017; Takahashi et al. 2014; Wang and Quadflieg 2015]
Activation in the medial prefrontal cortex (mPFC)	[Chaminade et al. 2012; Gobbini et al. 2011; Krach et al. 2008; Takahashi et al. 2014; Wang and Quadflieg 2015; Waytz et al. 2010b]
Mu-activation in the mirror neuron system (MNS)	[Hoenen et al. 2016]
Resting state gamma activation	[Bossi et al. 2020]

*All reported neurological data were collected using functional magnetic resonance imaging (fMRI) as measurement tool except electroencephalographic (EGG) data reported in Bossi et al. [2020] and Hoenen et al. [2016].

Table 12. Reported Findings on Human-Factor Determinants of Mental State Attribution to Robots (+, effect; −, no effect)

Determinant	Study
Human factors	
Age	
(+) Observed tendency in children	[Beran and Ramirez-Serrano 2010; Bernstein and Crowley 2008; Brink et al. 2019; Cross et al. 2019; Di Dio et al. 2018, 2020; Hannibal 2014; Henkel et al. 2017; Itakura et al. 2008; Jipson and Gelman 2007; Lemaignan et al. 2015; Levin et al. 2013; Manzi et al. 2020; Martin et al. 2020; Melson et al. 2009; Nigam and Klahr 2000; Okanda et al. 2021; Okita et al. 2005; Saylor and Levin 2005; Severson and Lemm 2016; Somanader et al. 2011; Spektor-Precel and Mioduser 2015; Subrahmanyam et al. 2002; van den Berghe et al. 2021; van Duuren and Scaife 1996]
(+) Stronger tendency in younger than older children	[Di Dio et al. 2018, 2020; Manzi et al. 2020; Okanda et al. 2021; Okita et al. 2005; Severson and Lemm 2016; Somanader et al. 2011]
(+) Stronger tendency in older than younger children	[van Duuren and Scaife 1996]
(+) Stronger tendency in children than adults	[Jipson and Gelman 2007; Okanda et al. 2021; Subrahmanyam et al. 2002]
(+) Observed tendency in infants	[Kamewari et al. 2005]
(+) Observed tendency in older adults	[Giusti and Marti 2006; Stafford et al. 2014; Subrahmanyam et al. 2002]
(+) Stronger tendency in older than younger adults	[Levin et al. 2013]
(−) Similar tendency in younger and older adults	[Alimardani and Qurashi 2019; Tan et al. 2018]
Cultural and socioeconomic background	
(+) Stronger tendency in Chinese than US	[Tan et al. 2018]
(+) Stronger tendency in Japanese than Germans	[Bernotat and Eyssel 2018]
(+) Stronger tendency in Japanese than Australians	[Haring et al. 2015]
(+) Stronger tendency in Japanese than Westerners	[Takahashi et al. 2016]
(+) Stronger tendency in Japanese than Italians	[Trovato and Eyssel 2017]
(−) Similar tendency in Japanese and Westerners	[Ishii and Watanabe 2019; Kamide et al. 2013]
(−) Similar tendency in people with different socioeconomic background	[Marchesi et al. 2019]
Gender	
(−) Similar tendency in men and women	[Raffard et al. 2016; Saerbeck and Bartneck 2010; Tan et al. 2018]
Interaction history	
(+) Stronger tendency when informed about a robot's interaction history	[Fu et al. 2021; Mahzoon et al. 2019]
(+) Stronger tendency with repeated interactions	[Fiore et al. 2013]
(+) Weaker tendency with reapeated interactions	[Abubshait and Wykowska 2020]
(−) No effect of interacting with robot	[Cross et al. 2019; van den Berghe et al. 2021]
Mental disorder	
(+) Stronger tendency in children without autism as compared with autism	[Zhang et al. 2019]
(−) Similar tendency in children with and without autism	[Akechi et al. 2018]
(−) Similar tendency in adults with and without schizophrenia	[Raffard et al. 2016]
Motivation	
(+) Stronger tendency when motivated to predict robot behavior	[Waytz et al. 2010b]
(+) Stronger tendency when anticipating future interaction with robot	[Eyssel et al. 2011]
(+) Stronger tendency when lonely	[Eyssel and Reich 2013]
(+) Stronger tendency when believing that robot is controlled by human	[Özdem et al. 2017]
(+) Stronger tendency when having high expectations about robot capabilities	[Perez-Osorio et al. 2019]
(+) Weaker tendency when informed about robot's capabilities	[van Straten et al. 2020; Van Straten et al. 2021]
(+) Stronger tendency when robot is perceived as being mistreated or subjected to harm	[Hoenen et al. 2016; Konijn and Hoorn 2020; Küster and Swiderska 2021; Ward et al. 2013]
(+) Stronger tendency when robot is perceived as being helped as opposed to treated neutrally	[Tanibe et al. 2017]
(+) Stronger tendency when loosing as compared to winning in a cooperative task with a robot	[Lefkeli et al. 2021]
(+) Stronger tendency when winning as compared to loosing against a robot	[Lefkeli et al. 2021]
(−) Similar tendency when motivated to reason about robot as when not	[Złotowski et al. 2018]
(−) Similar tendency when socially excluded or included	[Eyssel and Pfundmair 2015]
Species	
(−) Similar tendency in monkeys as in humans	[Kupferberg et al. 2013]

Table 13. Reported Findings on Robot-Factor Determinants of Mental State Attribution to Robots (+, effect; –, no effect)

Determinant	Study
Robot factors	
Appearance	
(+) Stronger tendency with (increasingly) human-like appearance	[Abubshait et al. 2020; Banks 2020; Broadbent et al. 2013; Krach et al. 2008; Manzi et al. 2020; Martini et al. 2016, 2015; Takahashi et al. 2014; Xu and Sar 2018]
(–) Similar tendency with (increasingly) human-like appearance	[Saerbeck and Bartneck 2010; Złotowski et al. 2017]
(+) Stronger tendency with increased amount of physical features	[Martini et al. 2015]
(+) Observed differences in states attributed depending on physical features	[Eimler et al. 2011; Paetzel et al. 2018; Terada et al. 2007]
(+) Stronger tendency when face visible	[Gray and Wegner 2012]
(+) Stronger tendency with facial wounds	[Swiderska and Küster 2018]
(+) Weaker tendency with cues of violent conflict present	[Holbrook 2018]
Behavior	
(+) Stronger tendency when exhibiting gaze behavior	[Abubshait and Wiese 2017; Ito et al. 2004; Levin et al. 2013; Takahashi et al. 2013; Zhao et al. 2016]
(+) Stronger tendency when exhibiting intelligent behavior	[Sturgeon et al. 2019]
(+) Stronger tendency when exhibiting unpredictable behavior	[Eyssel et al. 2011; Waytz et al. 2010b]
(+) Observed differences in states attributed depending on type of behavior	[Lakatos et al. 2014; Mutlu et al. 2009; Thellman and Ziemke 2020]
(+) Stronger tendency when exhibiting reactive behavior	[Terada et al. 2007]
(+) Stronger tendency when exhibiting complex behavior	[Imamura et al. 2015]
(+) Stronger tendency when exhibiting higher behavioral variability	[Terada and Yamada 2017]
(+) Stronger when exhibiting emotional behavior	[Złotowski et al. 2014, 2018]
(+) Stronger tendency when exhibiting social behavior	[Fraune et al. 2020; Straub 2016]
(–) Similar tendency when exhibiting social and non-social behavior	[Wallkötter et al. 2020]
(+) Stronger tendency when exhibiting cheating behavior	[Short et al. 2010]
(+) Stronger tendency when exhibiting gestures	[Salem et al. 2011]
(+) Stronger tendency when responding to user behavior with emotional facial expressions	[Eyssel et al. 2010]
(+) Stronger tendency when exhibiting reaching behavior	[Zhao et al. 2016]
(+) Weaker tendency when exhibiting norm-violating behavior	[Korman et al. 2019]
(+) Stronger tendency when moving	[Sirkin et al. 2015]
(+) Stronger tendency when moving with human-like speed	[Morewedge et al. 2007]
(+) Stronger tendency when moving toward displaced object	[Yamaji et al. 2010]
(+) Stronger tendency when moving fast	[Saerbeck and Bartneck 2010]
(+) Stronger tendency when moving slowly	[Fiore et al. 2013]
(+) Stronger tendency when exhibiting positively velenced movement	[Law et al. 2021]
(–) Similar tendency when disclosing personal information, thoughts, and feelings as when not	[Eyssel et al. 2017]
(–) Similar tendency when moving in synchrony with person	[Heijnen et al. 2019]
Capability	
(+) Stronger tendency with human-like traits of imagination	[Tatsukawa et al. 2019]
(+) Stronger tendency when a robot fails due to lack of effort relative to lack of ability	[van der Woerdt and Haselager 2019]
(+) Observed difficulties in attributing states to robots with different-from-human capabilities	[Thellman and Ziemke 2020]
(–) Similar tendency when described as having mental capabilities as when not	[Wallkötter et al. 2020]
Identity	
(+) Observed differences in attributed states depending on robot function or purpose	[Buckwalter and Phelan 2013; Dang and Liu 2021; Terada et al. 2007; Wang and Krumhuber 2018]
(+) Observed differences in attributed states depending on robot language and described country of origin	[Lee et al. 2005]
(+) Observed differences in attributed states depending on perceived robot gender	[Powers et al. 2005]
(+) Stronger tendency with in-group robot name and country of origin	[Eyssel and Kuchenbrandt 2012]
(+) Stronger tendency with in-group gendered robot voice	[Eyssel et al. 2012]
Presence	
(+) Stronger tendency when physically present than telepresent	[Kiesler et al. 2008; Straub 2016]

Table 14. Reported Findings on Consequences of Mental State Attribution to Robots (+, effect; −, no effect)

Consequence	Study
Psychological	
(+) Increased eeriness/uncanniness of robot	[Appel et al. 2020, 2016; Gray and Wegner 2012]
(+) Decreased eeriness/uncanniness of robot	[Brink et al. 2019; Quadflieg et al. 2016]
(−) No effect on uncanniness of robot	[Wang and Quadflieg 2015]
(+) Increased perceived predictability	[Waytz et al. 2010b]
(+) Increased trust in robot	[Mou et al. 2020; Xie et al. 2019]
(+) Increased drainage of cognitive resources	[Mandell et al. 2017; Wiese et al. 2019]
(+) Increased perceived threat of damage to humans and human identity	[Müller et al. 2020]
(+) Increased moral concern for robot	[Nijssen et al. 2019; Sommer et al. 2019]
(+) Increased ambivalence in attitudes toward robots	[Dang and Liu 2021]
(+) Reduced sense of agency	[Ciardo et al. 2020]
Behavioral	
(+) Observed ability to predict robot behavior	[Banks 2020; Levin et al. 2013, 2012; Rueben et al. 2021; Sciutti et al. 2013; Thellman and Ziemke 2020; Zhang et al. 2019]
(+) Observed difficulty to predict robot behavior	[Rueben et al. 2021; Thellman and Ziemke 2020]
(+) Observed ability to explain robot behavior	[Banks 2020; de Graaf and Malle 2019; Korman et al. 2019; Rueben et al. 2021]
(+) Observed difficulty to explain robot behavior	[Rueben et al. 2021]
(+) Stronger tendency to attend to robot gaze behavior	[Abubshait and Wykowska 2020; Wiese et al. 2012; Wykowska et al. 2014]
(+) Stronger tendency to avert from eye contact with a robot	[Desideri et al. 2021]
(+) Observed differences in the tendency to attend to robot gaze behavior	[Abubshait et al. 2020]
(+) Observed differences in the task assigned to a robot depending on ascribed mental capability	[Wiese et al. 2021]
(+) Decreased abuse against robot	[Bartneck and Hu 2008; Keijsers and Bartneck 2018; Keijsers et al. 2021]
(−) No effect on abuse against robot	[Keijsers et al. 2021]
(+) Hesitation to switch off robot	[Bartneck et al. 2007]
(+) Increased tendency to help robot	[Martin et al. 2020]
(+) Less likely to use robot	[Stafford et al. 2014]
(+) Observed differences in actions taken when negotiating with a robot	[Lee et al. 2021]
(−) No effect on the tendency to mimic robot facial expressions	[Hofree et al. 2014]

Table 15. Reported Findings from Comparative Studies of Mental State Attribution to Robots and Other Types of Agents (+, effect; –, no effect)

Contrast	Study
Human vs. robot	
(+) Stronger tendency toward human	[Abubshait and Wiese 2017; Banks 2020; Bernstein and Crowley 2008; Chaminade et al. 2018, 2012; Cross et al. 2019; de Graaf and Malle 2019; Di Dio et al. 2018, 2020; Fiala et al. 2014; Gobbini et al. 2011; Haring et al. 2019; Kahn Jr et al. 2012; Konijn and Hoorn 2020; Krach et al. 2008; Küster and Swiderska 2021; Levin et al. 2013; Mandell et al. 2017; Manzi et al. 2020; Marchesi et al. 2021; Miyake et al. 2019; Nijssen et al. 2019; Saylor and Levin 2005; Somanader et al. 2011; Swiderska and Küster 2018; Takahashi et al. 2013, 2014; Terada and Yamada 2017; van Duuren and Scaife 1996; Wang and Quadflieg 2015; Weisman et al. 2017; Xu and Sar 2018; Zhao et al. 2016]
(–) Similar tendency	[Carter et al. 2011; Carter and Pelphrey 2006; Kamewari et al. 2005; Kupferberg et al. 2018; Quadflieg et al. 2016; Sciutti et al. 2013; Thellman et al. 2017]
(+) Stronger tendency to attribute experience-related mental states to human	[Gray et al. 2007; Gray and Wegner 2012; Huebner 2010; Ishii and Watanabe 2019; Okanda et al. 2019; Peressini 2014]
(–) Similar tendency to attribute agency-related mental states	[Gray et al. 2007; Gray and Wegner 2012; Huebner 2010; Ishii and Watanabe 2019; Okanda et al. 2019; Peressini 2014]
(–) Similar types of attributed mental states	[Banks 2020; de Graaf and Malle 2019; Lee et al. 2005; Thellman et al. 2017]
(+) Stronger tendency to attribute valenced mental states to humans	[Sytsma and Machery 2010]
(+) Stronger performance in recognizing emotions in humans than robots	[Paetzel et al. 2018]
Human vs. robot (moderator)	
(+) Presence of gaze behavior increased tendency toward robot relative to human	[Abubshait and Wiese 2017; Levin et al. 2013]
(+) Longer response time increased tendency toward robot relative to human	[Levin et al. 2012]
(+) Tendency toward robot relative to human increased over time	[Abubshait and Wiese 2017]
Computer vs. robot	
(+) Weaker tendency toward computer	[Bernstein and Crowley 2008; Krach et al. 2008; Miyake et al. 2019; Takahashi et al. 2014; van Duuren and Scaife 1996]
Computer vs. robot (moderator)	
(+) Presence of gaze behavior increased tendency toward robot relative to computer	[Levin et al. 2013]

REFERENCES

Abdulaziz Abubshait, Ali Momen, and Eva Wiese. 2020. Pre-exposure to ambiguous faces modulates top-down control of attentional orienting to counterpredictive gaze cues. *Frontiers in Psychology* 11 (2020), 2234.

Abdulaziz Abubshait and Eva Wiese. 2017. You look human, but act like a machine: Agent appearance and behavior modulate different aspects of human–robot interaction. *Frontiers in Psychology* 8 (2017), 1393.

Abdulaziz Abubshait and Agnieszka Wykowska. 2020. Repetitive robot behavior impacts perception of intentionality and gaze-related attentional orienting. *Frontiers in Robotics and AI* 7 (2020), 150.

Hironori Akechi, Yukiko Kikuchi, Yoshikuni Tojo, Koichiro Hakarino, and Toshikazu Hasegawa. 2018. Mind perception and moral judgment in autism. *Autism Research* 11, 9 (2018), 1239–1244.

Maryam Alimardani and Sonia Qurashi. 2019. Mind perception of a sociable humanoid robot: A comparison between elderly and young adults. In *Proceedings of the Iberian Robotics Conference*. Springer, 96–108.

Markus Appel, David Izydorczyk, Silvana Weber, Martina Mara, and Tanja Lischetzke. 2020. The uncanny of mind in a machine: Humanoid robots as tools, agents, and experiencers. *Computers in Human Behavior* 102 (2020), 274–286.

Markus Appel, Silvana Weber, Stefan Krause, and Martina Mara. 2016. On the eeriness of service robots with emotional capabilities. In *Proceedings of the 2016 11th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 411–412.

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

Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind"? *Cognition* 21, 1 (1985), 37–46.

H. Clark Barrett. 2020. Towards a cognitive science of the human: Cross-cultural approaches and their urgency. *Trends in Cognitive Sciences* 24, 8 (2020), 620–638.

Christoph Bartneck and Jun Hu. 2008. Exploring the abuse of robots. *Interaction Studies* 9, 3 (2008), 415–433.

Christoph Bartneck, Michel Van Der Hoek, Omar Mubin, and Abdullah Al Mahmud. 2007. "Daisy, daisy, give me your answer do!" switching off a robot. In *Proceedings of the 2007 2nd ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 217–222.

Anthony W. Bateman and Peter Ed Fonagy. 2012. *Handbook of mentalizing in mental health practice*. American Psychiatric Publishing, Inc.

Tanya Beran and Alejandro Ramirez-Serrano. 2010. Do children perceive robots as alive? Children's attributions of human characteristics. In *Proceedings of the 2010 5th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 137–138.

Richard A. Bernardi. 2006. Associations between Hofstede's cultural constructs and social desirability response bias. *Journal of Business Ethics* 65, 1 (2006), 43–53.

Jasmin Bernotat and Friederike Eyssel. 2018. Can('t) wait to have a robot at home? Japanese and German users' attitudes toward service robots in smart homes. In *Proceedings of the 2018 27th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 15–22.

Debra Bernstein and Kevin Crowley. 2008. Searching for signs of intelligent life: An investigation of young children's beliefs about robot intelligence. *The Journal of the Learning Sciences* 17, 2 (2008), 225–247.

Francesco Bossi, Cesco Willemse, Jacopo Cavazza, Serena Marchesi, Vittorio Murino, and Agnieszka Wykowska. 2020. The human brain reveals resting state activity patterns that are predictive of biases inattitudes toward robots. *Science Robotics* 5, 46, Article eabb6652 (2020).

Kimberly A. Brink, Kurt Gray, and Henry M. Wellman. 2019. Creepiness creeps in: Uncanny valley feelings are acquired in childhood. *Child Development* 90, 4 (2019), 1202–1214.

Elizabeth Broadbent, Vinayak Kumar, Xingyan Li, John Sollers 3rd, Rebecca Q. Stafford, Bruce A. MacDonald, and Daniel M. Wegner. 2013. Robots with display screens: A robot with a more humanlike face display is perceived to have more mind and a better personality. *PLOS ONE* 8, 8 (2013), e72589.

Martin Brüne, Mona Abdel-Hamid, Caroline Lehmkämper, and Claudia Sonntag. 2007. Mental state attribution, neurocognitive functioning, and psychopathology: What predicts poor social competence in schizophrenia best? *Schizophrenia Research* 92, 1–3 (2007), 151–159.

Wesley Buckwalter and Mark Phelan. 2013. Function and feeling machines: A defense of the philosophical conception of subjective experience. *Philosophical Studies* 166, 2 (2013), 349–361.

Barry Buzan. 1993. From international system to international society: Structural realism and regime theory meet the English school. *International Organization* 47, 3 (1993), 327–352.

Elizabeth J. Carter, Jessica K. Hodgins, and David H. Rakison. 2011. Exploring the neural correlates of goal-directed action and intention understanding. *Neuroimage* 54, 2 (2011), 1634–1642.

Elizabeth J. Carter and Kevin A. Pelphrey. 2006. School-aged children exhibit domain-specific responses to biological motion. *Social Neuroscience* 1, 3–4 (2006), 396–411.

Nathan Caruana and Genevieve McArthur. 2019. The mind minds minds: The effect of intentional stance on the neural encoding of joint attention. *Cognitive, Affective, & Behavioral Neuroscience* 19, 6 (2019), 1479–1491.

Ravi Teja Chadalavada, Henrik Andreasson, Robert Krug, and Achim Lilienthal. 2015. That's on my mind! robot to human intention communication through on-board projection on shared floor space. In *Proceedings of the 2015 European Conference on Mobile Robots*. IEEE, 1–6.

Thierry Chaminade, Birgit Rauchbauer, Bruno Nazarian, Morgane Bourhis, Magalie Ochs, and Laurent Prévot. 2018. Investigating the dimensions of conversational agents' social competence using objective neurophysiological measurements. In *Proceedings of the 20th International Conference on Multimodal Interaction: Adjunct*. 1–7.

Thierry Chaminade, Delphine Rosset, David Da Fonseca, Bruno Nazarian, Ewald Lutscher, Gordon Cheng, and Christine Deruelle. 2012. How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. *Frontiers in Human Neuroscience* 6 (2012), 103.

Francesca Ciardo, Frederike Beyer, Davide De Tommaso, and Agnieszka Wykowska. 2020. Attribution of intentional agency towards robots reduces one's own sense of agency. *Cognition* 194 (2020), 104109.

Emily S. Cross, Katie A. Riddoch, Jaydan Pratts, Simon Titone, Bishakha Chaudhury, and Ruud Hortensius. 2019. A neurocognitive investigation of the impact of socializing with a robot on empathy for pain. *Philosophical Transactions of the Royal Society B* 374, 1771 (2019), 1–13.

Gergely Csibra and György Gergely. 2007. "Obsessed with goals": Functions and mechanisms of teleological interpretation of actions in humans. *Acta Psychologica* 124, 1 (2007), 60–78.

Jianning Dang and Li Liu. 2021. Robots are friends as well as foes: Ambivalent attitudes toward mindful and mindless AI robots in the United States and China. *Computers in Human Behavior* 115 (2021), 106612.

Maartje de Graaf and Bertram Malle. 2019. People's explanations of robot behavior subtly reveal mental state inferences. In *Proceedings of the 2019 ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 239–248.

Daniel Dennett. 1989. *The Intentional Stance*. MIT Press.

Lorenzo Desideri, Paola Bonifacci, Giulia Croati, Angelica Dalena, Maria Gesualdo, Gianfelice Molinario, Arianna Gherardini, Lisa Cesario, and Cristina Ottaviani. 2021. The mind in the machine: Mind perception modulates gaze aversion during child–robot interaction. *International Journal of Social Robotics* 13, 4 (2021), 599–614.

Cinzia Di Dio, Sara Isernia, Chiara Ceolaro, Antonella Marchetti, and Davide Massaro. 2018. Growing up thinking of God's beliefs: Theory of mind and ontological knowledge. *Sage Open* 8, 4 (2018), 2158244018809874.

Cinzia Di Dio, Federico Manzi, S. Itakura, Takayuki Kanda, Hiroshi Ishiguro, Davide Massaro, and Antonella Marchetti. 2020. It does not matter who you are: Fairness in pre-schoolers interacting with human and robotic partners. *International Journal of Social Robotics* 12, 5 (2020), 1045–1059.

Brian Duffy. 2003. Anthropomorphism and the social robot. *Robotics and autonomous systems* 42, 3–4 (2003), 177–190.

Sabrina C. Eimler, Nicole C. Krämer, and Astrid M. von der Pütten. 2011. Empirical results on determinants of acceptance and emotion attribution in confrontation with a robot rabbit. *Applied Artificial Intelligence* 25, 6 (2011), 503–529.

Nicholas Epley, Adam Waytz, and John T. Cacioppo. 2007. On seeing human: A three-factor theory of anthropomorphism. *Psychological Review* 114, 4 (2007), 864–886.

Jonathan Evans. 2003. In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences* 7, 10 (2003), 454–459.

Friederike Eyssel, Laura De Ruiter, Dieta Kuchenbrandt, Simon Bobinger, and Frank Hegel. 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. In *Proceedings of the 2012 7th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 125–126.

Friederike Eyssel, Frank Hegel, Gernot Horstmann, and Claudia Wagner. 2010. Anthropomorphic inferences from emotional nonverbal cues: A case study. In *Proceedings of the 2010 19th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 646–651.

Friederike Eyssel and Dieta Kuchenbrandt. 2012. Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology* 51, 4 (2012), 724–731.

Friederike Eyssel, Dieta Kuchenbrandt, and Simon Bobinger. 2011. Effects of anticipated human-robot interaction and predictability of robot behavior on perceptions of anthropomorphism. In *Proceedings of the 2011 6th ACM/IEEE International Conference on Human-Robot Interaction*. 61–68.

Friederike Eyssel and Natalia Reich. 2013. Loneliness makes the heart grow fonder (of robots) – On the effects of loneliness on psychological anthropomorphism. In *Proceedings of the 2013 8th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 121–122.

Friederike Eyssel, Ricarda Wullenkord, and Verena Nitsch. 2017. The role of self-disclosure in human-robot interaction. In *Proceedings of the 2017 26th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 922–927.

Friederike A. Eyssel and Michaela Pfundmair. 2015. Predictors of psychological anthropomorphization, mind perception, and the fulfillment of social needs: A case study with a zoomorphic robot. In *Proceedings of the 2015 24th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 827–832.

Oriel FeldmanHall and Amitai Shenhav. 2019. Resolving uncertainty in a social world. *Nature Human Behaviour* 3, 5 (2019), 426–435.

Brian Fiala, Adam Arico, and Shaun Nichols. 2014. You, robot. In *Proceedings of the Current controversies in experimental philosophy*. Edouard Machery and Elizabeth O'Neill (Eds.), Routledge, 31–47.

Stephen M. Fiore, Travis J. Wiltshire, Emilio J. C. Lobato, Florian G. Jentsch, Wesley H. Huang, and Benjamin Axelrod. 2013. Toward understanding social cues and signals in human–robot interaction: effects of robot gaze and proxemic behavior. *Frontiers in Psychology* 4 (2013), 859.

Susan T. Fiske and Shelley E. Taylor. 1991. *Social Cognition*. Mcgraw-Hill Book Company.

Jerry Fodor. 1975. *The Language of Thought*. Vol. 5. Harvard University Press.

Jerry Fodor. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Vol. 2. MIT Press.

Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A survey of socially interactive robots. *Robotics and Autonomous Systems* 42, 3–4 (2003), 143–166.

Marlena R. Fraune, Benjamin C. Oisted, Catherine E. Sembrowski, Kathryn A. Gates, Margaret M. Krupp, and Selma Šabanović. 2020. Effects of robot-human versus robot-robot behavior and entitativity on anthropomorphism and willingness to interact. *Computers in Human Behavior* 105 (2020), 106220.

Changzeng Fu, Yuichiro Yoshikawa, Takamasa Iio, and Hiroshi Ishiguro. 2021. Sharing experiences to help a robot present its mind and sociability. *International Journal of Social Robotics* 13, 2 (2021), 341–352.

Susan R. Fussell, Sara Kiesler, Leslie D. Setlock, and Victoria Yew. 2008. How people anthropomorphize robots. In *Proceedings of the 2008 3rd ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 145–152.

Leonardo Giusti and Patrizia Marti. 2006. Interpretative dynamics in human robot interaction. In *Proceedings of the 2006 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 111–116.

Maria Ida Gobbini, Claudio Gentili, Emiliano Ricciardi, Claudia Bellucci, Pericle Salvini, Cecilia Laschi, Mario Guazzelli, and Pietro Pietrini. 2011. Distinct neural systems involved in agency and animacy detection. *Journal of Cognitive Neuroscience* 23, 8 (2011), 1911–1920.

Heather Gray, Kurt Gray, and Daniel Wegner. 2007. Dimensions of mind perception. *Science* 315, 5812 (2007), 619–619.

Kurt Gray and Daniel M. Wegner. 2012. Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition* 125, 1 (2012), 125–130.

Anthony G. Greenwald and Mahzarin R. Banaji. 1995. Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review* 102, 1 (1995), 4–23.

Richard Griffin and Simon Baron-Cohen. 2002. The intentional stance: Developmental and neurocognitive perspectives. In *Daniel Dennett*, A. Brook and D. Ross (Eds.). Cambridge University Press, 83–116.

Michael Gusenbauer and Neal R. Haddaway. 2020. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods* 11, 2 (2020), 181–217.

Azra Habibovic, Victor Malmsten Lundgren, Jonas Andersson, Maria Klingegård, Tobias Lagström, Anna Sirkka, Johan Fagerlönn, Edgren Claes, Rikard Fredriksson, Stas Krupenia, Dennis Saluäär, and Pontus Larsson. 2018. Communicating intent of automated vehicles to pedestrians. *Frontiers in Psychology* 9 (2018), 1336.

Glenda Hannibal. 2014. 'Dynamic' categorization and rationalized ascription: A study on NAO. In *Proceedings of the Robophilosophy*. 343–347.

Kerstin Haring, Kristin Nye, Ryan Darby, Elizabeth Phillips, Ewart de Visser, and Chad Tossell. 2019. I'm not playing anymore! a study comparing perceptions of robot and human cheating behavior. In *Proceedings of the International Conference on Social Robotics*. Springer, 410–419.

Kerstin S. Haring, David Silvera-Tawil, Tomotaka Takahashi, Mari Velonaki, and Katsumi Watanabe. 2015. Perception of a humanoid robot: A cross-cultural comparison. In *Proceedings of the 2015 24th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 821–826.

Fritz Heider. 1958. *The Psychology of Interpersonal Relations*. New York: Wiley.

Fritz Heider and Marianne Simmel. 1944. An experimental study of apparent behavior. *The American Journal of Psychology* 57, 2 (1944), 243–259.

Saskia Heijnen, Roy De Kleijn, and Bernhard Hommel. 2019. The impact of human–robot synchronization on anthropomorphization. *Frontiers in Psychology* 9 (2019), 2607.

Zachary Henkel, Cindy L. Bethel, John Kelly, Alexis Jones, Kristen Stives, Zach Buchanan, Deborah K. Eakin, David C. May, and Melinda Pilkinton. 2017. He can read your mind: Perceptions of a character-guessing robot. In *Proceedings of the 2017 26th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 242–247.

Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences* 33, 2–3 (2010), 61–83.

Matthias Hoenen, Katrin T. Lübke, and Bettina M. Pause. 2016. Non-anthropomorphic robots as social entities on a neurophysiological level. *Computers in Human Behavior* 57 (2016), 182–186.

Galit Hofree, Paul Ruvolo, Marian Stewart Bartlett, and Piotr Winkielman. 2014. Bridging the mechanical and the human mind: Spontaneous mimicry of a physically present android. *PLOS ONE* 9, 7 (2014), e99934.

Colin Holbrook. 2018. Cues of violent intergroup conflict diminish perceptions of robotic personhood. *ACM Transactions on Interactive Intelligent Systems* 8, 4 (2018), 1–17.

Ruud Hortensius and Emily S. Cross. 2018. From automata to animate beings: The scope and limits of attributing socialness to artificial agents. *Annals of the New York Academy of Sciences* 1426, 1 (2018), 93–110.

Ruud Hortensius, Felix Hekele, and Emily S. Cross. 2018. The perception of emotion in artificial agents. *IEEE Transactions on Cognitive and Developmental Systems* 10, 4 (2018), 852–864.

Bryce Huebner. 2010. Commonsense concepts of phenomenal consciousness: Does anyone care about functional zombies? *Phenomenology and the Cognitive Sciences* 9, 1 (2010), 133–155.

Edwin Hutchins. 1995. *Cognition in the Wild*. Number 1995. MIT Press.

Yuto Imamura, Kazunori Terada, and Hideyuki Takahashi. 2015. Effects of behavioral complexity on intention attribution to robots. In *Proceedings of the 2015 3rd International Conference on Human-Agent Interaction*. 65–72.

Tatsunori Ishii and Katsumi Watanabe. 2019. How people attribute minds to non-living entities. In *Proceedings of the 2019 11th International Conference on Knowledge and Smart Technology*. IEEE, 213–217.

Shoji Itakura, Hiraku Ishida, Takayuki Kanda, Yohko Shimada, Hiroshi Ishiguro, and Kang Lee. 2008. How to build an intentional android: Infants' imitation of a robot's goal-directed actions. *Infancy* 13, 5 (2008), 519–532.

Akira Ito, Shunsuke Hayakawa, and Tazunori Terada. 2004. Why robots need body for mind communication-an attempt of eye-contact between human and robot. In *Proceedings of the 2004 13th IEEE International Workshop on Robot and Human Interactive Communication*. IEEE, 473–478.

Christopher Brett Jaeger and Daniel T. Levin. 2016. If asimo thinks, does roomba feel? the legal implications of attributing agency to technology. *Journal of Human-Robot Interaction* 5, 3 (2016), 3–25.

Jennifer L. Jipson and Susan A. Gelman. 2007. Robots and rodents: Children's inferences about living and nonliving kinds. *Child Development* 78, 6 (2007), 1675–1688.

Malte Jung and Pamela Hinds. 2018. Robots in the wild: A time for more robust theories of human-robot interaction. *ACM Transactions on Human-Robot Interaction* 7, 1 (2018), 5 pages.

Peter H. Kahn Jr, Takayuki Kanda, Hiroshi Ishiguro, Brian T. Gill, Jolina H. Ruckert, Solace Shen, Heather E. Gary, Aimee L. Reichert, Nathan G. Freier, and Rachel L. Severson. 2012. Do people hold a humanoid robot morally accountable for the harm it causes?. In *Proceedings of the 2012 7th ACM/IEEE International Conference on Human-Robot Interaction*. 33–40.

Kazunori Kamewari, Masaharu Kato, Takayuki Kanda, Hiroshi Ishiguro, and Kazuo Hiraki. 2005. Six-and-a-half-month-old children positively attribute goals to human action and to humanoid-robot motion. *Cognitive Development* 20, 2 (2005), 303–320.

Hiroko Kamide, Friederike Eyssel, and Tatsuo Arai. 2013. Psychological anthropomorphism of robots. In *Proceedings of the International Conference on Social Robotics*. Springer, 199–208.

Frank Kaptein, Joost Broekens, Koen Hindriks, and Mark Neerincx. 2017. Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults. In *Proceedings of the 2017 26th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 676–682.

Merel Keijsers and Christoph Bartneck. 2018. Mindless robots get bullied. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 205–214.

Merel Keijsers, Hussain Kazmi, Friederike Eyssel, and Christoph Bartneck. 2021. Teaching robots a lesson: Determinants of robot punishment. *International Journal of Social Robotics* 13, 1 (2021), 41–54.

Harold H. Kelley. 1967. Attribution theory in social psychology. In *Proceedings of the Nebraska Symposium on Motivation*. University of Nebraska Press.

Sara Kiesler, Aaron Powers, Susan R. Fussell, and Cristen Torrey. 2008. Anthropomorphic interactions with a robot and robot–like agent. *Social Cognition* 26, 2 (2008), 169–181.

Barbara Kitchenham and Stuart Charters. 2007. *Guidelines for Performing Systematic Literature Reviews in Software Engineering*. EBSE Technical report, Ver. 2.3.

Elly A. Konijn and Johan F. Hoorn. 2020. Differential facial articulacy in robots and humans elicit different levels of responsiveness, empathy, and projected feelings. *Robotics* 9, 4 (2020), 92.

Joanna Korman, Anthony Harrison, Malcolm McCurry, and Greg Trafton. 2019. Beyond programming: Can robots' norm-violating actions elicit mental state attributions?. In *Proceedings of the 2019 14th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 530–531.

Sören Krach, Frank Hegel, Britta Wrede, Gerhard Sagerer, Ferdinand Binkofski, and Tilo Kircher. 2008. Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PLOS ONE* 3, 7 (2008), e2597.

Aleksandra Kupferberg, Stefan Glasauer, and Judith M. Burkart. 2013. Do robots have goals? How agent cues influence action understanding in non-human primates. *Behavioural Brain Research* 246 (2013), 47–54.

Aleksandra Kupferberg, Marco Iacoboni, Virginia Flanagin, Markus Huber, Anna Kasparbauer, Thomas Baumgartner, Gregor Hasler, Florian Schmidt, Christoph Borst, and Stefan Glasauer. 2018. Fronto-parietal coding of goal-directed actions performed by artificial agents. *Human Brain Mapping* 39, 3 (2018), 1145–1162.

Dennis Küster and Aleksandra Swiderska. 2021. Seeing the mind of robots: Harm augments mind perception but benevolent intentions reduce dehumanisation of artificial entities in visual vignettes. *International Journal of Psychology* 56, 3 (2021), 454–465.

Gabriella Lakatos, Márta Gácsi, Veronika Konok, Ildikó Brúder, Boróka Bereczky, Péter Korondi, and Ádám Miklósi. 2014. Emotion attribution to a non-humanoid robot in different social situations. *PLOS ONE* 9, 12 (2014), e114207.

Paul J. Lavrakas. 2008. *Encyclopedia of Survey Research Methods*. Sage Publications.

Theresa Law, Josh de Leeuw, and John H. Long. 2021. How movements of a non-humanoid robot affect emotional perceptions and trust. *International Journal of Social Robotics* 8 (2021), 1967–1978.

Minha Lee, Gale Lucas, and Jonathan Gratch. 2021. Comparing mind perception in strategic exchanges: Human-agent negotiation, dictator and ultimatum games. *Journal on Multimodal User Interfaces* 15, 2 (2021), 201–214.

Sau-lai Lee, Ivy Yee-man Lau, Sara Kiesler, and Chi-Yue Chiu. 2005. Human mental models of humanoid robots. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*. IEEE, 2767–2772.

Deniz Lefkeli, Yagmur Ozbay, Zeynep Gürhan-Canli, and Terry Eskenazi. 2021. Competing with or against cozmo, the robot: Influence of interaction context and outcome on mind perception. *International Journal of Social Robotics* 4 (2021), 715–754.

Séverin Lemaignan, Julia Fink, Francesco Mondada, and Pierre Dillenbourg. 2015. You're doing it wrong! studying unexpected behaviors in child-robot interaction. In *Proceedings of the International Conference on Social Robotics*. Springer, 390–400.

Daniel T. Levin, Stephen S. Killingsworth, Megan M. Saylor, Stephen M. Gordon, and Kazuhiko Kawamura. 2013. Tests of concepts about different kinds of minds: Predictions about the behavior of computers, robots, and people. *Human–Computer Interaction* 28, 2 (2013), 161–191.

Daniel T. Levin, Megan M. Saylor, and Simon D. Lynn. 2012. Distinguishing first-line defaults and second-line conceptualization in reasoning about humans, robots, and computers. *International Journal of Human-Computer Studies* 70, 8 (2012), 527–534.

Jamy Li. 2015. The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies* 77 (2015), 23–37.

Christian List and Philip Pettit. 2011. *Group agency: The Possibility, Design, and Status of Corporate Agents*. Oxford University Press.

Hamed Mahzoon, Kohei Ogawa, Yuichiro Yoshikawa, Michiko Tanaka, Kento Ogawa, Ryouta Miyazaki, Yusaku Ota, and Hiroshi Ishiguro. 2019. Effect of self-representation of interaction history by the robot on perceptions of mind and positive relationship: a case study on a home-use robot. *Advanced Robotics* 33, 21 (2019), 1112–1128.

Arielle R. Mandell, Melissa Smith, and Eva Wiese. 2017. Mind perception in humanoid agents has negative effects on cognitive processing. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. SAGE Publications Sage CA: Los Angeles, CA, 1585–1589.

Federico Manzi, Giulia Peretti, Cinzia Di Dio, Angelo Cangelosi, Shoji Itakura, Takayuki Kanda, Hiroshi Ishiguro, Davide Massaro, and Antonella Marchetti. 2020. A robot is not worth another: exploring children's mental state attribution to different humanoid robots. *Frontiers in Psychology* 11 (2020), 2011.

Serena Marchesi, Davide Ghiglino, Francesca Ciardo, Jairo Perez-Osorio, Ebru Baykara, and Agnieszka Wykowska. 2019. Do we adopt the intentional stance toward humanoid robots? *Frontiers in Psychology* 10 (2019), 450.

Serena Marchesi, Nicolas Spatola, Jairo Perez-Osorio, and Agnieszka Wykowska. 2021. Human vs. Humanoid. A behavioral investigation of the individual tendency to adopt the intentional stance. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 332–340.

Antonella Marchetti, Federico Manzi, Shoji Itakura, and Davide Massaro. 2018. Theory of mind and humanoid robots from a lifespan perspective. *Zeitschrift für Psychologie* 226, 2 (2018), 98–109.

Dorothea Ulrike Martin, Conrad Perry, Madeline Isabel MacIntyre, Luisa Varcoe, Sonja Pedell, and Jordy Kaufman. 2020. Investigating the nature of children's altruism using a social humanoid robot. *Computers in Human Behavior* 104 (2020), 106149.

Molly C. Martini, Christian A. Gonzalez, and Eva Wiese. 2016. Seeing minds in others–Can agents with robotic appearance have human-like preferences? *PLOS ONE* 11, 1 (2016), e0146310.

Molly C. Martini, Rabia Murtza, and Eva Wiese. 2015. Minimal physical features required for social robots. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. SAGE Publications Sage CA: Los Angeles, CA, 1438–1442.

Gail F. Melson, Peter H. Kahn, Jr, Alan Beck, and Batya Friedman. 2009. Robotic pets in human lives: Implications for the human–animal bond and for human relationships with personified technologies. *Journal of Social Issues* 65, 3 (2009), 545–567.

Steven J. Mithen. 1998. *The Prehistory of the Mind: A Search for the Origins of Art, Religion and Science*. Phoenix London.

Tomohito Miyake, Yuji Kawai, Jihoon Park, Jiro Shimaya, Hideyuki Takahashi, and Minoru Asada. 2019. Mind perception and causal attribution for failure in a game with a robot. In *Proceedings of the 2019 28th IEEE International Conference on Robot and Human Interactive Communication*. IEEE, 1–6.

Carey K. Morewedge, Jesse Preston, and Daniel M. Wegner. 2007. Timescale bias in the attribution of mind. *Journal of Personality and Social Psychology* 93, 1 (2007), 1.

Masahiro Mori, Karl F. MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine* 19, 2 (2012), 98–100.

Wenxuan Mou, Martina Ruocco, Debora Zanatto, and Angelo Cangelosi. 2020. When would you trust a robot? A study on trust and theory of mind in human-robot interactions. In *Proceedings of the 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 956–962.

Barbara C. N. Müller, Xin Gao, Sari R. R. Nijssen, and Tom G. E. Damen. 2020. I, robot: How human appearance and mind attribution relate to the perceived danger of robots. *International Journal of Social Robotics* 13, 4 (2020), 691–701.

Bilge Mutlu, Fumitaka Yamaoka, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. Nonverbal leakage in robots: Communication of intentions through seemingly unintentional behavior. In *Proceedings of the 2009 4th ACM/IEEE International Conference on Human-Robot Interaction*. 69–76.

Shaun Nichols and Stephen P. Stich. 2003. *Mindreading: An Integrated Account of Pretence, Self-awareness, and Understanding Other Minds*. Clarendon Press/Oxford University Press.

Milena K. Nigam and David Klahr. 2000. If robots make choices, are they alive?: Children's judgments of the animacy of intelligent artifacts. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Sari R. R. Nijssen, Barbara C. N. Müller, Rick B. van Baaren, and Markus Paulus. 2019. Saving the robot or the human? Robots who feel deserve moral care. *Social Cognition* 37, 1 (2019), 41–S2.

Richard E. Nisbett and Timothy D. Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84, 3 (1977), 231.

Mako Okanda, Kosuke Taniguchi, and Shoji Itakura. 2019. The role of animism tendencies and empathy in adult evaluations of robot. In *Proceedings of the 7th International Conference on Human-Agent Interaction*. 51–58.

Mako Okanda, Kosuke Taniguchi, Ying Wang, and Shoji Itakura. 2021. Preschoolers' and adults' animism tendencies toward a humanoid robot. *Computers in Human Behavior* 118 (2021), 106688.

Sandra Y. Okita, Daniel L. Schwartz, Takanori Shibata, and Hideyuki Tokuda. 2005. Exploring young children's attributions through entertainment robots. In *Proceedings of the ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.* IEEE, 390–395.

Ceylan Özdem, Eva Wiese, Agnieszka Wykowska, Hermann Müller, Marcel Brass, and Frank Van Overwalle. 2017. Believing androids–fMRI activation in the right temporo-parietal junction is modulated by ascribing intentions to non-human agents. *Social Neuroscience* 12, 5 (2017), 582–593.

Maike Paetzel, Ginevra Castellano, Giovanna Varni, Isabelle Hupont, Mohamed Chetouani, and Christopher Peters. 2018. The attribution of emotional state-how embodiment features and social traits affect the perception of an artificial agent. In *Proceedings of the 2018 27th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 495–502.

Delroy L. Paulhus and Simine Vazire. 2007. The self-report method. In *Proceedings of the Handbook of Research Methods in Personality Psychology*, Richard W. Robins, R. Chris Fraley, and Robert F. Krueger (Eds.), Guilford Press, 224–239.

Anthony Peressini. 2014. Blurring two conceptions of subjective experience: Folk versus philosophical phenomenality. *Philosophical Psychology* 27, 6 (2014), 862–889.

Jairo Perez-Osorio, Serena Marchesi, Davide Ghiglino, Melis Ince, and Agnieszka Wykowska. 2019. More than you expect: Priors influence on the adoption of intentional stance toward humanoid robots. In *Proceedings of the Social Robotics*. Miguel A. Salichs, Shuzhi Sam Ge, Emilia Ivanova Barakova, John-John Cabibihan, Alan R. Wagner, Álvaro Castro-González, and Hongsheng He (Eds.), Springer International Publishing, Cham, 119–129.

Jairo Perez-Osorio and Agnieszka Wykowska. 2020. Adopting the intentional stance toward natural and artificial agents. *Philosophical Psychology* 33, 3 (2020), 369–395.

Aaron Powers, Adam Kramer, Shirlene Lim, Jean Kuo, Sau-lai Lee, and Sara Kiesler. 2005. Eliciting information from people with a gendered humanoid robot. In *Proceedings of the 2005 IEEE International Workshop on Robot and Human Interactive Communication*. IEEE, 158–163.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 1, 4 (1978), 515–526.

Susanne Quadflieg, Israr Ul-Haq, and Nikolaos Mavridis. 2016. Now you feel it, now you don't: How observing human-robot interactions and human-human interactions can make you feel eerie. *Interaction Studies* 17, 2 (2016), 211–247.

Stéphane Raffard, Catherine Bortolon, Mahdi Khoramshahi, Robin N. Salesse, Marianna Burca, Ludovic Marin, Benoit G. Bardy, Aude Billard, Valérie Macioce, and Delphine Capdevielle. 2016. Humanoid robots versus humans: How is emotional valence of facial expressions recognized by individuals with schizophrenia? An exploratory study. *Schizophrenia Research* 176, 2–3 (2016), 506–513.

Matthew Rueben, Jeffrey Klow, Madelyn Duer, Eric Zimmerman, Jennifer Piacentini, Madison Browning, Frank J. Bernieri, Cindy M. Grimm, and William D. Smart. 2021. Mental models of a mobile shoe rack: Exploratory findings from a long-term in-the-wild study. *ACM Transactions on Human-Robot Interaction* 10, 2 (2021), 1–36.

Martin Saerbeck and Christoph Bartneck. 2010. Perception of affect elicited by robot motion. In *Proceedings of the 2010 5th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 53–60.

Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2011. Effects of gesture on the perception of psychological anthropomorphism: A case study with a humanoid robot. In *Proceedings of the International Conference on Social Robotics*. Springer, 31–41.

Megan M. Saylor and Daniel T. Levin. 2005. Thinking and seeing in intentional and mechanical systems. In *Proceedings of the ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.* IEEE, 710–715.

Brian Scassellati. 2002. Theory of mind for a humanoid robot. *Autonomous Robots* 12, 1 (2002), 13–24.

Elef Schellen and Agnieszka Wykowska. 2019. Intentional mindset toward robots–open questions and methodological challenges. *Frontiers in Robotics and AI* 5 (2019), 139.

Matthias Scheutz. 2002. *Computationalism: New Directions*. MIT Press.

Alessandra Sciutti, Ambra Bisio, Francesco Nori, Giorgio Metta, Luciano Fadiga, and Giulio Sandini. 2013. Robots can be perceived as goal-oriented agents. *Interaction Studies* 14, 3 (2013), 329–350.

John R. Searle. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3 (1980), 417–457.

Rachel L. Severson and Kristi M. Lemm. 2016. Kids see human too: Adapting an individual differences measure of anthropomorphism for a child sample. *Journal of Cognition and Development* 17, 1 (2016), 122–141.

Elaine Short, Justin Hart, Michelle Vu, and Brian Scassellati. 2010. No fair!! an interaction with a cheating robot. In *Proceedings of the 2010 5th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 219–226.

David Sirkin, Brian Mok, Stephen Yang, and Wendy Ju. 2015. Mechanical ottoman: How robotic furniture offers and withdraws support. In *Proceedings of the 2015 10th ACM/IEEE International Conference on Human-Robot Interaction*. 11–18.

Brian Cantwell Smith. 2019. *The Promise of Artificial Intelligence: Reckoning and Judgment*. MIT Press.

Mark C. Somanader, Megan M. Saylor, and Daniel T. Levin. 2011. Remote control and children's understanding of robots. *Journal of Experimental Child Psychology* 109, 2 (2011), 239–247.

Kristyn Sommer, Mark Nielsen, Madeline Draheim, Jonathan Redshaw, Eric J. Vanman, and Matti Wilks. 2019. Children's perceptions of the moral worth of live agents, robots, and inanimate objects. *Journal of Experimental Child Psychology* 187 (2019), 104656.

Nicolas Spatola and Olga A. Wudarczyk. 2021. Implicit attitudes towards robots predict explicit attitudes, semantic distance between robots and humans, anthropomorphism, and prosocial behavior: From attitudes to human–robot interaction. *International Journal of Social Robotics* 13, 5 (2021), 1149–1159.

Karen Spektor-Precel and David Mioduser. 2015. The influence of constructing robot's behavior on the development of theory of mind (ToM) and theory of artificial mind (ToAM) in young children. In *Proceedings of the 14th International Conference on Interaction Design and Children*. 311–314.

Rebecca Q. Stafford, Bruce A. MacDonald, Chandimal Jayawardena, Daniel M. Wegner, and Elizabeth Broadbent. 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 (2014), 17–32.

Ilona Straub. 2016. 'It looks like a human!' The interrelation of social presence, interaction and agency ascription: A case study about the effects of an android robot on social agency ascription. *AI & Society* 31, 4 (2016), 553–571.

Stephanie Sturgeon, Andrew Palmer, Janelle Blankenburg, and David Feil-Seifer. 2019. Perception of social intelligence in robots performing false-belief tasks. In *Proceedings of the 2019 28th IEEE International Conference on Robot and Human Interactive Communication*. IEEE, 1–7.

Kaveri Subrahmanyam, Rochel Gelman, and Alyssa Lafosse. 2002. Animates and other separably moveable objects. In *Proceedings of the Category Specificity in Brain and Mind*. Emer Forde and Glyn Humphreys (Eds.), Psychology Press New York, NY, 341–373.

Harry Surden and Mary-Anne Williams. 2016. Technological opacity, predictability, and self-driving cars. *Cardozo Law Review* 38, 1 (2016), 121.

Aleksandra Swiderska and Dennis Küster. 2018. Avatars in pain: Visible harm enhances mind perception in humans and robots. *Perception* 47, 12 (2018), 1139–1152.

Justin Sytsma and Edouard Machery. 2010. Two conceptions of subjective experience. *Philosophical Studies* 151, 2 (2010), 299–327.

Hideyuki Takahashi, Midori Ban, and Minoru Asada. 2016. Semantic differential scale method can reveal multi-dimensional aspects of mind perception. *Frontiers in Psychology* 7 (2016), 1717.

Hideyuki Takahashi, Chinatsu Saito, Hiroyuki Okada, and Takashi Omori. 2013. An investigation of social factors related to online mentalizing in a human-robot competitive game. *Japanese Psychological Research* 55, 2 (2013), 144–153.

Hideyuki Takahashi, Kazunori Terada, Tomoyo Morita, Shinsuke Suzuki, Tomoki Haji, Hideki Kozima, Masahiro Yoshikawa, Yoshio Matsumoto, Takashi Omori, Minoru Asada, and Eiichi Naito. 2014. Different impressions of other agents obtained through social interaction uniquely modulate dorsal and ventral pathway activities in the social human brain. *Cortex* 58 (2014), 289–300.

Haodan Tan, Dakuo Wang, and Selma Sabanovic. 2018. Projecting life onto robots: The effects of cultural factors and design type on multi-level evaluations of robot anthropomorphism. In *Proceedings of the 2018 27th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 129–136.

Tetsushi Tanibe, Takaaki Hashimoto, and Kaori Karasawa. 2017. We perceive a mind in a robot when we help it. *PLOS ONE* 12, 7 (2017), e0180952.

Kyohei Tatsukawa, Hideyuki Takahashi, Yuichiro Yoshikawa, and Hiroshi Ishiguro. 2019. Android pretending to have similar traits of imagination as humans evokes stronger perceived capacity to feel. *Frontiers in Robotics and AI* 6 (2019), 88.

Kazunori Terada, Takashi Shamoto, Akira Ito, and Haiying Mei. 2007. Reactive movements of non-humanoid robots cause intention attribution in humans. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 3715–3720.

Kazunori Terada and Seiji Yamada. 2017. Mind-reading and behavior-reading against agents with and without anthropomorphic features in a competitive situation. *Frontiers in Psychology* 8 (2017), 1071.

Sam Thellman. 2021. *Social Robots as Intentional Agents*. Ph.D. Dissertation. Linköping University Electronic Press.

Sam Thellman, Asenia Giagtzidou, Annika Silvervarg, and Tom Ziemke. 2020. An implicit, non-verbal measure of belief attribution to robots. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 473–475.

Sam Thellman, Annika Silvervarg, and Tom Ziemke. 2017. Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots. *Frontiers in Psychology* 8 (2017), 1962.

Sam Thellman and Tom Ziemke. 2019. The intentional stance toward robots: Conceptual and methodological considerations. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society*. 1097–1103.

Sam Thellman and Tom Ziemke. 2020. Do you see what i see? Tracking the perceptual beliefs of robots. *iScience* 23, 10 (2020), 101625.

Sam Thellman and Tom Ziemke. 2021. The perceptual belief problem: Why explainability is a tough challenge in social robotics. *ACM Transactions on Human-Robot Interaction* 10, 3 (2021), 15 pages.

Gabriele Trovato and Friederike Eyssel. 2017. Mind attribution to androids: A comparative study with Italian and Japanese adolescents. In *Proceedings of the 2017 26th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 561–566.

Alan Turing. 1950. Computing machinery and intelligence. *Mind* 59 (1950), 433–460.

Rianne van den Berghe, Mirjam de Haas, Ora Oudgenoeg-Paz, Emiel Krahmer, Josje Verhagen, Paul Vogt, Bram Willemsen, Jan de Wit, and Paul Leseman. 2021. A toy or a friend? Children's anthropomorphic beliefs about robots and how these relate to second-language word learning. *Journal of Computer Assisted Learning* 37, 2 (2021), 396–410.

Sophie van der Woerdt and Pim Haselager. 2019. When robots appear to have a mind: The human perception of machine agency and responsibility. *New Ideas in Psychology* 54 (2019), 93–100.

Mike van Duuren and Michael Scaife. 1996. "Because a robot's brain hasn't got a brain, it just controls itself"–Children's attributions of brain related behaviour to intelligent artefacts. *European Journal of Psychology of Education* 11, 4 (1996), 365–376.

Frank Van Overwalle and Kris Baetens. 2009. Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. *Neuroimage* 48, 3 (2009), 564–584.

Iris van Rooij and Giosuè Baggio. 2020. Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science* 16, 4 (2020), 682–687.

Caroline L. van Straten, Jochen Peter, Rinaldo Kühne, and Alex Barco. 2020. Transparency about a robot's lack of human psychological capacities: Effects on child-robot perception and relationship formation. *ACM Transactions on Human-Robot Interaction* 9, 2 (2020), 1–22.

Caroline L. Van Straten, Jochen Peter, Rinaldo Kühne, and Alex Barco. 2021. The wizard and I: How transparent teleoperation and self-description (do not) affect children's robot perceptions and child-robot relationship formation. *AI & Society* 37, 1 (2021), 383–399.

Sebastian Wallkötter, Rebecca Stower, Arvid Kappas, and Ginevra Castellano. 2020. A robot by any other frame: Framing and behaviour influence mind perception in virtual but not real-world environments. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 609–618.

Xijing Wang and Eva G. Krumhuber. 2018. Mind perception of robots varies with their economic versus social function. *Frontiers in Psychology* 9 (2018), 1230.

Yin Wang and Susanne Quadflieg. 2015. In our own image? Emotional and neural processing differences when observing human–human vs. human–robot interactions. *Social Cognitive and Affective Neuroscience* 10, 11 (2015), 1515–1524.

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

Adam Waytz, Kurt Gray, Nicholas Epley, and Daniel M. Wegner. 2010a. Causes and consequences of mind perception. *Trends in Cognitive Sciences* 14, 8 (2010), 383–388.

Adam Waytz, Carey K. Morewedge, Nicholas Epley, George Monteleone, Jia-Hong Gao, and John T. Cacioppo. 2010b. Making sense by making sentient: Effectance motivation increases anthropomorphism. *Journal of Personality and Social Psychology* 99, 3 (2010), 410.

Kara Weisman, Carol S. Dweck, and Ellen M. Markman. 2017. Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences* 114, 43 (2017), 11374–11379.

Alexander Wendt. 1999. *Social Theory of International Politics*. Vol. 67. Cambridge University Press.

Eva Wiese, Arielle Mandell, Tyler Shaw, and Melissa Smith. 2019. Implicit mind perception alters vigilance performance because of cognitive conflict processing. *Journal of Experimental Psychology: Applied* 25, 1 (2019), 25–40.

Eva Wiese, Tyler Shaw, Daniel Lofaro, and Carryl Baldwin. 2017. Designing artificial agents as social companions. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. SAGE Publications Sage CA: Los Angeles, CA, 1604–1608.

Eva Wiese, Patrick P. Weis, Yochanan Bigman, Kyra Kapsaskis, and Kurt Gray. 2021. It's a match: Task assignment in human–robot collaboration depends on mind perception. *International Journal of Social Robotics* 14, 1 (2021), 141–148.

Eva Wiese, Agnieszka Wykowska, Jan Zwickel, and Hermann J. Müller. 2012. I see what you mean: How attentional selection is shaped by ascribing intentions to others. *PLOS ONE* 7, 9 (2012), e45391.

Tom Williams, Daniel Szafir, Tathagata Chakraborti, and Heni Ben Amor. 2018. Virtual, augmented, and mixed reality for human-robot interaction. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 403–404.

Agnieszka Wykowska, Eva Wiese, Aaron Prosser, and Hermann J. Müller. 2014. Beliefs about the minds of others influence how we process sensory information. *PLOS ONE* 9, 4 (2014), e94339.

Yaqi Xie, Indu P. Bodala, Desmond C. Ong, David Hsu, and Harold Soh. 2019. Robot capability and intention in trust-based decisions across tasks. In *Proceedings of the 2019 14th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 39–47.

Xiaoyu Xu and Sela Sar. 2018. Do we see machines the same way as we see humans? A survey on mind perception of machines and human beings. In *Proceedings of the 2018 27th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 472–475.

Yuto Yamaji, Taisuke Miyake, Yuta Yoshiike, P. Ravindra S. De Silva, and Michio Okada. 2010. STB: Intentional stance grounded child-dependent robot. In *Proceedings of the International Conference on Social Robotics*. Springer, 71–80.

Yaoxin Zhang, Wenxu Song, Zhenlin Tan, Yuyin Wang, Cheuk Man Lam, Sio Pan Hoi, Qianhan Xiong, Jiajia Chen, and Li Yi. 2019. Theory of robot mind: False belief attribution to social robots in children with and without autism. *Frontiers in Psychology* 10 (2019), 1732.

Xuan Zhao, Corey Cusimano, and Bertram F. Malle. 2016. Do people spontaneously take a robot's visual perspective?. In *Proceedings of the 2016 11th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 335–342.

Tom Ziemke. 2020. Understanding robots. *Science Robotics* 5, 46, Article abe2987 (2020).

Jakub Złotowski, Diane Proudfoot, Kumar Yogeeswaran, and Christoph Bartneck. 2015. Anthropomorphism: Opportunities and challenges in human–robot interaction. *International Journal of Social Robotics* 7, 3 (2015), 347–360.

Jakub Złotowski, Ewald Strasser, and Christoph Bartneck. 2014. Dimensions of anthropomorphism: From humanness to humanlikeness. In *Proceedings of the 2014 9th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 66–73.

Jakub Złotowski, Hidenobu Sumioka, Christoph Bartneck, Shuichi Nishio, and Hiroshi Ishiguro. 2017. Understanding anthropomorphism: Anthropomorphism is not a reverse process of dehumanization. In *Proceedings of the International Conference on Social Robotics*. Springer, 618–627.

Jakub Złotowski, Hidenobu Sumioka, Friederike Eyssel, Shuichi Nishio, Christoph Bartneck, and Hiroshi Ishiguro. 2018. Model of dual anthropomorphism: The relationship between the media equation effect and implicit anthropomorphism. *International Journal of Social Robotics* 10, 5 (2018), 701–714.

Received June 2021; revised January 2022; accepted February 2022